



PHD

## The Helmholtz Equation in Heterogeneous and Random Media: Analysis and Numerics

Pembrey, Owen

*Award date:*  
2020

*Awarding institution:*  
University of Bath

[Link to publication](#)

### Alternative formats

If you require this document in an alternative format, please contact:  
[openaccess@bath.ac.uk](mailto:openaccess@bath.ac.uk)

Copyright of this thesis rests with the author. Access is subject to the above licence, if given. If no licence is specified above, original content in this thesis is licensed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs 4.0 International (CC BY-NC-ND 4.0) Licence (<https://creativecommons.org/licenses/by-nc-nd/4.0/>). Any third-party copyright material present remains the property of its respective owner(s) and is licensed under its existing terms.

#### Take down policy

If you consider content within Bath's Research Portal to be in breach of UK law, please contact: [openaccess@bath.ac.uk](mailto:openaccess@bath.ac.uk) with the details. Your claim will be investigated and, where appropriate, the item will be removed from public view as soon as possible.

The Helmholtz Equation in  
Heterogeneous and Random  
Media: Analysis and Numerics

OWEN RHYS PEMBERY

*Thesis submitted for the degree of  
Doctor of Philosophy*

UNIVERSITY OF BATH

*Department of Mathematical Sciences*

FEBRUARY 2020

## COPYRIGHT

Attention is drawn to the fact that copyright of this thesis/portfolio rests with the author and copyright of any previously published materials included may rest with third parties. A copy of this thesis/portfolio has been supplied on condition that anyone who consults it understands that they must not copy it or use material from it except as licenced, permitted by law or with the consent of the author or other copyright owners, as applicable.

In particular, Chapter 3 is based on the paper ‘The Helmholtz equation in random media: well-posedness and a priori bounds’ [176] which was published in the SIAM/ASA Journal on Uncertainty Quantification and the copyright of this paper is owned by SIAM.

**Declaration of authorship**

I am the author of this thesis, and the work described therein was carried out by myself personally, with the exception of Chapter 4 where around 20% of the work was carried out by Dr. Euan Spence; who proved Theorem 4.11 (generalising my proof of Theorem 4.12), Theorem 4.15, and Lemma 4.16, and Appendix A, where Example A.1 was devised by Federico Cornalba.

**Declaration of any previous submission of the work**

The material presented here for examination for the award of a higher degree by research has not been incorporated into a submission for another degree.

Owen Rhys Pembroly .....



## Summary

The Helmholtz equation is the simplest possible model of wave propagation, describing time-harmonic solutions of the wave equation. We study the Helmholtz equation with heterogeneous and random coefficients, corresponding to wave propagation through a medium with spatially variable and random physical characteristics.

In Chapter 2 we study  $h$ -finite-element approximations of the (deterministic) heterogeneous Helmholtz equation. We prove the first sharp error bounds (explicit in the frequency) for the higher-order  $h$ -finite-element method for the Helmholtz equation in heterogeneous media

In Chapter 3 we move on to the stochastic Helmholtz equation, i.e., the Helmholtz equation with random field coefficients. We prove the first frequency-independent well-posedness results for the stochastic Helmholtz equation. (I.e. we prove existence and uniqueness of a solution, and we prove a frequency-explicit bound on the solution, under a suitable assumption on the coefficients that corresponds to the problem being ‘nontrapping’ almost surely.) To prove this well posedness, we develop general, abstract arguments that allow us to prove well-posedness for three different formulations of stochastic PDEs; these arguments can be applied more widely than just to the stochastic Helmholtz equation.

In Chapter 4 we study nearby preconditioning, a computational technique for speeding up solving many linear systems arising from discretisations of realisations of the stochastic Helmholtz equation. This speedup is achieved by reusing the preconditioner corresponding to one realisation for many other realisations. We prove rigorous results on when nearby preconditioning is effective, and investigate this effectiveness numerically in a range of situations.

Also in Chapter 4, we combine nearby preconditioning with a Quasi-Monte-Carlo method (i.e., a high-dimensional integration rule) to compute quantities of interest of the stochastic Helmholtz equation. We see that nearby preconditioning offers a significant computational saving, with 98% of linear-system solves being made with a previously-calculated preconditioner. As a by-product of these results we also provide some preliminary computational evidence, of independent interest, that when using QMC methods for the Helmholtz equation the total number of realisations used must increase with the frequency.

Finally, in Chapter 5, we study the Multi-Level Monte-Carlo method for the Helmholtz equation, where one reduces the variance in an Uncertainty Quantification calculation by solving the underlying PDE on a hierarchy of meshes. We generalise the standard Multi-Level Monte-Carlo convergence theory to the frequency-dependent case and prove rigorous frequency-explicit bounds on the computational cost of Monte Carlo and Multi-Level Monte-Carlo methods for the Helmholtz equation. We see that in many cases the Multi-Level Monte-Carlo method gives theoretical speedup over the Monte-Carlo Method.



## Acknowledgements

I feel incredibly blessed that I have many people to thank for their support and help through the years I've spent working on this thesis.

My thanks first go to my supervisors, Ivan Graham and Euan Spence. Ivan—thank you so much for all the support you've given me throughout my mathematical career, from being my personal tutor as an undergraduate, through supervising my undergraduate project, and now supervising this thesis. I've appreciated your honesty, the way you strive for excellence, and the way you've shown me that simple examples can often illuminate bigger mathematical truths. Euan—thank you for pushing me hard throughout my PhD, for forcing me to read Krantz and Higham, and for making me into (I hope) a much better writer and mathematician. Thank you also for generously funding me to travel to lots of conferences, for making those conferences a lot of fun, and for introducing me to lots of the 'waves' community. Both—thank you for your pastoral support, especially through the more demanding parts of this PhD; I truly feel that I have had one of the best supervisory teams.

There are lots of people to thank in the Department of Mathematical Sciences at the University of Bath; all of the SAMBa team past and present—Susie, Jess, Anna, Andreas, Paul, and many others—thank you for building a centre that is a fun place to study, and a great place to learn. My thanks also go to the Numerical Analysis group, for being a great blend of a friendly bunch of people who are also a stimulating group in which to do research. I'd also like to thank the High-Performance Computing team at Bath for their support in running my code on Balena, Bath's high-performance computer.

There are also numerous individuals to thank from Bath—Kieran Jarrett helped me get my head around measure theory for Chapter 3, Tony Shardlow gave us really helpful feedback on that same chapter. Tom Finn and Dan Ng helped me understand stopping times, and if Federico Cornalba hadn't discovered Example A.1, then Chapter 3 may never have been written! I've also been blessed with numerous office mates who have provided both stimulating conversation and welcome distraction (mathematical or otherwise); my thanks in particular go to Aoibheann Brady, Matt Durey, Matt Parkinson, and Kate Powers.

A special thanks goes to Will Saunders and Jack Betteridge, for putting up with my endless programming questions, often spending lots of their own time fixing my code, and convincing me to use version control and write tests for my code. A special thanks goes to Jack for being my Student-Led-Symposium-organising partner-in-crime and SAMBa-numerical-analysis-friend.

Looking at the wider academic community, I want to thank Rob Scheichl and Ralf Hiptmair for examining my thesis. Rob—thank you for lots of probing yet supportive questions and comments throughout my PhD and during my viva. Ralf—thank you for taking the time to read and examine my thesis, and for your very generous hospitality when I travelled to Zürich for my viva. My thanks also go to Stefan Sauter, Nilima Nigam and Théophile Chaumont-Frelet, for their suggestions which led to me proving the weaker-norm bounds in Section 4.5. I'm also very grateful to the Firedrake team for building the Firedrake software (without which, I wouldn't have been able to do all the numerical experiments in this thesis) and for very patiently answering

my questions when I got stuck. I'm also very thankful to EPSRC for funding my studentship, through SAMBa.

I'm also very blessed to have many people outside of mathematics who love me, support me, and tell me when I need to stop doing maths. A huge thanks should go to everyone at my church, St Bart's Bath, but especially to our very dear friends Will, Hannah, Emma, and Rich. Thank you all for lots of fun, laughter, love, support, and prayers. I'm also phenomenally grateful to my parents; Mum and Dad, thank you for pushing me to do my best when I was younger, but more than that, thank you for your love, care, support, and prayers in many, many ways, over many, many years. To my sister Anwen and brother-in-law Paul, thank you both for your friendship, companionship and love over the years, I'm incredible grateful. Also, Paul, thank you very much for helping me with the design of this thesis, so that it looks good!

Above else, my thanks go to God. He is the incredible Trinity who has created me, has created mathematics and has enabled me to spend my time thinking about and wrestling with His creation. But more than that, LORD, thank You for revealing Yourself in the Bible. Father, thank You for choosing me before the foundation of the world, not because of anything I have done, but because You love me. Jesus, thank You for being incarnate as a man, and loving me so much that You would die on the cross for me, taking all the punishment and pain that I deserve as a sinner. Holy Spirit, thank You for dwelling within me, and reassuring me that I have the hope of a glorious future with You. LORD, Your love for me is beyond compare.

And lastly, but by no means leastly, my thanks go to my wife, Rebecca. Reb, when we married I had already committed to studying for this PhD, and you have been such a wonderful support throughout it all. You've loved me dearly, supported and cared for me when I've needed it, persuaded me to keep going when I've wanted to give up, prayed for me, and had the wisdom to tell me when I should keep working and when I should stop for the evening. I don't think I'd have got here without you. As I finish this thesis, I'm reminded again of how blessed I am to be loved by you.

*Rebecca, this thesis is dedicated to you.*

# Contents

|                                                                                                                  |           |
|------------------------------------------------------------------------------------------------------------------|-----------|
| Contents                                                                                                         | 5         |
| List of Figures                                                                                                  | 9         |
| List of Tables                                                                                                   | 11        |
| List of Algorithms                                                                                               | 13        |
| <b>1 Introduction</b>                                                                                            | <b>15</b> |
| 1.1 The Subjects of the Thesis                                                                                   | 15        |
| 1.1.1 Motivation from Applications for the High-Frequency Stochastic Helmholtz Equation                          | 15        |
| 1.1.2 Solving the Helmholtz Equation Numerically                                                                 | 17        |
| 1.2 The Main Achievements of the Thesis                                                                          | 21        |
| 1.3 The Structure of the thesis                                                                                  | 23        |
| <b>2 PDE theory for the deterministic Helmholtz equation and the theory of its finite-element discretisation</b> | <b>25</b> |
| 2.1 Introduction                                                                                                 | 25        |
| 2.2 Wavenumber- and coefficient-explicit PDE theory of the deterministic Helmholtz equation                      | 25        |
| 2.2.1 Deterministic Helmholtz problems                                                                           | 26        |
| 2.2.2 Well-posedness and a priori bounds                                                                         | 28        |
| 2.2.3 Discussion of results on well-posedness and a priori bounds for the Helmholtz equation                     | 32        |
| 2.3 The theory of the $b$ -finite-element discretisation of the Helmholtz equation                               | 38        |
| 2.3.1 Variational formulations for the Helmholtz equation                                                        | 38        |
| 2.3.2 Background concepts in finite-element theory                                                               | 39        |
| 2.3.3 Discussion of the finite-element method for the Helmholtz equation                                         | 42        |
| 2.3.4 Extended discussion of proof techniques for finite-element errors for the Helmholtz equation               | 51        |
| 2.4 New finite-element-error bounds for the heterogeneous Helmholtz equation                                     | 62        |
| 2.4.1 Main result: new finite-element-error bounds                                                               | 63        |
| 2.4.2 Decomposition of solution and best approximation bound                                                     | 68        |
| 2.4.3 Routine analysis results                                                                                   | 75        |
| 2.4.4 Error bounds for Galerkin projections                                                                      | 76        |
| 2.4.5 Discrete Sobolev spaces                                                                                    | 81        |

|          |                                                                                                                  |            |
|----------|------------------------------------------------------------------------------------------------------------------|------------|
| 2.4.6    | Proof of Theorem 2.39 . . . . .                                                                                  | 90         |
| 2.4.7    | Constants from Section 2.4 . . . . .                                                                             | 101        |
| 2.5      | Summary and future work . . . . .                                                                                | 108        |
| 2.5.1    | Summary . . . . .                                                                                                | 108        |
| 2.5.2    | Future work . . . . .                                                                                            | 108        |
| <b>3</b> | <b>Well-posedness of formulations of the stochastic Helmholtz equation</b>                                       | <b>109</b> |
| 3.1      | Introduction . . . . .                                                                                           | 109        |
| 3.1.1    | Statement of main results . . . . .                                                                              | 111        |
| 3.1.2    | Random fields satisfying Condition 3.8 . . . . .                                                                 | 117        |
| 3.1.3    | Outline of the chapter . . . . .                                                                                 | 119        |
| 3.1.4    | Discussion of the main results in the context of other work on UQ for<br>time-harmonic wave equations . . . . .  | 119        |
| 3.2      | General results proving a priori bounds and well-posedness of stochastic varia-<br>tional formulations . . . . . | 120        |
| 3.2.1    | Notation and definitions of the variational formulations . . . . .                                               | 120        |
| 3.2.2    | Conditions on $\mathcal{A}$ , $\mathcal{L}$ , and $c$ . . . . .                                                  | 121        |
| 3.2.3    | Results on the equivalence of Problems MAS, SOAS, and SV . . . . .                                               | 122        |
| 3.3      | Proof of the results in Section 3.2 . . . . .                                                                    | 125        |
| 3.3.1    | Preliminary lemmas . . . . .                                                                                     | 125        |
| 3.3.2    | Proofs of Theorems 3.23, 3.25, and 3.26 and Lemmas 3.24 and 3.28 . . . . .                                       | 126        |
| 3.4      | Proofs of Theorems 3.7 and 3.10 . . . . .                                                                        | 130        |
| 3.4.1    | Placing the Helmholtz stochastic EDP into the framework of Section 3.2                                           | 130        |
| 3.4.2    | Verifying the Helmholtz stochastic EDP satisfies the general conditions<br>in Section 3.2 . . . . .              | 131        |
| 3.4.3    | Proofs of Theorems 3.7 and 3.10 . . . . .                                                                        | 135        |
| 3.5      | Summary and future work . . . . .                                                                                | 136        |
| 3.5.1    | Summary . . . . .                                                                                                | 136        |
| 3.5.2    | Future work . . . . .                                                                                            | 136        |
| <b>4</b> | <b>Nearby preconditioning for the Helmholtz equation</b>                                                         | <b>137</b> |
| 4.1      | Introduction and Motivation from UQ . . . . .                                                                    | 137        |
| 4.1.1    | Motivation from uncertainty quantification for the Helmholtz equation .                                          | 137        |
| 4.1.2    | Outline of the chapter . . . . .                                                                                 | 138        |
| 4.2      | Statement of the main results . . . . .                                                                          | 138        |
| 4.2.1    | Definition of variational problems and conditions used to prove main<br>results . . . . .                        | 138        |
| 4.2.2    | Definition of finite-element matrices, weighted norms, and weighted<br>GMRES . . . . .                           | 140        |
| 4.2.3    | Main results . . . . .                                                                                           | 142        |
| 4.2.4    | PDE analogues to Theorems 4.11 and 4.12 . . . . .                                                                | 144        |

|          |                                                                                                                                 |            |
|----------|---------------------------------------------------------------------------------------------------------------------------------|------------|
| 4.3      | Numerical experiments verifying and investigating the sharpness of Theorems 4.11 and 4.12 . . . . .                             | 146        |
| 4.4      | Proofs of Theorems 4.11, 4.12, and 4.15 and Lemma 4.16 . . . . .                                                                | 147        |
| 4.4.1    | Proof of the main ingredient of the proofs of Theorems 4.11 and 4.12 . . . . .                                                  | 147        |
| 4.4.2    | Proofs of the finite-element results Theorems 4.11 and 4.12 . . . . .                                                           | 159        |
| 4.4.3    | Proofs of the PDE results Theorem 4.15 and Lemma 4.16 . . . . .                                                                 | 160        |
| 4.5      | Extension of the nearby preconditioning results to weaker norms . . . . .                                                       | 163        |
| 4.5.1    | Theory in weaker norms . . . . .                                                                                                | 164        |
| 4.5.2    | Numerics in weaker norms . . . . .                                                                                              | 168        |
| 4.6      | Applying nearby preconditioning to a Quasi-Monte-Carlo method for the Helmholtz equation . . . . .                              | 172        |
| 4.6.1    | Brief description of QMC . . . . .                                                                                              | 172        |
| 4.6.2    | Methods for applying nearby preconditioning to QMC . . . . .                                                                    | 173        |
| 4.6.3    | Numerical Experiments . . . . .                                                                                                 | 181        |
| 4.7      | Review of related techniques in the literature . . . . .                                                                        | 186        |
| 4.8      | Probabilistic nearby preconditioning results . . . . .                                                                          | 194        |
| 4.8.1    | Probabilistic theory for nearby preconditioning . . . . .                                                                       | 195        |
| 4.8.2    | Numerical probabilistic results for nearby preconditioning . . . . .                                                            | 197        |
| 4.9      | Summary and future work . . . . .                                                                                               | 199        |
| 4.9.1    | Summary . . . . .                                                                                                               | 199        |
| 4.9.2    | Future work . . . . .                                                                                                           | 199        |
| <b>5</b> | <b>Monte-Carlo and Multi-Level Monte-Carlo methods for the stochastic Helmholtz equation . . . . .</b>                          | <b>201</b> |
| 5.1      | Introduction . . . . .                                                                                                          | 201        |
| 5.2      | Background on both Monte-Carlo and Multi-Level Monte-Carlo methods . . . . .                                                    | 202        |
| 5.2.1    | The ideas of Monte-Carlo and Multi-Level Monte-Carlo methods . . . . .                                                          | 202        |
| 5.2.2    | Challenges in Monte-Carlo and Multi-Level Monte-Carlo methods for the Helmholtz equation . . . . .                              | 203        |
| 5.2.3    | Literature Review of Multi-Level Monte-Carlo methods . . . . .                                                                  | 204        |
| 5.3      | An abstract setting for both the Multi-Level Monte-Carlo and Monte-Carlo methods, motivated by the Helmholtz equation . . . . . | 206        |
| 5.4      | Monte-Carlo methods . . . . .                                                                                                   | 210        |
| 5.5      | Multi-level Monte-Carlo methods . . . . .                                                                                       | 211        |
| 5.6      | Placing the stochastic Helmholtz equation in the abstract $k$ -dependent setting . . . . .                                      | 218        |
| 5.6.1    | Model problem and quantities of interest . . . . .                                                                              | 219        |
| 5.6.2    | Main result on Monte-Carlo and Multi-Level Monte-Carlo methods for the Helmholtz equation . . . . .                             | 220        |
| 5.6.3    | Proof of Theorem 5.22 . . . . .                                                                                                 | 224        |
| 5.7      | Summary and future work . . . . .                                                                                               | 225        |
| 5.7.1    | Summary . . . . .                                                                                                               | 225        |

|          |                                                                     |            |
|----------|---------------------------------------------------------------------|------------|
| 5.7.2    | Future work . . . . .                                               | 225        |
| <b>A</b> | <b>Failure of Fredholm theory</b>                                   | <b>227</b> |
| <b>B</b> | <b>Recap of basic material on measure theory and Bochner spaces</b> | <b>229</b> |
| B.1      | Recap of measure theory results . . . . .                           | 229        |
| B.2      | Recap of results on Bochner spaces . . . . .                        | 230        |
| <b>C</b> | <b>Measurability of series expansions (used in Section 3.1.2)</b>   | <b>233</b> |
| <b>D</b> | <b>Error estimators for complex random variables</b>                | <b>237</b> |
| <b>E</b> | <b>Numerical investigation of QMC</b>                               | <b>241</b> |
| <b>F</b> | <b>Additional probabilistic results</b>                             | <b>265</b> |
| <b>G</b> | <b>Computational set-up</b>                                         | <b>269</b> |
|          | <b>Bibliography</b>                                                 | <b>271</b> |

# List of Figures

|      |                                                                                                                                                                           |     |
|------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 1.1  | Number of unpreconditioned GMRES iterations for the homogeneous Helmholtz equation . . . . .                                                                              | 18  |
| 1.2  | The increasing interpolation error if the mesh is not refined with increasing frequency. . . . .                                                                          | 19  |
| 1.3  | The pollution effect for the 1-d Helmholtz equation. . . . .                                                                                                              | 20  |
| 2.1  | An example of the sets in the definitions of the Helmholtz Exterior Dirichlet Problem and Truncated Exterior Dirichlet Problem. . . . .                                   | 29  |
| 2.2  | An example of an impenetrable obstacle with a cavity containing trapped rays. . . . .                                                                                     | 35  |
| 2.3  | A schematic of the expected behaviour of an $hk^a$ -accurate finite-element method. . . . .                                                                               | 46  |
| 2.4  | A schematic of the sets $D_{\text{int}}$ and $\tilde{D}_{\text{int}}$ from the proof of Theorem 2.51. . . . .                                                             | 70  |
| 2.5  | A schematic of the set $D_{\text{scat}}$ from the proof of Theorem 2.51. . . . .                                                                                          | 71  |
| 2.6  | A schematic of the set $D_{\text{trunc}}$ from the proof of Theorem 2.51. . . . .                                                                                         | 71  |
| 3.1  | The relationship between the different variational formulations of stochastic PDEs                                                                                        | 123 |
| 4.1  | Maximum GMRES iteration counts when $\ A^{(1)} - A^{(2)}\ _{L^\infty(D; \mathbb{R}^{d \times d})} = 0.5 \times k^{-\beta}$ for $\beta = 0, 0.1, 0.2, 0.3$ . . . . .       | 148 |
| 4.2  | Maximum GMRES iteration counts when $\ A^{(1)} - A^{(2)}\ _{L^\infty(D; \mathbb{R}^{d \times d})} = 0.5 \times k^{-\beta}$ for $\beta = 0.4, 0.5, 0.6, 0.7$ . . . . .     | 149 |
| 4.3  | Maximum GMRES iteration counts when $\ A^{(1)} - A^{(2)}\ _{L^\infty(D; \mathbb{R}^{d \times d})} = 0.5 \times k^{-\beta}$ for $\beta = 0.8, 0.9, 1$ . . . . .            | 150 |
| 4.4  | Maximum GMRES iteration counts when $\ n^{(1)} - n^{(2)}\ _{L^\infty(D; \mathbb{R})} = 0.5 \times k^{-\beta}$ for $\beta = 0, 0.1, 0.2, 0.3$ . . . . .                    | 151 |
| 4.5  | Maximum GMRES iteration counts when $\ n^{(1)} - n^{(2)}\ _{L^\infty(D; \mathbb{R})} = 0.5 \times k^{-\beta}$ for $\beta = 0.4, 0.5, 0.6, 0.7$ . . . . .                  | 152 |
| 4.6  | Maximum GMRES iteration counts when $\ n^{(1)} - n^{(2)}\ _{L^\infty(D; \mathbb{R})} = 0.5 \times k^{-\beta}$ for $\beta = 0.8, 0.9, 1$ . . . . .                         | 153 |
| 4.7  | GMRES iteration counts when $\ n^{(1)} - n^{(2)}\ _{L^q(D; \mathbb{R})} = 0.2 \times k^{-\beta}$ , for any $1 \leq q < \infty$ and $\beta = 0, 0.1, 0.2, 0.3$ . . . . .   | 169 |
| 4.8  | GMRES iteration counts when $\ n^{(1)} - n^{(2)}\ _{L^q(D; \mathbb{R})} = 0.2 \times k^{-\beta}$ , for any $1 \leq q < \infty$ and $\beta = 0.4, 0.5, 0.6, 0.7$ . . . . . | 170 |
| 4.9  | GMRES iteration counts when $\ n^{(1)} - n^{(2)}\ _{L^q(D; \mathbb{R})} = 0.2 \times k^{-\beta}$ , for any $1 \leq q < \infty$ and $\beta = 0.8, 0.9, 1$ . . . . .        | 171 |
| 4.10 | The computed Quasi-Monte-Carlo convergence rate for $Q(u) = \int_D u$ . . . . .                                                                                           | 187 |
| 4.11 | The computed Quasi-Monte-Carlo convergence rate for $Q(u) = u(\mathbf{0})$ . . . . .                                                                                      | 188 |

|      |                                                                                                                                               |     |
|------|-----------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 4.12 | The computed Quasi-Monte-Carlo convergence rate for $Q(u) = u(1, 1)$ . . . . .                                                                | 189 |
| 4.13 | The computed Quasi-Monte-Carlo convergence rate for $Q(u) = \nabla u(1, 1)$ . . . . .                                                         | 190 |
| 4.14 | The number of LU factorisations in the sequential nearby-preconditioning-QMC algorithm as a percentage of the total number of solves. . . . . | 191 |
| 4.15 | The empirical probability that GMRES applied to a nearby-preconditioned linear system converges in at most 12 iterations. . . . .             | 198 |
| E.1  | Quasi-Monte-Carlo error, for $Q = \int_D u$ and $k = 10$ . . . . .                                                                            | 241 |
| E.2  | Quasi-Monte-Carlo error, for $Q = \int_D u$ and $k = 20$ . . . . .                                                                            | 242 |
| E.3  | Quasi-Monte-Carlo error, for $Q = \int_D u$ and $k = 30$ . . . . .                                                                            | 243 |
| E.4  | Quasi-Monte-Carlo error, for $Q = \int_D u$ and $k = 40$ . . . . .                                                                            | 244 |
| E.5  | Quasi-Monte-Carlo error, for $Q = \int_D u$ and $k = 50$ . . . . .                                                                            | 245 |
| E.6  | Quasi-Monte-Carlo error, for $Q = \int_D u$ and $k = 60$ . . . . .                                                                            | 246 |
| E.7  | Quasi-Monte-Carlo error, for $Q = u(0)$ and $k = 10$ . . . . .                                                                                | 247 |
| E.8  | Quasi-Monte-Carlo error, for $Q = u(0)$ and $k = 20$ . . . . .                                                                                | 248 |
| E.9  | Quasi-Monte-Carlo error, for $Q = u(0)$ and $k = 30$ . . . . .                                                                                | 249 |
| E.10 | Quasi-Monte-Carlo error, for $Q = u(0)$ and $k = 40$ . . . . .                                                                                | 250 |
| E.11 | Quasi-Monte-Carlo error, for $Q = u(0)$ and $k = 50$ . . . . .                                                                                | 251 |
| E.12 | Quasi-Monte-Carlo error, for $Q = u(0)$ and $k = 60$ . . . . .                                                                                | 252 |
| E.13 | Quasi-Monte-Carlo error, for $Q = u((1, 1))$ and $k = 10$ . . . . .                                                                           | 253 |
| E.14 | Quasi-Monte-Carlo error, for $Q = u((1, 1))$ and $k = 20$ . . . . .                                                                           | 254 |
| E.15 | Quasi-Monte-Carlo error, for $Q = u((1, 1))$ and $k = 30$ . . . . .                                                                           | 255 |
| E.16 | Quasi-Monte-Carlo error, for $Q = u((1, 1))$ and $k = 40$ . . . . .                                                                           | 256 |
| E.17 | Quasi-Monte-Carlo error, for $Q = u((1, 1))$ and $k = 50$ . . . . .                                                                           | 257 |
| E.18 | Quasi-Monte-Carlo error, for $Q = u((1, 1))$ and $k = 60$ . . . . .                                                                           | 258 |
| E.19 | Quasi-Monte-Carlo error, for $Q = \nabla u((1, 1))$ and $k = 10$ . . . . .                                                                    | 259 |
| E.20 | Quasi-Monte-Carlo error, for $Q = \nabla u((1, 1))$ and $k = 20$ . . . . .                                                                    | 260 |
| E.21 | Quasi-Monte-Carlo error, for $Q = \nabla u((1, 1))$ and $k = 30$ . . . . .                                                                    | 261 |
| E.22 | Quasi-Monte-Carlo error, for $Q = \nabla u((1, 1))$ and $k = 40$ . . . . .                                                                    | 262 |
| E.23 | Quasi-Monte-Carlo error, for $Q = \nabla u((1, 1))$ and $k = 50$ . . . . .                                                                    | 263 |
| E.24 | Quasi-Monte-Carlo error, for $Q = \nabla u((1, 1))$ and $k = 60$ . . . . .                                                                    | 264 |
| F.1  | An upper bound on the number of GMRES iterations required for a nearby-preconditioned linear system. . . . .                                  | 267 |
| G.1  | A sample mesh, similar to those used in all the computations in this thesis. . . . .                                                          | 269 |

# List of Tables

|     |                                                                                                                                                                                                          |     |
|-----|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----|
| 1.1 | The number of degrees of freedom required to obtain various properties of finite-element approximations of the solution of the Helmholtz equation. . . . .                                               | 21  |
| 2.1 | All the results in the literature on $(hk^a, hk^b)$ -accuracy for $h$ -finite-element discretisations of the Helmholtz equation. . . . .                                                                 | 52  |
| 2.2 | All the results in the literature on $(hk^a, hk^b)$ -data-accuracy for $h$ -finite-element discretisations of the Helmholtz equation. . . . .                                                            | 53  |
| 2.3 | All the results in the literature on $hk^a$ -quasi-optimality for $h$ -finite-element discretisations of the Helmholtz equation. . . . .                                                                 | 54  |
| 2.4 | The constants from Section 2.4 . . . . .                                                                                                                                                                 | 102 |
| 4.1 | GMRES iteration counts when $\ n^{(1)} - n^{(2)}\ _{L^q(D, \mathbb{R})} = 0.2 \times k^{-\beta}$ , for any $1 \leq q < \infty$ and $\beta = 0, 0.1, \dots, 1$ . . . . .                                  | 172 |
| 4.2 | The value of $\alpha_0$ and $\alpha_1$ for different QoIs, where the QMC error $\approx CN_{\text{QMC}}^{-(\alpha_0 - \alpha_1 \ln(k))}$ . . . . .                                                       | 184 |
| 4.3 | The ideal and actual number of QMC points $N_{\text{QMC}}$ used in the numerical experiments summarised in Tables 4.4 and 4.5, chosen so that the QMC error is empirically bounded for all $k$ . . . . . | 185 |
| 4.4 | Results for the sequential nearby-preconditioning-Quasi-Monte-Carlo algorithm. . . . .                                                                                                                   | 192 |
| 4.5 | Results applying our parallel nearby-preconditioning-Quasi-Monte-Carlo algorithm with the target proportion of preconditioners as $(-0.04 + 0.02k)\%$ . . . . .                                          | 192 |
| 5.1 | Computational complexity of Monte-Carlo and Multi-Level Monte-Carlo algorithms . . . . .                                                                                                                 | 222 |



# List of Algorithms

- 4.1 The sequential nearby-preconditioning-Quasi-Monte-Carlo algorithm . . . . . 177
- 4.2 The main part of the parallel nearby-preconditioning-Quasi-Monte-Carlo algorithm. 180



# Introduction

## 1.1 THE SUBJECTS OF THE THESIS

The subjects of this thesis are rigorous theory and fast methods for the stochastic Helmholtz equation

$$\nabla \cdot (A \nabla u) + k^2 n u = -f, \quad (1.1)$$

where  $A$ ,  $n$ , and  $f$  are random fields (i.e., they are spatially heterogeneous and random). We are particularly interested in theory and methods that are applicable for large values of the wavenumber  $k$ , as the case of large  $k$  is of interest in applications, and theoretically and computationally demanding.

### 1.1.1 Motivation from Applications for the High-Frequency Stochastic Helmholtz Equation

The Helmholtz equation is the simplest possible model of wave propagation. Indeed, if one seeks time-harmonic solutions of the scalar wave equation

$$n \frac{\partial^2 U}{\partial t^2} - \nabla \cdot (A \nabla U) = F, \quad (1.2)$$

that is, solutions of the form  $U(t, \mathbf{x}) = e^{-ikt} u(\mathbf{x})$ , where  $F(t, \mathbf{x}) = e^{-ikt} f(\mathbf{x})$ , then the spatial part  $u$  satisfies (1.1); equivalently, (1.1) is the Fourier transform in time of (1.2). In certain scenarios, the time-harmonic Maxwell's equations reduce to (1.1), see, e.g., [152, Remark 2.1] for this derivation.

The physical motivation for studying (1.1) is, therefore, any physical scenario in which wave propagation can be modelled by either (1.2) or Maxwell's equations. One prominent example of the usage of (1.2) is in subsurface imaging, where the rock structures of the earth's crust are imaged by generating waves at the earth's surface, and recording the reflections of these waves from the rock structures. The waves within the earth's rock structures satisfy the elastic wave equation, but under the so-called acoustic approximation, this equation can be approximated by (1.2); see [118, Sections 1.1 and 1.2] for discussion of the acoustic and elastic wave equations, and the PhD thesis [43] for a detailed physical derivation of the elastic wave equation [43, Section 1.2], and a derivation and discussion of the acoustic approximation [43, Section 1.2.6]. Other physical scenarios involving waves modelled by either (1.2) or Maxwell's equations are the propagation of sound in an inviscid fluid [50, Section 2.1], and Microwave imaging (see, e.g., [27, Section 6.4]).

A mathematical and computational motivation for studying (1.1) is that many of the difficulties one encounters when studying and numerically solving more complex wave-propagation models, such as the elastic wave equation, are also encountered with (1.1). Therefore (1.1) is an appropriate

starting point for mathematical study of, and numerical algorithms for, wave propagation models.

We now consider three characteristics of the above examples that will drive the theoretical and computational work in this thesis: high effective frequency, heterogeneity, and stochasticity.

The real-life examples above can have high effective frequency, that is, the wavenumber  $k$  is large. The wavenumber may be large because either (i) the physical frequency is large, or (ii) the waves are low-frequency, but propagate over a large domain. Situation (i) arises in, e.g., non-destructive testing, where waves of frequency  $1 \times 10^6 - 1 \times 10^7$  Hz (see, e.g., [31]) are passed through materials to image their interior, and situation (ii) arises in, e.g., seismic imaging, where the wave frequencies are in the range 1-100 Hz (see, e.g., [193]) but the domain of interest is on the kilometre scale<sup>1</sup>. These low-frequency, large-domain problems have many wavelengths in the domain, and hence, when they are scaled to a domain of size  $\approx 1$ , they give rise to problems with large (effective) frequency.

Also, many of the examples above are modelled by the Helmholtz equation with *heterogeneous* coefficients. For example, in subsurface imaging of the Earth's crust, waves will pass through the sea, different types of rock, and materials contained within these rocks, such as water or oil. Each of these materials will have different properties, such as density and Lamé parameters, and therefore the coefficients  $A$  and  $n$  in (1.1) will be heterogeneous (see, e.g. [43, Section 1.2.4] for an explanation of how the density and Lamé parameters manifest themselves in (1.2)).

To understand the presence of stochasticity in the above examples, we consider two possible problems associated with (1.1):

1. The *forward problem*, where one knows the coefficients  $A$  and  $n$ , and wishes to find properties of the solution  $u$ , and
2. The *inverse problem*, where one knows properties of the solution  $u$  and wishes to find the coefficients  $A$  and  $n$ .

The Helmholtz equation with *random* coefficients arises when we model physical situations with uncertainty in the material parameters; this uncertainty can arise in both the forward and inverse problem. In the inverse problem, where one has sent an incident wave into an unknown medium, recorded the scattered wave, and reconstructs the medium, there will be uncertainties inherent in the process. For example, (i) the scattered wave will only be recorded at discrete points in space, rather than everywhere, and these recordings will be subject to measurement error, and (ii) the operator giving the properties of the solution  $u$  may inherently lose information (e.g., if it is the far-field operator, see the discussion in [50, pp. 37–38] of the ill-posedness of the inverse problem with the far-field operator). These uncertainties in the inference will result in uncertainties in the inferred properties of the medium. Alternatively, uncertainty arises in the forward problem, when we are already aware of uncertainty in our knowledge of the medium, and we wish to quantify the uncertainty in the wave passing through the uncertain medium. This occurs, for example, in radar imaging of ice sheets, where one wishes to know properties of the wave scattered by the ice, as in [126].

---

<sup>1</sup>E.g., the SEG Overthrust model [4], a common benchmark for seismic imaging applications, has domain size  $20\text{km} \times 20\text{km} \times 4.65\text{ km}$ .

This thesis will only focus on UQ for the forward problem. The forward problem and inverse problem share the common computational difficulty of needing to solve many (deterministic) realisations of (1.1). Whether the uncertainty in  $A$  and  $n$  has arisen as a result of the inverse or forward problem, most UQ algorithms will require many samples of the (random) solution of (1.1). As will be discussed below, obtaining one sample of the solution of (1.1) is a considerable computational task, and so obtaining many (and ‘many’ could easily mean thousands) of such samples is an even harder task. Reducing the computational cost of obtaining many samples of the solution of (1.1) will be a main focus of the algorithms developed and studied in this thesis.

### 1.1.2 Solving the Helmholtz Equation Numerically

We have just stated that it is hard to solve the (deterministic) Helmholtz equation

$$\nabla \cdot (A_{\text{det}} \nabla u_{\text{det}}) + k^2 n_{\text{det}} u_{\text{det}} = -f_{\text{det}}, \quad (1.3)$$

i.e., a single realisation of (1.1), numerically; we now provide some background on why this is the case. When solving (1.3) numerically we discretise it to obtain a linear system

$$Au = f. \quad (1.4)$$

*Issues from finite elements* We are exclusively concerned with discretisation via finite elements, see Chapter 2 for the details of such a discretisation. The linear systems (1.4) arising from standard finite-element discretisations of (1.3) are hard to solve, as the matrices  $A$  are:

1. non-Hermitian,
2. indefinite, and
3. large.

We will now briefly discuss each of these properties in turn, outlining why the matrices  $A$  have these properties, and how these properties affect the numerical solution of (1.4).

For 1, the matrices  $A$  are non-Hermitian because the underlying boundary-value problems are not self-adjoint (see [200, Section 4.2] for a discussion of the non-self-adjointness of exterior-boundary-value problems for the Helmholtz equation). This lack of self-adjointness means the sesquilinear forms arising in standard variational formulations of (1.1) are not conjugate symmetric, and this lack of conjugate symmetry is inherited by the discretisation matrices  $A$ . If one uses an iterative solver for the linear system (1.4), then the non-Hermitian nature of the matrices means that a solver that is suitable for such matrices, such as GMRES, must be used.

For 2, the matrices  $A$  are indefinite because the sesquilinear forms arising from standard variational formulations of (1.1) are not coercive. This indefiniteness means that GMRES applied to (1.4) may exhibit poor convergence properties, especially as the wavenumber  $k$  is increased—see Figure 1.1. In addition, the standard convergence theory for GMRES does not apply to indefinite systems, and so proving convergence results is also challenging.

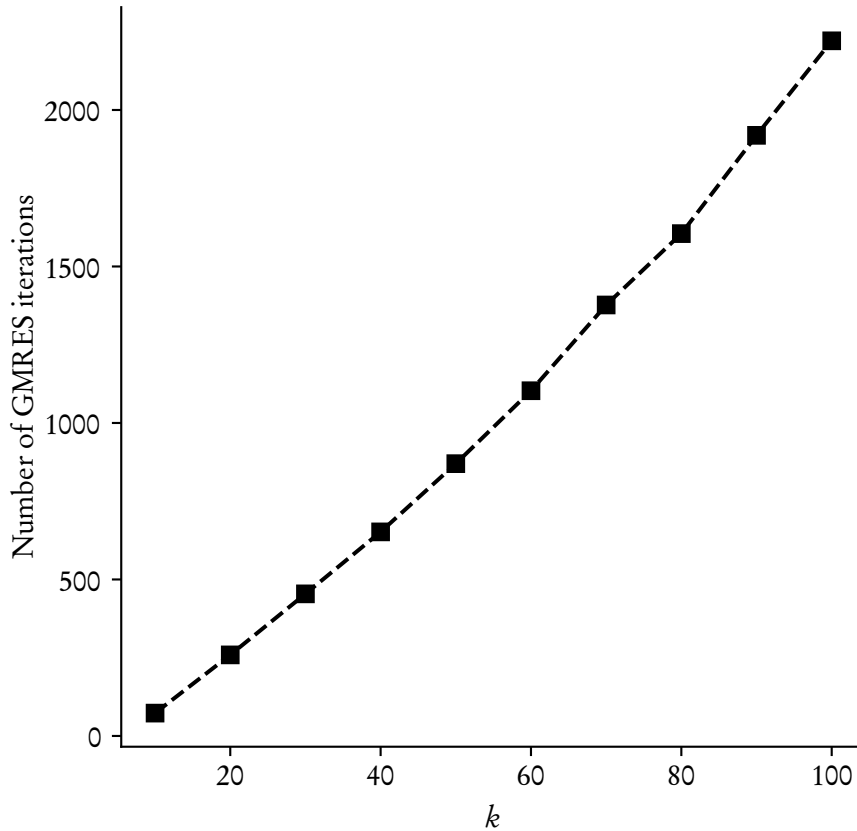


Figure 1.1: Number of unpreconditioned GMRES iterations for the homogeneous Helmholtz equation on the unit square, with zero impedance boundary condition and  $f = 1$ .

For 3, the matrices  $A$  are large because the number of degrees of freedom must increase as  $k$  increases. One can see this from interpolating the solution of (1.3); solutions  $u$  of (1.3) oscillate on a scale  $1/k$ , and therefore the number of degrees of freedom (interpolation points) must increase like  $k^d$  (where  $d$  is the spatial dimension) in order to keep the interpolation error for  $u$  bounded. This need for increasing degrees of freedom with  $k$  is illustrated in Figure 1.2, where we see the interpolation error grows if the number of degrees of freedom is not increased with  $k$ . In practice one typically chooses to use 6–10 discretisation points in each dimension for each wavelength in the domain—this choice empirically keeps the interpolation error at a reasonable size, but means the linear systems (1.4) will have  $\mathcal{O}(k^d)$  unknowns.

However, discretising (1.3) with a fixed number of points per wavelength is *not* enough to keep the error in the finite-element solution of (1.3) bounded as  $k \rightarrow \infty$  when using fixed-order methods. This is because standard-finite-element methods applied to the Helmholtz equation suffer from pollution, where the numerically calculated wave has a different wavelength to the true solution  $u$ , and so ‘drifts’ away from  $u$ ; moreover, this error increases as  $k$  increases. See Figure 1.3 for an illustration of this phenomenon, and Section 2.3.3 for an extended discussion of

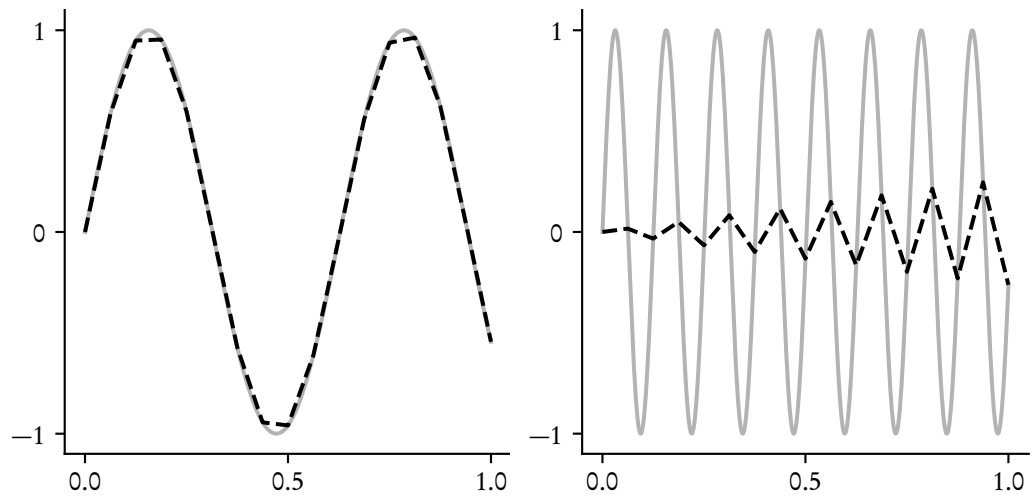


Figure 1.2: The increasing interpolation error if the mesh is not refined with increasing frequency. The left-hand plot show the interpolant of  $\sin(10x)$  on a mesh with 10 points per wavelength. The right-hand plot shows the interpolant of  $\sin(50x)$  on the same mesh.

how this phenomenon is reflected in the accuracy of finite-element methods for the Helmholtz equation.

In summary, numerically solving the Helmholtz equation gives rise to large (size at least  $\sim k^d$ ) linear systems, and the size of these linear systems increases as  $k$  increases. Table 1.1 presents the sizes of the linear systems one obtains for different values of  $k$ , depending on the spatial dimension, and the properties of the finite-element solution that one requires.

*Issues from linear solvers* We now turn our attention to how one might solve the large, indefinite, non-Hermitian linear systems (1.4). One option is to solve the linear systems (1.4) using a direct solver (solvers that, up to machine precision, invert the linear system (1.4) exactly, see, e.g., [62]). Such solvers are incredibly competitive for solving (1.3) in 2-D, if (1.4) has up to  $10^6$  unknowns; however, for larger linear systems (1.4), such as those obtained from 3-D discretisations, direct solvers are not as competitive as so-called iterative solvers, see, e.g., [68, p. 70]. An iterative solver is one that does not solve (1.4) exactly, but rather produces a sequence of approximations to the solution of (1.4). A standard iterative solver to use for non-Hermitian linear systems is GMRES; this is the solver we will use throughout this thesis. However, as seen in Figure 1.1, GMRES applied to (1.4) can perform very badly (the number of iterations to achieve convergence can grow dramatically with  $k$ , and moreover, one cannot apply the standard convergence results for GMRES (originally presented in [66] and given in a helpful form in [19, Section 1]) to (1.4) because the matrices  $A$  are typically indefinite. An explanation of how the wave-nature of the solution of the Helmholtz equation causes slow convergence of iterative methods for (1.4) is explained in [69, Section 2.1], using a finite-difference approximation of the Helmholtz equation as an example.

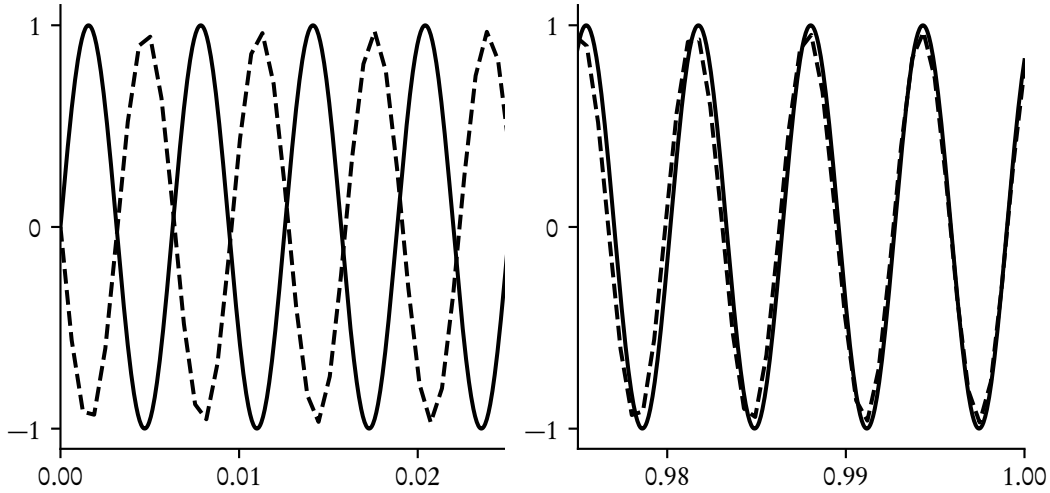


Figure 1.3: The pollution effect for the 1-d Helmholtz equation  $u'' + k^2 u = 0$  for  $k = 1000$ , with a zero Dirichlet boundary condition on the left endpoint, and an impedance boundary condition on the right endpoint, chosen so that the exact solution  $u = \sin(kx)$ . The solid line denotes the true solution, the dashed line denotes the finite-element approximation with 10 points per wavelength.

As GMRES applied to (1.4) performs badly, we consider preconditioning (1.4), that is, solving the equivalent linear system

$$(\mathbf{P})^{-1} \mathbf{A} \mathbf{u} = (\mathbf{P})^{-1} \mathbf{f} \quad (1.5)$$

for some matrix  $\mathbf{P}$ . The goal of preconditioning<sup>2</sup> is to choose the preconditioner  $\mathbf{P}$  such that:

1. The equation (1.5) is easy to solve iteratively, and
2. The action of  $(\mathbf{P})^{-1}$  is cheap to compute.

The ideal preconditioner from the point of view of Requirement 1 is  $\mathbf{P} = (\mathbf{A})^{-1}$ , however, in view of Requirement 2, if we could cheaply compute the action of  $(\mathbf{A})^{-1}$ , we could cheaply solve (1.4), and there would be no need for preconditioning. Hence, one needs to balance the Requirements 1 and 2 so that one obtains a good preconditioner for  $\mathbf{A}$  that is cheap to apply. There are several groups around the world working on the construction of good preconditioners for the Helmholtz equation, and this is an open research area. The design of such preconditioners is not the focus of this thesis, but we refer to, e.g., [85] for an overview of many recent preconditioners for the Helmholtz equation.

*Issues from UQ* On top of all the above issues in solving the deterministic Helmholtz equation (1.3), when seeking to perform UQ calculations for the stochastic Helmholtz equation (1.1) one

<sup>2</sup>We have only considered left-preconditioning, that is, multiplying  $\mathbf{A}$  from the left by  $(\mathbf{P})^{-1}$ . However, one can also consider right-preconditioning, that is, solving the linear system  $\mathbf{A}(\mathbf{P})^{-1} \tilde{\mathbf{u}} = \mathbf{f}$ , the solution  $\mathbf{u}$  is then given by  $\mathbf{u} = (\mathbf{P})^{-1} \tilde{\mathbf{u}}$ .

|      | Interpolation error bounded<br>( $h = \pi/5 \times k^{-1}$ ) |                         | Finite-element error bounded<br>( $h = k^{-3/2}$ ) |                            | Quasi-optimality<br>( $h = k^{-2}$ ) |           |
|------|--------------------------------------------------------------|-------------------------|----------------------------------------------------|----------------------------|--------------------------------------|-----------|
|      | 2-D                                                          | 3-D                     | 2-D                                                | 3-D                        | 2-D                                  | 3-D       |
| $k$  |                                                              |                         |                                                    |                            |                                      |           |
| 10   | $\approx 2.5 \times 10^2$                                    | $\approx 4 \times 10^3$ | $10^3$                                             | $\approx 3 \times 10^4$    | $10^4$                               | $10^6$    |
| 100  | $\approx 2.5 \times 10^4$                                    | $\approx 4 \times 10^6$ | $10^6$                                             | $10^9$                     | $10^8$                               | $10^{12}$ |
| 1000 | $\approx 2.5 \times 10^6$                                    | $\approx 4 \times 10^9$ | $10^9$                                             | $\approx 3 \times 10^{13}$ | $10^{12}$                            | $10^{18}$ |

Table 1.1: The number of degrees of freedom that would be required to obtain various properties of piecewise-linear finite-element approximations of the solution  $u$  of (1.3), for various values of  $k$ , in 2-D and 3-D. All errors etc. would be measured in the weighted  $H^1$  norm  $(\|\cdot\|_{H^1}^2 + k^2\|\cdot\|_{L^2}^2)^{1/2}$ . See Section 2.3 for discussion of why one chooses the mesh conditions used here.

often needs to solve many realisations of (1.1). E.g., if one wants to calculate  $\mathbb{E}[Q(u)]$ , where  $Q$  is some quantity of interest, then one can use the sample average of many realisations of  $u$ ;

$$\mathbb{E}[Q(u)] = \int_{\Omega} Q(u(\omega)) d\mathbb{P}(\omega) \approx \frac{1}{N} \sum_{j=1}^N Q(u(\omega^{(j)})),$$

where  $\omega \in \Omega$  denotes a random sample. Observe that to compute the sample average, one needs to solve many (which we emphasise again, could easily be thousands) different deterministic Helmholtz problems which, as has just been discussed, are each individually difficult to solve. This situation arises when using either sampling-based methods (such as Monte-Carlo methods) or interpolation-based methods (such as Stochastic-Collocation methods) to compute properties of the solution  $u$  of (1.1), or approximations to  $u$  itself. Rigorously studying (1.1), devising computational techniques to reduce the cost of such UQ calculations, and rigorously justifying this reduction, is the subject of this thesis.

## 1.2 THE MAIN ACHIEVEMENTS OF THE THESIS

The main achievements of the thesis are as follows:

1. New error bounds for the higher-order  $h$ -finite-element method applied to the (deterministic) Helmholtz equation in heterogeneous media; these bounds are explicit in their dependence on the wavenumber  $k$  and on the coefficient  $n$ .

These error bounds are the first for the Helmholtz equation in a heterogeneous medium; similar bounds have been proved for the Helmholtz equation in a homogeneous medium in, e.g., [216, 59, 44] and for a homogeneous medium with a small, frequency-dependent nonlinear perturbation in [217]. These bounds (and their explicit dependence on  $k$  and  $n$ ) are crucial for the analysis of the numerical methods developed in this thesis.

2. Well-posedness results and a priori bounds on the solution of the Helmholtz equation in random media, where the results and bounds obtained are frequency-independent. The arguments behind these results are written in a sufficiently general way that they can be used, in principle, to conclude similar results for a range of stochastic PDEs.

This work is an advance on the only previous work in the literature, [80], in which similar results and bounds were proved, but under restrictions that became more stringent as the frequency increased; in contrast the results in this thesis are frequency-independent. To prove such frequency-independent results, we will consider classes of random media that are almost-surely nontrapping. This nontrapping assumption will allow us to obtain our frequency-independent results. These results on well-posedness and a priori bounds are crucial for analysing the numerical methods that follow. Our results show that the problems we are solving are well-posed, and will enable us to rigorously prove properties of the numerical methods considered in this thesis. We expect that the arguments behind these results can be used in other cases where the bilinear form given by the PDE is indefinite, such as for the time-harmonic Maxwell's equations.

3. A computational strategy, which we call ‘nearby preconditioning’, that reduces the computational cost of solving many realisations of the Helmholtz equation in random media.
4. Numerical experiments that indicate that nearby preconditioning is, in practice, more effective than can be rigorously proved at present. We demonstrate the effectiveness of nearby preconditioning when applied to a Quasi-Monte-Carlo method for the Helmholtz equation.

The nearby preconditioning strategy seeks to reduce the computational cost of assembling preconditioners for many deterministic Helmholtz problems. This reduction is achieved by re-using a preconditioner from one deterministic Helmholtz problem for other Helmholtz problems, provided the coefficients are close in some metric (hence the term ‘nearby’). As far as we are aware, this is the first time such a strategy has been rigorously studied for the Helmholtz equation.

5. Theoretical analysis of the Monte-Carlo and Multi-Level Monte Carlo methods applied to the Helmholtz equation in random media.
6. Computational investigation into Quasi-Monte-Carlo methods applied to the Helmholtz equation in random media.

Multi-Level Monte-Carlo is a variance reduction technique that uses computations on a sequence of meshes to reduce the variance in UQ calculations, and therefore to reduce the number of realisations of the Helmholtz equation that need to be solved. We extend the existing abstract Multi-Level Monte-Carlo analysis in the literature to the case where the finite-element error is dependent on an additional parameter (here this parameter is the wavenumber  $k$ ), and then apply this abstract analysis to the Helmholtz equation, showing that we obtain theoretical speedup over Monte-Carlo. We also perform preliminary computations giving insight into the behaviour of QMC methods for the Helmholtz equation; these results are the first available in the literature.

### 1.3 THE STRUCTURE OF THE THESIS

In Chapter 2 we give background material on the (deterministic) Helmholtz equation and its discretisation via finite elements; this material will be necessary to understand the rest of this thesis. We present recent results in [105] concerning the well-posedness of the deterministic heterogeneous Helmholtz equation and a priori bounds on its solution of which the author of this thesis was a co-author. These results are used in some of the developments in this thesis, especially in Chapter 3. We also give an overview of the theory of finite-element discretisations of the Helmholtz equation, and prove the results in Achievement 1.

In Chapter 3 we prove the results in Achievement 2 for three formulations of the stochastic exterior Dirichlet problem (SEDP) for the Helmholtz equation in random media. These results are underpinned by the well-posedness results and a priori bounds obtained in [105] for the heterogeneous (but non-random) Helmholtz equation. Chapter 3 is a lightly-edited version of [176], accepted for publication in the SIAM/ASA Journal on Uncertainty Quantification.

In Chapter 4 we develop the nearby preconditioning strategy described in Achievement 3, prove the results on its effectiveness, and perform the numerical investigation of its effectiveness as in Achievement 4. We then provide preliminary numerical investigation into QMC methods for the Helmholtz equation, as in Achievement 6 before applying nearby preconditioning to a QMC method for the Helmholtz equation, obtaining speedup, as in Achievement 4.

In Chapter 5 we develop abstract Multi-Level Monte-Carlo theory as in Achievement 5, and then apply this to the Helmholtz equation, observing theoretical speedup compared to the Monte-Carlo method.



# PDE theory for the deterministic Helmholtz equation and the theory of its finite-element discretisation

## 2.1 INTRODUCTION

This chapter has two main foci:

1. Recapping theory for the deterministic Helmholtz equation in heterogeneous media, especially well-posedness results and a priori bounds on the solution, and
2. Recapping and extending theory of the finite-element method (FEM) for the deterministic heterogeneous Helmholtz equation, especially error bounds.

In Section 2.2.1 we define two Helmholtz problems, one on an infinite exterior domain, and the other on a truncated domain, and discuss their physical relevance. In Section 2.2.2 we recap in some detail the well-posedness results and a priori bounds from [105]; these results will be crucial for our analysis of stochastic Helmholtz problems in Chapter 3. Then in Section 2.2.3 we set these results in their wider context with a review of the related literature. We then move on to the FEM for these problems; in Section 2.3.1 we give the variational formulations of our two Helmholtz problems. In Section 2.3.2 we recap basic concepts of the FEM, and in Sections 2.3.3 and 2.3.4 we give an overview of the literature on error bounds and quasi-optimality for the Helmholtz equation, including an extended discussion of proof techniques for these results. Finally in Section 2.4 we prove new error bounds for the FEM for the heterogeneous Helmholtz equation.

The material from [105] presented in Section 2.2 is not presented as original work in this thesis, even though the author of this thesis is one of the authors of [105]; this material is given to set the scene for the original work in the rest of this thesis. We also note that the literature review in Section 2.2.3 is based on the literature reviews in in [40, Section 1.1] and [105, Sections 1 and 2.4]. The work in Sections 2.3 and 2.4 is, however, presented as original work.

## 2.2 WAVENUMBER- AND COEFFICIENT-EXPLICIT PDE THEORY OF THE DETERMINISTIC HELMHOLTZ EQUATION

We begin by defining the two deterministic Helmholtz problems that we consider in this thesis; we consider their stochastic counterparts in subsequent chapters.

### 2.2.1 Deterministic Helmholtz problems

We first state our problems of interest, largely following the presentation in [105]. For  $D \subseteq \mathbb{R}^d$ , we define matrix-value function spaces by letting  $L^\infty(D; \mathbb{R}^{d \times d})$  be the set of all measurable matrix-valued functions  $A : D \rightarrow \mathbb{R}^{d \times d}$  such that  $A_{i,j} \in L^\infty(D; \mathbb{R})$  for all  $i, j = 1, \dots, d$ . We define  $W^{1,\infty}(D; \mathbb{R}^{d \times d})$  and  $C^{0,1}(D; \mathbb{R}^{d \times d})$  (the spaces of componentwise  $W^{1,\infty}$  and Lipschitz functions respectively) analogously. We let SPD be the set of all symmetric-positive-definite matrices in  $\mathbb{R}^{d \times d}$ , and then define  $L^\infty(D; \text{SPD}) = \{A : D \rightarrow \text{SPD} : A \in L^\infty(D; \mathbb{R}^{d \times d})\}$ , and  $W^{1,\infty}(D; \text{SPD}) = \{A : D \rightarrow \text{SPD} : A \in W^{1,\infty}(D; \mathbb{R}^{d \times d})\}$ . Observe, however, that SPD is not a vector space, and therefore  $L^\infty(D; \text{SPD})$  and  $W^{1,\infty}(D; \text{SPD})$  are not vector spaces. For other function spaces, where the range of functions is  $\mathbb{C}$  we suppress the second argument, e.g. we write  $L^2(D)$  for  $L^2(D; \mathbb{C})$ .

**Problem 2.1** (Exterior Dirichlet Problem (EDP)). *Let  $D_-$  be a bounded Lipschitz open set such that the open complement  $D_+ := \mathbb{R}^d \setminus \overline{D_-}$  is connected and let  $\Gamma_D := \partial D_-$ . Let  $\gamma_D : H^1(D) \rightarrow H^{1/2}(\Gamma_D)$  denote the Dirichlet trace operator on  $\Gamma_D$ . Given*

- $k > 0$ ,
- $f \in L^2(D_+)$  with compact support,
- $g_D \in H^{1/2}(\Gamma_D)$ ,
- $n \in L^\infty(D_+; \mathbb{R})$  such that  $\text{supp}(1 - n)$  is compact in  $\mathbb{R}^d$  and there exist  $0 < n_{\min} < n_{\max} < \infty$  such that

$$n_{\min} \leq n(\mathbf{x}) < n_{\max} \text{ for almost every } \mathbf{x} \in D_+, \quad (2.1)$$

and

- $A \in L^\infty(D_+; \mathbb{R}^{d \times d})$  such that  $\text{supp}(I - A)$  is compact in  $\mathbb{R}^d$ ,  $A$  is symmetric, and there exist  $0 < A_{\min} < A_{\max} < \infty$  such that

$$A_{\min} |\xi|^2 \leq (A(\mathbf{x})\xi) \cdot \bar{\xi} < A_{\max} |\xi|^2 \text{ for all } \xi \in \mathbb{C}^d \text{ for almost every } \mathbf{x} \in D_+, \quad (2.2)$$

we say  $u \in H_{\text{loc}}^1(D_+)$  satisfies the exterior Dirichlet problem if

$$\nabla \cdot (A \nabla u) + k^2 n u = -f \text{ in } D_+, \quad (2.3)$$

$$\gamma_D u = g_D, \quad (2.4)$$

and  $u$  satisfies the Sommerfeld radiation condition

$$\frac{\partial u}{\partial r}(\mathbf{x}) - i k u(\mathbf{x}) = o\left(\frac{1}{r^{(d-1)/2}}\right) \text{ as } r := |\mathbf{x}| \rightarrow \infty, \text{ uniformly in } \hat{\mathbf{x}} := \mathbf{x}/|\mathbf{x}|. \quad (2.5)$$

As in [105, pp. 2874-2875], we note that (2.3) is understood in the sense that

$$\int_{D_+} (A \nabla u) \cdot \nabla \bar{\phi} - k^2 n u \bar{\phi} = \int_{D_+} f \bar{\phi} \text{ for all } \phi \in C_0^\infty(D_+),$$

and we can impose the radiation condition (2.5) on  $u$  because  $u$  is  $C^\infty$  outside some ball, by elliptic regularity.

One particular case of Problem 2.1 is the sound-soft scattering problem, where  $u$  is the acoustic pressure field resulting from the scattering of an incoming wave  $u_i$  by the scatterer<sup>1</sup>  $D_-$ , c.f., [118, Section 1.1] and [50, Section 1.1] In this problem, the total field  $u_T = u + u_i$  satisfies  $\Delta u_T + k^2 n u_T = 0$  in  $D_+$ , with  $\gamma_D u_T = 0$ . If there are no sources in the domain, and the incident field satisfies  $\Delta u_i + k^2 u_i = 0$  in  $D_+$ , then  $u$  satisfies Problem 2.1 with  $f = \nabla \cdot ((A - I)\nabla u_i) + k^2(n - 1)u_i$ , and  $g_D = -\gamma_D u_i$ .

Physically, the Sommerfeld radiation condition (2.5) ensures that the solutions of Problem 2.1 correspond to ‘outgoing’ waves (see, e.g., [118, Section 1.1.3]), and mathematically, it guarantees the uniqueness of the solution to Problem 2.1, see, e.g., [39, Corollary 2.9]. Observe that Problem 2.1 is defined on an infinite spatial domain; if one discretises Problem 2.1 using domain-based methods (such as FEMs) the infiniteness of the domain causes an issue. Therefore a common approach is to truncate Problem 2.1 with an artificial boundary that is sufficiently large to contain  $D_-$  and all the inhomogeneities in  $A$ ,  $n$ , and  $f$ .

If one was able to compute the Dirichlet-to-Neumann operator<sup>2</sup> for the *homogeneous* Helmholtz equation in the exterior of the artificial boundary, then one could discretise Problem 2.1 exactly. See Problem 2.10 for the variational formulation of Problem 2.1, which is posed on a finite domain and includes the exact Dirichlet-to-Neumann operator. In practice, however, the Dirichlet-to-Neumann operator is expensive to compute, and so is approximated with a different boundary condition on the truncated boundary. Options for the truncated boundary condition include a perfectly matched layer, first introduced in [22] for Maxwell’s equations, which mimics the whole of the external domain, or FEM-BEM coupling (a numerical method, where BEM stands for boundary-element method), as in, e.g., [116], where a boundary element method is used to approximate the solution in the exterior of the truncated domain. However, in this thesis, as a model problem we consider imposing an *impedance boundary condition*

$$\partial_\nu u - ik u = g_I \quad (2.6)$$

on the truncated boundary. If  $g_I = 0$ , then (2.6) can be seen as a first-order approximation to (2.5) (see, e.g., [82, p. 353], where it is shown that in certain asymptotic limits, the Dirichlet-to-Neumann map for the homogeneous Helmholtz equation is equal to multiplication by  $ik$ ). Moreover, we note that a common Helmholtz model problem in the numerical-analysis community is the *interior impedance problem (IIP)*, where an impedance boundary condition (2.6) is used, and it is assumed that  $D_- = \emptyset$ . Truncating Problem 2.1 with an impedance boundary condition gives rise to the following deterministic Helmholtz problem.

**Problem 2.2** (Truncated Exterior Dirichlet Problem (TEDP)). *Let  $D_-$  be an open bounded Lipschitz set such that the open complement  $D_+ := \mathbb{R}^d \setminus \overline{D_-}$  is connected. Let  $\tilde{D}$  be a bounded connected Lipschitz open set such that  $\overline{D_-} \subset \subset \tilde{D}$ . Let  $D := \tilde{D} \setminus \overline{D_-}$ ,  $\Gamma_D := \partial D_-$ , and  $\Gamma_I := \partial \tilde{D}$ . Let*

<sup>1</sup>In the literature the scattered field is sometimes denoted  $u_s$ , in which case  $u$  usually denotes the total field  $u_i + u_s$ .

<sup>2</sup>The operator  $T$  such that  $\partial_\nu u = T\gamma u$ .

$\gamma_I : H^1(D) \rightarrow H^{1/2}(\Gamma_I)$  denote the Dirichlet trace operator on  $\Gamma_I$ , and  $\partial_\nu : H^1(D) \rightarrow H^{-1/2}(\Gamma_I)$  the Neumann trace operator. Given

- $k > 0$ ,
- $f \in L^2(D)$ ,
- $g_D \in H^{1/2}(\Gamma_D)$ ,
- $g_I \in L^2(\Gamma_I)$ ,
- $n \in L^\infty(D; \mathbb{R})$  such that  $\text{supp}(1 - n)$  is compact in  $\mathbb{R}^d$ , satisfying (2.1) with  $D_+$  replaced by  $D$ , and
- $A \in L^\infty(D; \mathbb{R}^{d \times d})$  such that  $\text{supp}(I - A)$  is compact in  $\mathbb{R}^d$  and  $A$  is symmetric, satisfying (2.2) with  $D_+$  replaced by  $D$ ,

we say  $u \in H^1(D)$  satisfies the truncated exterior Dirichlet problem if

$$\nabla \cdot (A \nabla u) + k^2 n u = -f \text{ in } D, \quad (2.7)$$

$$\gamma_D u = g_D, \text{ on } \Gamma_D \text{ and}$$

$$\partial_\nu u - i k \gamma_I u = g_I \text{ on } \Gamma_I. \quad (2.8)$$

Observe that, by construction,  $\partial D = \Gamma_I \cup \Gamma_I$  and  $\Gamma_D \cap \Gamma_I = \emptyset$ .

Whilst the impedance boundary condition (2.8) is only an approximation to the Sommerfeld radiation condition (2.5), the solutions of Problem 2.2 are still ‘wave-like’, and we see below that the  $k$ -dependence of the solution operator of Problem 2.2 is the same as that of Problem 2.1.

### 2.2.2 Well-posedness and a priori bounds

We now recap the well-posedness results and a priori bounds for Problems 2.1 and 2.2 from [105]; these results will be crucial for proving well-posedness results and a priori bounds for the stochastic analogues of Problems 2.1 and 2.2 in Chapter 3. The novelty of the bounds in [105] is that the results are for all  $k$  and are explicit in  $A$ ,  $n$  and  $k$ ; this explicitness is necessary in order to prove similar a priori bounds for stochastic  $A$  and  $n$ . We prove these results under conditions on  $A$  and  $n$  that require  $A$  and  $n$  to be, in some sense, ‘nontrapping’. Informally, a medium is ‘nontrapping’ if all rays travelling through the medium escape in a uniform time; this definition, and the sense in which our conditions are ‘nontrapping’, is discussed in Section 2.2.3 below.

We first define the classes of  $A$  and  $n$  that we consider. We say, for  $A_0 \in \mathbb{R}^{d \times d}$  and  $\mu > 0$  that  $A_0 \geq \mu$  in the sense of quadratic forms if

$$\xi \cdot (A_0 \bar{\xi}) \geq \mu |\xi|^2 \text{ for all } \xi \in \mathbb{C}^d.$$

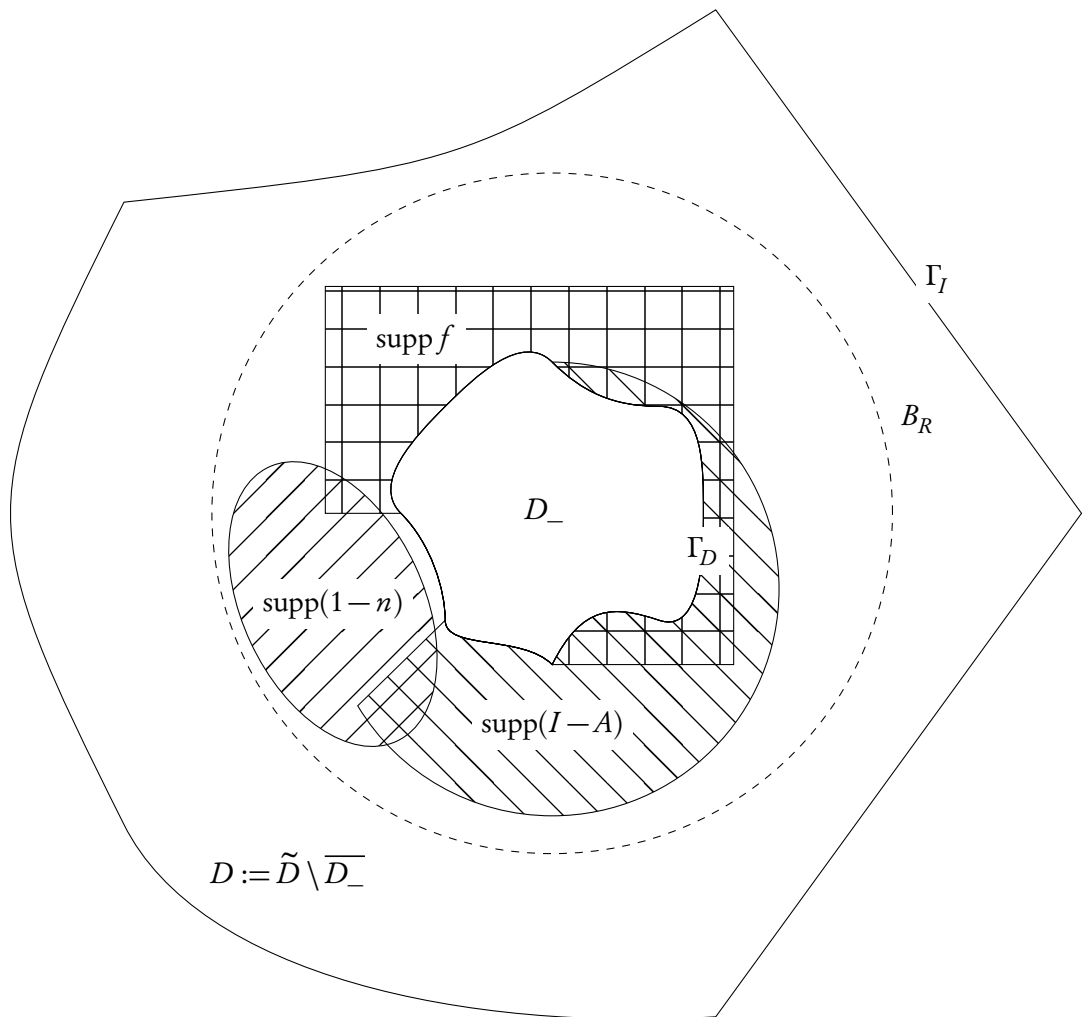


Figure 2.1: An example of the sets  $D_-$ ,  $\Gamma_D$ ,  $B_R$ ,  $\tilde{D}$ ,  $D$ , and  $\Gamma_I$  and  $\text{supp } f$ ,  $\text{supp}(I-A)$ , and  $\text{supp}(1-n)$  from Problems 2.1, 2.2, 2.10, and 2.12.

**Definition 2.3** (Class of nontrapping media). Let  $A \in C^{0,1}(\overline{D_+}; \mathbb{R}^{d \times d})$ ,  $n \in C^{0,1}(\overline{D_+}; \mathbb{R})$ , and  $\mu_1, \mu_2 > 0$ . We say that  $A \in \text{NT}_{\text{mat}, D_+}(\mu_1)$  if

$$A(\mathbf{x}) - (\mathbf{x} \cdot \nabla)A(\mathbf{x}) \geq \mu_1, \quad (2.9)$$

in the sense of quadratic forms, for almost every  $\mathbf{x} \in D_+$ . We say that  $n \in \text{NT}_{\text{scal}, D_+}(\mu_2)$  if

$$n(\mathbf{x}) + \mathbf{x} \cdot \nabla n(\mathbf{x}) \geq \mu_2 \quad (2.10)$$

for almost every  $\mathbf{x} \in D_+$ .

If  $D$  is as in Problem 2.2, then we define  $\text{NT}_{\text{mat}, D}(\mu_1)$  and  $\text{NT}_{\text{scal}, D}(\mu_2)$  analogously.

Definition 2.3 defines a sufficient, but not necessary, condition for a medium to be nontrapping; see [105, Section 7] for the connection between a related condition on  $A$  and  $n$  and nontrapping media.

**Remark 2.4** (Definition 2.3 makes sense). Both  $A$  and  $n$  are supported inside some bounded Lipschitz open set  $D$ , and on such sets,  $C^{0,1}(D) = W^{1,\infty}(D)$  (see, e.g., [74, Section 4.2.3, Theorem 5]). Since  $A$  and  $n$  are both Lipschitz functions (from Definition 2.3), it follows that they are in  $W^{1,\infty}(D; \mathbb{R}^{d \times d})$  and  $W^{1,\infty}(D; \mathbb{R})$  respectively. By construction  $A = I$  and  $n = 1$  outside  $D$ , and so it follows that  $A \in W^{1,\infty}(\overline{D_+}; \mathbb{R}^{d \times d})$  and  $n \in W^{1,\infty}(\overline{D_+}; \mathbb{R})$ , i.e.,  $A$  and  $n$  have weak first-order derivatives.

Our well-posedness results require the scatterer  $D_-$  to be star-shaped, and our results for Problem 2.2 require the truncation domain  $\tilde{D}$  to be star-shaped with respect to a ball. We now recall these definitions.

**Definition 2.5** (Star-shaped, star-shaped with respect to a ball). We say that  $D_-$  is star-shaped with respect to the point  $\mathbf{x}_0$  if for all  $\mathbf{x} \in D_-$ , the line segment  $[\mathbf{x}_0, \mathbf{x}] \in D_-$ .

We say that  $D_-$  is star-shaped with respect to the ball  $B$  if  $D_-$  is star shaped with respect to  $\mathbf{x}_0$ , for all  $\mathbf{x}_0 \in B$ .

We can now state well-posedness results and a priori bounds for the Helmholtz equation in the class of heterogeneous media we have just defined. We denote the ball of radius  $R$  about the point  $\mathbf{x}_0$  by  $B_R(\mathbf{x}_0)$ . We denote  $B_R(\mathbf{0})$  by  $B_R$ .

**Theorem 2.6** (Well-posedness and bound for Problem 2.1). If  $D_-, A, n$ , and  $f$  satisfy the requirements in Problem 2.1,  $g_D = 0$ ,  $D_-$  is star-shaped with respect to the origin, and there exist  $\mu_1, \mu_2 > 0$  such that  $A \in \text{NT}_{\text{mat}, D_+}(\mu_1)$  and  $n \in \text{NT}_{\text{scal}, D_+}(\mu_2)$ , then the solution of Problem 2.1 exists and is unique. Furthermore, given  $R > 0$  such that  $\text{supp}(I - A)$ ,  $\text{supp}(1 - n)$ , and  $\text{supp} f$  are compactly contained in  $D_R = D \cap B_R$ , then

$$\mu_1 \|\nabla u\|_{L^2(D_R)}^2 + \mu_2 k^2 \|u\|_{L^2(D_R)}^2 \leq C_1 \|f\|_{L^2(D_R)}^2,$$

for all  $k > 0$ , where

$$C_1 := 4 \left( \frac{R^2}{\mu_1} + \frac{1}{\mu_2} \left( R + \frac{d-1}{2k} \right)^2 \right).$$

For the proof of Theorem 2.6, see [105, Theorem 2.5].

The following result is the analogue of Theorem 2.6 for the solution of Problem 2.2. However, the statement is slightly more complicated than the statement of Theorem 2.6 due to the presence of the impedance boundary  $\Gamma_I$ . In particular, we have additional data  $g_I$  on  $\Gamma_I$ , and we bound the norm of  $u$  on  $\Gamma_I$  and well as on  $D$ .

**Theorem 2.7** (Well-posedness and bound for Problem 2.2). *If  $D_-, A, n, f$ , and  $g_I$  satisfy the requirements in Problem 2.2,  $g_D = 0$ ,  $D_-$  is star-shaped with respect to the origin,  $\tilde{D}$  is star-shaped with respect to a ball, and there exist  $\mu_1, \mu_2 > 0$  such that  $A \in \text{NT}_{\text{mat}, D}(\mu_1)$  and  $n \in \text{NT}_{\text{scal}, D}(\mu_2)$ , then the solution of Problem 2.2 exists and is unique. Let:*

- $L_I := \max_{\mathbf{x} \in \Gamma_I} |\mathbf{x}|$  and
- $aL_I$  be the radius of the ball with respect to which  $\tilde{D}$  is star-shaped.

Then

$$\begin{aligned} \mu_1 \|\nabla u\|_{L^2(D)}^2 + \mu_2 k^2 \|u\|_{L^2(D)}^2 + aL_I \|\nabla_{\Gamma_I} \gamma_I u\|_{L^2(\Gamma_I)}^2 + 2L_I k^2 \|\gamma_I u\|_{L^2(\Gamma_I)}^2 \\ \leq C_2 \|f\|_{L^2(D_R)}^2 + \tilde{C}_2 \|g_I\|_{L^2(\Gamma_I)}^2 \end{aligned} \quad (2.11)$$

for all  $k > 0$ , where  $\nabla_{\Gamma_I}$  is the surface gradient on  $\Gamma_I$ ,

$$\begin{aligned} C_2 &:= 4 \left( \frac{L_I^2}{\mu_1} + \frac{1}{\mu_2} \left( \beta + \frac{d-1}{2k} \right)^2 \right), \\ \tilde{C}_2 &:= 2 \left( 2 \left( 1 + \frac{2}{a} \right) + \frac{\beta}{L_I} + \frac{(d-1)^2}{4} \right) L_I, \end{aligned}$$

and

$$\beta := L_I \left( 2 + \frac{1}{(kL_I)^2} + 2 \left( 1 + \frac{2}{a} \right) \right).$$

For the proof of Theorem 2.7, see [105, Theorem A.6 (i)].

Observe that the above results are stated only in the case that  $g_D = 0$ . Whilst there is no mathematical difficulty in proving analogous results in the case  $g_D \neq 0$ , the calculations in this case are more involved, as one must consider the surface gradient on  $\Gamma_D$  and this surface gradient depends on  $A$ . In the case  $A = I$ , these calculations are significantly simplified, and so in the case  $A = I$  and  $g_D \neq 0$  results analogous to Theorems 2.6 and 2.7 are proved in [105, Theorem 2.19(ii)] (for Problem 2.1) and [105, Theorem A.6(iv)] (for Problem 2.2); although the proofs of these results require  $g_D \in H^1(\Gamma_D)$ .

We highlight that Theorems 2.6 and 2.7 and the similar results in [105] are significant for the following two reasons.

1. These are the first  $A$ ,  $n$ , and  $k$ -explicit bounds on the solution of the Helmholtz equation in the case where both  $A$  and  $n$  are heterogeneous. As will be discussed in more detail

in Section 2.2.3 below, previous results were either not  $A$ ,  $n$ , and  $k$ -explicit, or did not have  $A$  and  $n$  varying. The  $k$ -explicitness of these results is crucial for understanding how the solution of the Helmholtz equation (and numerical methods for its approximation) behave for large  $k$ ; the  $A$ -and- $n$ -explicitness is crucial for proving bounds on the stochastic Helmholtz equation, as in Chapter 3, and for understanding how numerical methods are affected by the heterogeneity in  $A$  and  $n$ .

2. These are the first bounds explicit in  $A$  and  $n$  where the bound *and* the restrictions on  $A$  and  $n$  are independent of  $k$ . Previous results in the literature only proved such bounds by imposing conditions on  $A$  and  $n$  that became *more stringent* as  $k \rightarrow \infty$ ; again, this literature will be more fully discussed in Section 2.2.3 below.

**Remark 2.8** (Extensions of Theorems 2.6 and 2.7). *Theorems 2.6 and 2.7 are extended to wider classes of heterogeneous  $A$  and  $n$  and to the case  $g_D \neq 0$  in [105]. As stated above, the case  $g_D \neq 0$  (with  $A = I$ ) is treated in [105, Theorem 2.19(ii)] (for Problem 2.1) and [105, Theorem A.6(iv)] (for Problem 2.2), and the case  $n = 1$  is covered in [105, Theorem 2.19(i)] (for Problem 2.1) and [105, Theorem A.6(ii)]. We highlight that when either  $A = I$  or  $n = 1$  (but not both) the condition on the non-constant coefficient can be slightly weakened from those in Definition 2.3. When  $A$  and  $n$  are discontinuous, [105, Condition 2.6] gives analogues of the conditions in Definition 2.3, and then the result corresponding to Theorem 2.6 is proved in [105, Theorem 2.7]. Letting  $A$  and  $n$  be  $L^\infty$ -perturbations of nontrapping media is discussed in [105, Remark 2.15], and relaxing the Lipschitz assumption on  $\Gamma_D$  is outlined in [105, Remark 2.13], with the caveat that when  $\Gamma_D$  is non-Lipschitz, we instead formulate Problem 2.1 as a variational problem, which is discussed in Section 2.3.1 below. The above extensions and generalisations can all be applied to Problem 2.2, as mentioned in [105, p. 2916].*

### 2.2.3 Discussion of results on well-posedness and a priori bounds for the Helmholtz equation

We now review the historical development of well-posedness results and a priori bounds for the Helmholtz equation.

#### Well-posedness results

By ‘well-posedness’, we mean that a solution of the problem under consideration exists, is unique, and continuously depends on the data ( $f$ ,  $g_D$ , and  $g_I$ ).

We note that proving well-posedness results and a priori bounds for the Helmholtz equation is much more involved than proving such results for the stationary diffusion equation

$$\nabla \cdot (A \nabla u) = -f \text{ in } D. \quad (2.12)$$

In (2.12) if  $A$  is bounded above and bounded away from zero, then the associated bilinear form is bounded and coercive. Then the Lax–Milgram Theorem applies, and one immediately obtains well-posedness and an a priori bound (in  $H^1(D)$ ) that is explicit in  $A$ .

However, for Helmholtz problems, the situation is much more subtle. Even if  $A$  and  $n$  are bounded above and bounded away from zero, in general one cannot prove a bound

$$\|u\|_{H_k^1(D_+)} \leq C \|f\|_{L^2(D_+)}, \quad (2.13)$$

where  $C$  depends explicitly on  $A$ ,  $n$ , and the wavenumber  $k$  and

$$\|v\|_{H_k^1(D)} := \left( \|\nabla v\|_{L^2(D)} + k^2 \|v\|_{L^2(D)} \right)^{1/2} \quad (2.14)$$

is the weighted  $H^1$  norm used frequently when studying Helmholtz problems<sup>3</sup>. A fundamental cause of this difficulty is the fact that the sesquilinear form  $a$  associated with the standard variational formulation of the Helmholtz equation is not coercive. However,  $a$  does satisfy a Gårding inequality

$$\Re a(v, v) + k^2 (A_{\min} + n_{\max}) \|v\|_{L^2(D)}^2 \geq A_{\min} \|v\|_{H_k^1(D)}^2, \quad (2.15)$$

where  $\Re$  denotes the real part. The Gårding inequality means  $\Re a(v, v)$  is ‘coercive’ if an appropriate multiple of the  $L^2$ -norm is added to it.

Because of the Gårding inequality, if the solution of the Helmholtz equation is unique, then existence and an a priori bound on the solution follow from Fredholm Theory (see, e.g. [200, Theorems 5.10 and 5.18]). Therefore the challenge of proving well-posedness reduces to proving uniqueness. However, we note that the a priori bound one obtains using Fredholm theory is generally *not* explicit in  $k$ ,  $A$  or  $n$ .

For homogeneous problems (with  $A = I$  and  $n = 1$ ) uniqueness follows from the Sommerfeld radiation condition (see, e.g., [39, Corollary 2.9]); for heterogeneous problems, the Unique Continuation Principle (UCP) gives uniqueness, under some additional smoothness assumptions on  $A$  and  $n$ . The UCP was first applied to Helmholtz problems by Melenk [147, Remark 8.1.1], following [137]; see, e.g., [137, Section 4.3], [105, p. 2871] for a discussion of the UCP and [99, Section 2] for a more detailed application of the UCP to show uniqueness for heterogeneous Helmholtz problems. Therefore, as well-posedness results for the Helmholtz equation are essentially well-understood<sup>4</sup>, we now turn our attention to a priori bounds on the solution that are explicit in  $k$ ,  $A$ , and  $n$ .

### $k$ -, $A$ -, and $n$ -explicit a priori bounds

All the bounds we now discuss will, unless otherwise stated, be for the weighted  $H^1$  norm  $\|\cdot\|_{H_k^1(D)}$  defined in (2.14). We only consider the case where the scatterer  $D_-$  is compact, and the

<sup>3</sup>The norm  $\|\cdot\|_{H_k^1(D)}$  is used because solutions of the Helmholtz equation typically have  $\nabla u \sim ku$ ; therefore the norm  $\|\cdot\|_{H_k^1(D)}$  should contain terms of roughly the same size. This relationship between the solution and its gradient is exactly the case for plane waves  $u = \exp(ik\mathbf{x} \cdot \mathbf{d})$  (for some  $\mathbf{d} \in \mathbb{R}^d$ ), where  $\nabla u = ik\mathbf{d}u$ .

<sup>4</sup>Observe that if one can prove an a-priori bound of the form (2.13), then one can conclude uniqueness (as the solution of the Helmholtz equation with zero data must therefore be the zero function). Therefore, if one can prove such a priori bounds *without* the restrictions on  $A$  and  $n$  needed to apply the UCP, one can conclude uniqueness (and well-posedness, as outlined above) in a wider class of media; see [105, pp. 2873, 2883] for more details on how the results in [105] can be used in this way.

inhomogeneities in  $A$  and  $n$  are compactly supported, as in Problem 2.1. Research into so-called rough surface scattering, where either  $D_-$  or the inhomogeneities in  $A$  and  $n$  are not compactly supported, is itself a rich area of research (see, e.g., the literature reviews in [205]), but this area is not the concern of this thesis. Throughout this section we use  $\lesssim$  notation— we say  $a \lesssim b$  if  $a \leq Cb$ , where  $C$  is independent of  $k$ . We define  $\gtrsim$  similarly, and say  $a \sim b$  if  $a \lesssim b$  and  $a \gtrsim b$ .

*Techniques for proving a priori bounds* There are two main classes of techniques for proving a priori bounds on the Helmholtz equation. The first class uses techniques from semiclassical analysis (a branch of microlocal analysis), and studies the behaviour of rays through the medium. For this approach to be used  $D_-$ ,  $A$ , and  $n$  must all be smooth, so that rays and the notion of reflections from the scatterer  $D_-$  can be defined (the notion of a reflection is difficult to define rigorously if the scatterer has a corner).

When one uses microlocal analysis tools, the key geometric condition on  $A$ ,  $n$ , and  $D_-$  is that of being nontrapping. The problem is *nontrapping* if, for any bounded set  $S \subseteq D_+$  there exists a time  $t(S)$  such that any ray starting in  $S$  and evolving according to the laws of geometrical optics leaves  $S$  by time  $t(S)$ . The rigorous definition is more technical; see [105, Section 6] for an overview. The problem is called *trapping* if it is not nontrapping. Once one has proved the problem is nontrapping, one combines the paramatrix argument of Vainberg [211] with the propagation of singularities results of Melrose and Sjöstrand [150] to conclude a bound with the same  $k$ -dependence as Theorem 2.6. Observe that trapping behaviour can be caused by an impenetrable obstacle (where, informally, rays ‘bounce’ off the obstacle), a penetrable obstacle (where rays penetrate and are then ‘trapped’ inside), or variations in the medium (that can also ‘trap’ rays).

We see below that one can prove  $k$ -independent bounds even when rays cannot be defined, typically when  $D_-$ ,  $A$ , and  $n$  are not smooth. In such situations, one usually uses the multiplier techniques discussed below. In an abuse of terminology, we call all situations where a bound holds with the same  $k$ -dependence as Theorem 2.6 ‘nontrapping’.

The second class of techniques is multiplier techniques, where the PDE (2.3) is multiplied by carefully chosen multiples of, e.g.,  $u$  and  $\mathbf{x} \cdot \nabla u$ , and the resulting expression is then integrated by parts and rearranged. Whilst conceptually simpler than semiclassical analysis tools, multiplier techniques allow one to prove bounds in situations that are inaccessible to semiclassical analysis, e.g., when the scatterer or coefficients are not smooth. However, multiplier techniques typically require more severe restrictions on the geometry of the scatterer (and truncation boundary, in the case of Problem 2.2) than semiclassical analysis techniques<sup>5</sup>. Multiplier methods were first used for wave problems by Morawetz in the 1960s for studying energy decay for the wave equation. See [84] for an overview of this, and other aspects of Morawetz’s work and [212, Theorem 1.1] for the connection between energy decay for the wave equation and a priori bounds on the Helmholtz equation.

---

<sup>5</sup>One can choose more complicated multipliers to mitigate some of these restrictions, as in [156], but most of the works we discuss below do not.

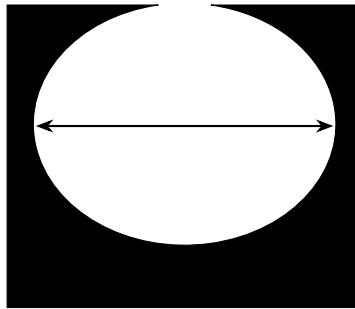


Figure 2.2: An example of an impenetrable obstacle with a cavity containing trapped rays.

*Situations in which to prove a priori bounds* We now summarise the state of the field regarding a priori bounds of the form (2.13), and especially the dependence of the constant  $C$  on  $k$ ,  $A$ ,  $n$ , and  $D_-$ . Some of these results are proved in the context of energy decay results for the time-domain wave equation; for simplicity's sake, we do not distinguish in our comments between these results, and those proving bounds (2.13) directly. This section borrows heavily from the literature reviews in [40, Section 1.1] and [105, Sections 1 and 2.4].

As noted above trapping behaviour can be caused either by the obstacle (either an impenetrable obstacle  $D_-$  or a penetrable obstacle, modelled by a jump in  $A$  and  $n$ ) or by variations in the medium, defined by heterogeneous  $A$  and  $n$ . For example, if an impenetrable obstacle  $D_-$  contains a cavity in which rays can be 'trapped', see, e.g., Figure 2.2, then trapping occurs. Similarly, if  $A$  and  $n$  jump, modelling a penetrable obstacle, the jumps can cause rays to be trapped in a manner analogous to the concept of total internal reflection.

This review will focus mainly on scattering induced by inhomogeneities in the medium, as this setting is the main concern of the results in Theorems 2.6 and 2.7 above (where the scatterer is assumed to be star-shaped). This is also the setting of the corresponding stochastic results in Chapter 3, where the medium is stochastic, and not the boundary of the scatterer. For an overview of results around impenetrable obstacle scattering, where  $C$  can grow logarithmically, polynomially, or exponentially in  $k$  depending on the scatterer, we refer the reader to the recent literature reviews in [40, Sections 1.1 and 1.3].

*The 'worst case' a priori bounds* In the worst case, when  $A$ ,  $n$ , and  $D_-$  are trapping, the constant  $C$  can depend exponentially on  $k$ ; i.e.,

$$C = C_1 \exp(kC_2), \quad (2.16)$$

for some constants  $C_1$  and  $C_2$  depending on  $D_-$ ,  $A$ , and  $n$ . This worst case bound was proved for a general impenetrable obstacle and  $A$  being  $C^\infty$  (with  $n = 1$ ) by Burq [33] and for a penetrable obstacle (defined by  $A$  and  $n$ , jumping across a shared  $C^\infty$  interface) by Bellasoued [21]. This worst case bound was proved for trapping by the medium when  $A = I$  and  $n$  is a Lipschitz perturbation of 1 by Shapiro [196].

Moreover, the bound (2.16) is sharp. This was shown to be sharp through a sequence of wavenumbers by Betcke, Chandler-Wilde, Graham, Langdon, and Lindner in [23, Equation 2.22].

For Problem 2.1 with constant media and an impenetrable scatterer whose boundary contains a certain part of an ellipse, they found a sequence  $(k_m)_{m \in \mathbb{N}}$  (with corresponding solutions  $u_m$  and right-hand sides  $f_m$  of (2.3)) such that

$$k \|u_m\|_{L^2(D_+)} \gtrsim \exp(\gamma k_m) \|f_m\|_{L^2(D_+)},$$

for some  $\gamma > 0$ . Similarly, Popov and Vodev [178] used semiclassical analysis techniques to prove the existence of a sequence  $k_m$  such that the growth in  $\|u_m\|_{L^2(D_+)}$  is superalgebraic for scattering by a penetrable obstacle (given by  $A = I$ , and  $n$  jumping downwards across a  $C^\infty$  interface).

However, the recent works [36, 152, 134] have shown that the exponential growth of  $C$  (2.16) is realised at very few frequencies. Moiola and Spence [152] provided numerical evidence (for the transmission problem through a sphere) that the realisation of super-algebraic growth is very sensitive to the value of  $k$ . More rigorously, Capdeboscq [36] obtained  $k$ -independent bounds for a penetrable circular obstacle in 2-d after excluding a small set of frequencies; and Lafontaine, Spence, and Wunsch [134] used microlocal analysis techniques to show that one can exclude a set of frequencies of arbitrarily small measure, and then obtain merely algebraic growth of  $C$ .

*'Nontrapping' a priori bounds* In contrast to the results above, in the best case the constant  $C$  in (2.13) has the same  $k$ -dependence as in Theorem 2.6, i.e.,  $C \sim 1$  for all  $k \geq k_0$ . These 'best case' results hold in a wide variety of settings, that we outline below, and in a slight abuse of terminology, we call all of these settings *nontrapping*. This is a slight abuse of terminology as for nonsmooth  $A$ ,  $n$ , and  $D_-$  we cannot always define 'nontrapping' in the sense of rays given above. In what follows, unless otherwise specified, the boundary conditions on impenetrable obstacles are Dirichlet boundary conditions.

In the full-space problem (i.e., Problem 2.1) when  $A$ ,  $n$ , and  $D_-$  are smooth, bounds with  $C \sim 1$  are proved with semiclassical analysis techniques by using (a) Melrose and Sjöstrand's results on propagation of singularities [150] combined with either (i) Vainberg's paramatrix argument from [211], or (ii) Lax-Phillips theory [136] or (b) Burq's defect measure argument [34]. An explicit value for  $C$  is given in [83].

For the full-space problem where  $A$  and  $n$  are not smooth, one typically uses multiplier techniques. These techniques were introduced by Morawetz and her collaborators in the 1960s and 1970s, who obtained bounds with  $C \sim 1$  for a variety of obstacle types [153, 155, 154, 156] in constant media. Multiplier techniques were also used to prove bounds with  $C \sim 1$  by Bloom and Kazarinoff [24, 25] and Perthame and Vega [177] when  $A$  or  $n$  are not compactly supported and not  $C^\infty$ , but decay sufficiently quickly at infinity, and possess sufficiently many derivatives for multiplier techniques to be used. Graham, the author of this thesis, and Spence also used multiplier techniques for a certain class of  $A$  and  $n$  possessing first-order derivatives (see Theorem 2.6 and [105]). In general whilst multiplier techniques do not work over the full range of nontrapping obstacles<sup>6</sup>, they allow one to conclude bounds when  $A$ ,  $n$ , and  $D_-$  are less than  $C^\infty$ .

<sup>6</sup>An exception is in [154], where the condition placed on the obstacle in [154, Equation (1.3)] is shown later in that paper (see [154, Equation (1.3a)]) to be equivalent to nontrapping in two dimensions.

For the full-space problem where  $A$  and  $n$  jump (i.e. are discontinuous), it was proved that  $C \sim 1$  for Problem 2.1 by Cardoso, Popov, and Vodev [37, 179] using semiclassical analysis techniques when  $A = I$  and  $n$  jumps up across a strictly convex, smooth boundary; and by Moiola and Spence [152] using multiplier techniques when both  $A$  and  $n$  jump, under assumptions on the jumps. Also, Theorem 2.6 and its related extensions discussed in Remark 2.8 prove  $C \sim 1$  for Problem 2.1 in a variety of situations, including when  $A$  and  $n$  jump.

For the truncated problem, i.e., Problem 2.2 or the IIP (Problem 2.2 with  $D_- = \emptyset$ ) a bound with  $C \sim 1$  is proved using semiclassical analysis techniques by Baskin, Spence, and Wunsch for the IIP with  $A = I$  and  $n = 1$  when  $\Gamma_I$  is  $C^\infty$  in [17]<sup>7</sup>. Bounds for the case  $A = I$  and  $n = 1$  (but with a scatterer  $D_-$  and/or less smoothness on  $\Gamma_I$ ) were proved by Melenk [147], Cummings and Feng [52], and Hetmaniuk [115]. Recently, similar bounds for the *PML problem*, that is, Problem 2.1 truncated with a perfectly matched layer have been obtained by Li and Wu [139] (for no obstacle) and Chaumont-Frelet, Gallistl, Nicaise, and Tomezyk (for a star-shaped impenetrable obstacle) [46].

For the truncated problem with variable, possibly jumping, media, multiplier techniques have been used to prove bounds with  $C \sim 1$  in a variety of recent work. Feng, Lin, and Lorton [80] proved a bound for random media (although the techniques in their proof are, in essence, for deterministic media), under the  $k$ -dependent assumption that  $A = I$  and  $n = 1 + \eta$ , with  $\eta$  a random field and  $\|\eta\|_{L^\infty(D)} \lesssim 1/k$  almost surely. Brown, Gallistl, and Peterseim proved a bound in [30], under conditions related to, but more restrictive than, those in [105]. Barucq, Chaumont-Frelet and Gout [16] proved a bound for 2-D piecewise-constant media, under a suitable condition on  $n$ . Graham and Sauter [99] took a very similar approach to [105], proving a bound for heterogeneous media when  $A = I$  under conditions on  $n$  that are analogous to those in [105]. In related results, for the 1-dimensional Helmholtz equation in heterogeneous media, Chaumont-Frelet [43, Section 2.1.5, Theorem 3] used multiplier methods with specially-chosen test functions to show a bound for piecewise constant media, under assumptions on the media that limit the number of ‘pieces’. Sauter and Torres [190] used properties of the 1-dimensional Green’s function to prove a bound for the 1-dimensional Helmholtz equation in piecewise-constant media with arbitrarily many ‘pieces’, and with  $C$  independent of  $k$ , but dependent on the number of pieces. Also, all of the results proved in [105] for Problem 2.1 have analogues for Problem 2.2.

Finally, we note that there is a small collection of work with  $k$  growing polynomially in  $k$ . For the IIP with general Lipschitz boundary, Spence [199] used bounds on layer potentials to show

$$\|\mathcal{U}\|_{H_k^1(D)} \lesssim k\|f\|_{L^2(D)} + k^{\frac{1}{2}}\|g_I\|_{L^2(\Gamma)},$$

building on work by Esterhazy and Melenk [73] and Feng and Sheen [76]. Ohlberger and Verfürth [163], studied the case where  $n = 1$ ,  $A$  is heterogeneous and scalar-valued, and the heterogeneity is given by many small inclusions. They proved a bound with  $C \sim k^3$  in this case. We suspect that both of these bounds are not sharp in their  $k$ -dependence, and that future work may improve the estimates of  $C$  in these cases.

<sup>7</sup>And these arguments can be generalised to certain classes of heterogeneous coefficients, see [17, Remark 5.6].

**Remark 2.9** (Bounds explicit in the parameters). *As stated previously, bounds on the heterogeneous Helmholtz equation that are explicit in all parameters of interest (such as  $k$ ,  $A$ , and  $n$ ) are crucial for proving  $k$ -explicit bounds on the corresponding stochastic Helmholtz equation; such bounds on the stochastic Helmholtz equation are the subject of Chapter 3. We observe in passing that of the works we described above, the only ones that have bounds explicit in all the parameters of interest are those of Moiola and Spence [152]; Galkowski, Spence, and Wunsch [83]; and Graham, the author of this thesis, and Spence [105].*

## 2.3 THE THEORY OF THE $h$ -FINITE-ELEMENT DISCRETISATION OF THE HELMHOLTZ EQUATION

We now shift our attention to the numerical analysis of the Helmholtz equation in heterogeneous media; in particular, we study the conforming finite-element method for the Helmholtz equation. We first state the variational formulations of the Helmholtz equation, define the finite-element method, and recall results on the approximation properties of finite-element spaces. We then prove our main result, error bounds for the finite-element method for the heterogeneous Helmholtz equation, where the bounds hold for arbitrary (fixed) degree finite elements, and are explicit in  $A$  and  $n$  in a sense made clear in Section 2.4 below.

### 2.3.1 Variational formulations for the Helmholtz equation

The finite-element method is based on the variational formulation of the Helmholtz equation; for simplicity of exposition, we state the variational formulation of Problems 2.1 and 2.2 in the case  $g_D = 0$ , although these can be generalised to the case  $g_D \neq 0$ .

**Problem 2.10** (Variational formulation of EDP when  $g_D = 0$ ). *Let  $D_+$ ,  $A$ ,  $n$ , and  $f$  be as in Problem 2.1. Choose  $R > 0$  such that  $\text{supp } f$ ,  $\text{supp}(I - A)$ ,  $\text{supp}(1 - n) \subset\subset B_R$ , and define  $D_R := D_+ \cap B_R$ .*

*We say  $u \in H_{0,D}^1(D_R)$  satisfies the variational formulation of the exterior Dirichlet problem with  $g_D = 0$  if*

$$a(u, v) = L(v) \quad \text{for all } v \in H_{0,D}^1(D_R),$$

where

$$a(w, v) := \int_{D_R} ((A\nabla w) \cdot \nabla \bar{v} - k^2 n w \bar{v}) - (T_R \gamma_R w, \gamma_R v)_{\Gamma_R} \quad (2.17)$$

and

$$L(v) := \int_{D_R} f \bar{v}, \quad (2.18)$$

where  $T_R : H^{1/2}(\Gamma_R) \rightarrow H^{-1/2}(\Gamma_R)$  is the Dirichlet-to-Neumann map for the homogeneous Helmholtz equation  $\Delta u + k^2 u = 0$  combined with the Sommerfeld radiation condition in the exterior of  $B_R$ ; and  $(\cdot, \cdot)_{\Gamma_R}$  is the duality pairing on  $\Gamma_R$ .

**Lemma 2.11** (Equivalence of formulations for the EDP). *Problems 2.1 and 2.10 are equivalent in the following sense. If  $u \in H_{\text{loc}}^1(D_+)$  solves Problem 2.1, then  $u|_{D_R} \in H_{0,D}^1(D_R)$  and  $u|_{D_R}$  solves*

*Problem 2.10* (for  $R$  as in Problem 2.10). Conversely, if  $u \in H_{0,D}^1(D_R)$  solves Problem 2.10, then  $u$  solves Problem 2.1, if  $u$  is extended to  $H_{\text{loc}}^1(D_+)$  by the solution of the exterior Dirichlet problem (in the exterior of  $D_R$ ) for the homogeneous Helmholtz equation with the Sommerfeld radiation condition (with Dirichlet data  $\gamma u$  on  $\partial B_R$ ).

For a proof of Lemma 2.11, see [105, Lemma 3.3].

**Problem 2.12** (Variational formulation of TEDP when  $g_D = 0$ ). Let  $D, A, n, f$ , and  $g_I$  be as in Problem 2.2. We say  $u \in H_{0,D}^1(D)$  satisfies the variational formulation of the truncated exterior Dirichlet problem with  $g_D = 0$  if

$$a_T(u, v) = L_T(v) \text{ for all } v \in H_{0,D}^1(D), \quad (2.19)$$

where

$$a_T(w, v) := \int_D ((A \nabla w) \cdot \nabla \bar{v} - k^2 n w \bar{v}) - ik(\gamma_I w, \gamma_I v)_{\Gamma_I} \quad (2.20)$$

and

$$L_T(v) := \int_D f \bar{v} + (g_I, \gamma_I v)_{\Gamma_I}$$

**Lemma 2.13** (Equivalence of formulations for the TEDP). *Problems 2.2 and 2.12 are equivalent, i.e.,  $u \in H_{0,D}^1(D)$  solves Problem 2.2 if, and only if,  $u$  solves Problem 2.12.*

For a proof of Lemma 2.13, see [105, Lemma A.7].

### 2.3.2 Background concepts in finite-element theory

We now give a brief summary of elementary concepts in finite-element theory. For brevity, we focus only on those concepts that we need to prove the new error bounds for finite-element discretisations of the Helmholtz equation in Theorem 2.39 below. We denote the spatial domain by  $D$ .

Throughout this thesis, we will assume that our finite-element space is defined on a conforming triangulation of simplices in the sense of Ciarlet [48, Paragraphs (FEM1) p. 61 and ( $\mathcal{T}_b$ 5) p. 71] which we now recall.

**Definition 2.14** (Conforming triangulation of simplices). *We say that  $\mathcal{T}$  is a conforming triangulation of simplices over  $D$  (or simply triangulation) if  $\bar{D}$  is subdivided into a finite number of simplices  $T \in \mathcal{T}$  such that*

1. For each  $T \in \mathcal{T}$ , the set  $T$  is closed and the interior  $\overset{\circ}{T}$  is nonempty and connected.
2. The equality

$$\bar{D} = \bigcup_{T \in \mathcal{T}} T$$

holds.

3. For each  $T_1 \neq T_2 \in \mathcal{T}$ , we have  $\overset{\circ}{T}_1 \cap \overset{\circ}{T}_2 = \emptyset$ .

4. For any  $T_1 \in \mathcal{T}$ , any face of  $T_1$  is either a subset of  $\partial D$  or a face of a different simplex  $T_2 \in \mathcal{T}$ .

We also define the mesh size of a triangulation.

**Definition 2.15** (Mesh size). *The mesh size of a triangulation  $\mathcal{T}$  is given by*

$$h := \max_{T \in \mathcal{T}} \text{diam } T.$$

We will frequently denote a triangulation with mesh size  $h$  by  $\mathcal{T}_h$ .

Having defined a triangulation, we are in a position to define the finite-element spaces that we will consider throughout this thesis.

**Definition 2.16** (Finite-element space). *Let  $\mathcal{T}_h$  be a triangulation of  $D$ , and let  $p \geq 1$  be an integer. We let  $V_{h,p}$  be the set of continuous piecewise-polynomials of degree  $p$  on  $\mathcal{T}_h$ , i.e.*

$$V_{h,p} := \{v_h \in C^0(D) : \text{for all } T \in \mathcal{T}_h, v_h|_T \text{ is a polynomial of degree at most } p\}.$$

**Remark 2.17** (Do  $\mathcal{T}_h$  and  $V_{h,p}$  exist?). *Observe that one can only construct a triangulation  $\mathcal{T}_h$  of  $D$  if  $D$  is a polyhedron (since  $\partial D$  must be the union of faces of simplices).*

*However, if  $D$  is not a polyhedron (for example, if  $D$  is Lipschitz, or  $D$  is smooth), then one cannot construct a triangulation  $\mathcal{T}_h$ , and therefore cannot construct  $V_{h,p}$ . The solution to this problem is to modify the elements near to the boundary, using, e.g., isoparametric finite elements. When using isoparametric finite elements, the reference element is mapped to elements near the boundary using finite-element functions of degree  $p$ , instead of using affine functions (as for standard finite elements). The result of this higher-order mapping is that one constructs an approximation of  $D$ , such that the distance from boundary of the approximation to  $\partial D$  is  $\mathcal{O}(h^{p+1})$  (see, e.g., [29, Section 4.7]). One can then construct a finite-element space on this ‘curved’ mesh, although this finite-element space will still be nonconforming (i.e.,  $V_{h,p} \not\subset H^1(D)$ ), and so one must analyse the resulting nonconforming error, as in, e.g., [29, Section 10.4].*

*However, in this thesis, we do not work with isoparametric finite elements, because:*

- *The analysis of isoparametric finite elements is standard (see, e.g.<sup>8</sup>, [29, Section 10.4]) and would complicate the presentation of our main results on the  $h$ - and  $k$ -dependence of the finite-element-error bounds. Moreover, a similar approach is taken in other literature on the  $h$ -finite-element method for the Helmholtz equation, see, e.g., [59, Top of p.785] and [135, Top of p.5].*
- *The only properties of  $V_{h,p}$  that we use are the existence of a best approximation (see Lemma 2.22 below) and the existence of an inverse inequality (see Lemma 2.71 below). Since one can prove an analogous best approximation result for isoparametric finite elements (see, e.g., [29, Theorem 4.7.3]), and we expect one can also prove an analogous inverse inequality (c.f. the proof of the standard inverse inequality in [29, Section 4.5] with the definition of isoparametric finite elements in [29, Section 4.7]), we expect that our original results in Section 2.4.1 below will also hold in the case of isoparametric finite elements.*

---

<sup>8</sup>Although [29, Section 10.4] works with the stationary diffusion equation, rather than the Helmholtz equation.

Therefore, throughout this thesis, we make the following assumption on our triangulations and finite-element spaces, for any spatial domain  $D$  arising in this thesis.

**Assumption 2.18** (Conforming triangulation and subspace). *We work with a family of triangulations  $(\mathcal{T}_h)_{h>0}$  and their corresponding finite-element spaces  $(V_{h,p})_{h>0}$ , where for any  $h > 0$ ,  $\mathcal{T}_h$  is a triangulation of  $D$  (in the sense of Definition 2.14), and  $V_{h,p}$  as defined in Definition 2.16 is a subspace of  $H^1(D)$ .*

Throughout this thesis, we only consider the  $h$ -finite-element method, i.e., the degree  $p$  of the polynomials associated with the space is assumed fixed, and we consider refining  $h$ . This is in contrast to the  $p$ -finite-element method, where  $h$  is fixed and  $p$  is increased, and the  $hp$ -finite-element method, where  $h$  is decreased and  $p$  is increased according to some rule. Throughout this section we use ‘finite-element method’ to refer to the  $h$ -finite-element method.

**Remark 2.19** (Why not consider the  $hp$ -finite element method?).  *$hp$ -finite-element methods for the homogeneous Helmholtz equation were analysed by Melenk and Sauter in [148, 149], who showed that the  $hp$ -finite-element method is quasi-optimal if  $kh/p$  is sufficiently small, and  $p \gtrsim 1 + \log(k)$  (i.e., we take  $p$  growing logarithmically with  $k$ , and a ( $p$ -dependent) fixed number of points per wavelength). Such methods can be very effective; the error for such methods can converge exponentially with respect to the number of degrees of freedom (see, e.g., [194, Theorem 4.51]). However, their analysis relies on a fully  $p$ -explicit analysis of the best approximation error of the Helmholtz equation in  $hp$ -finite-element spaces; such analysis is currently not available for the heterogeneous problem. Moreover, in practical applications one often works with fixed  $p$ , since implementing higher-order finite-elements can be challenging. Therefore it is still of interest to analyse the  $h$ -finite-element method.*

With the concept of a finite-element space established, we can now define the finite-element approximation to the variational problems Problems 2.10 and 2.12.

**Problem 2.20** (Finite-element approximation of Problem 2.12). *Find  $u_h \in V_{h,p}$  such that*

$$a_T(u_h, v_h) = L_T(v_h) \text{ for all } v_h \in V_{h,p}. \quad (2.21)$$

We say that  $u_h$  is the *finite-element approximation of  $u$*  (the solution to Problem 2.12). The finite-element approximation of Problem 2.10 is defined analogously.

To state an approximation bound for  $V_{h,p}$ , we first need to define the notion of a shape-regular mesh.

**Definition 2.21** (Shape-regular). *For  $h > 0$  and  $\tau \in \mathcal{T}_h$ , let  $B_\tau$  be the largest ball contained in  $\tau$ . The mesh family  $(\mathcal{T}_h)_{h>0}$  is shape-regular if there exists  $\rho > 0$  such that for all  $h > 0$  and for all  $\tau \in \mathcal{T}_h$*

$$\text{diam } B_\tau \geq \rho \text{ diam } \tau.$$

The following lemma shows how the approximation properties of the space  $V_{h,p}$  depend on  $h$  and  $p$ :

**Lemma 2.22** (Existence of approximation). *Let the mesh family  $(\mathcal{T}_h)_{h>0}$  underlying  $(V_{h,p})_{h>0}$  be shape-regular. Let  $v \in H^m(D)$ ,  $m \geq 1$ , and  $p \in \{1, \dots, m\}$ . Then there exists a constant  $C_{\text{BA},m} > 0$  independent of  $v$  and  $\tilde{v}_{h,p} \in V_{h,p}$  such that*

$$\|v - \tilde{v}_{h,p}\|_{L^2(D)} \leq C_{\text{BA},m} h^s \|v\|_{H^s(D)}$$

and

$$\|v - \tilde{v}_{h,p}\|_{H^1(D)} \leq C_{\text{BA},m} h^{s-1} \|v\|_{H^s(D)}$$

for  $1 \leq s \leq \min\{p+1, m\}$ .

For a proof of Lemma 2.22 see, e.g., [29, Theorems 4.4.4 and 4.4.20 and Remark 4.4.27].

### 2.3.3 Discussion of the finite-element method for the Helmholtz equation

We now discuss three possible properties of the  $h$ -finite-element method for the Helmholtz equation that we wish to investigate:

- The relative error  $\frac{\|u - u_b\|_{H_k^1(D)}}{\|u\|_{H_k^1(D)}}$  is bounded,
- The error  $\|u - u_b\|_{H_k^1(D)}$  is bounded in terms of norms of the data  $f$  and  $g_I$ , and
- The finite-element method is quasi-optimal,  $\|u - u_b\|_{H_k^1(D)} \lesssim \inf_{v_b \in V_{h,p}} \|u - v_b\|_{H_k^1(D)}$ ,

with a hidden constant independent of  $k$  and  $h$ .

The key question we ask is: What mesh conditions lead to each of the three properties above? (I.e., how must one refine  $h$  as  $k$  increases in order for the finite-element method to satisfy each of the three properties above?)

We will now summarise the state of the literature regarding each of the three properties above. We will:

1. Define each of the three properties formally,
2. State the sharpest results known in the literature,
3. Give a complete overview of all results in the literature concerning the three properties, and
4. Give a detailed discussion of proof techniques for each of the three properties.

Regarding Items 2 and 3: there is not a complete overview of all of these three properties anywhere in the literature. Moreover, research interest in these three properties has recently increased (e.g., the recent papers [44, 46, 45, 139] are all concerned with one or more of these properties, and have all been completed within the last two years), and so we believe it will be helpful and timely to provide a complete overview of the state of the literature. Regarding Item 4: although the different proof techniques are related, we do not know of anywhere in the literature

where these techniques are expounded, compared, and contrasted. We therefore hope that such a review will be valuable for the research community.

*Intuition for fixed number of points per wavelength* Before discussing the three properties listed above, we first give a brief discussion of the commonly-used heuristic ‘take a fixed number of points (or elements) per wavelength’. Recall from Section 1.1.2 that if one takes the mesh size in the finite element method  $h \sim 1/k$ , then one expects that the interpolation (or best approximation) error is bounded uniformly in  $k$ . More rigorously, as solutions of the Helmholtz equation typically<sup>9</sup> have  $\|u\|_{H^2(D)} \sim k$ , one can bound the  $H^1$ -norm of the interpolation error independently of  $k$  if  $h \sim 1/k$  using Lemma 2.22:

$$\left\| u - \tilde{u}_{h,p} \right\|_{H^1(D)} \lesssim h \|u\|_{H^2(D)} \lesssim hk \sim 1.$$

As explained in Section 1.1.2, this restriction ensures there are a fixed number of discretisation points per wavelength of the solution, since the wavelength  $\lambda = 2\pi/k$ .

An alternative motivation for taking  $h \sim 1/k$  is the Nyquist–Shannon sampling theorem (proved by Shannon in his seminal paper in information theory [195, Theorem 1]). The theorem states that any function  $v$  (in 1-d) whose Fourier transform lies inside  $[-k, k]$ , for some  $k > 0$ , is completely determined (via its Fourier series) by the point values  $v(0)$ ,  $v(\pm\mu)$ ,  $v(\pm 2\mu), \dots$ , for any  $\mu < 1/(2k)$ . Observe that  $k$  is then the highest frequency present in  $v$ . That is, if we interpolate  $v$  at points spaced less than  $\lambda/(4\pi) = 1/(2k)$  apart, where  $\lambda = 2\pi/k$  is the wavelength associated with oscillations of frequency  $k$ , we can reconstruct  $v$  from its Fourier transform. See, e.g., [9, §5.21] for an explicit formula for reconstructing  $v$ . Therefore, in the 1-d case we may reasonably expect that interpolating  $v$  with points spaced less than  $\lambda/(4\pi)$  apart will be a good approximation of  $v$ , as the Nyquist–Shannon theorem suggests such a spacing of points ‘captures’ all of the oscillations in  $v$ .

### Formal definition of the properties of the finite-element method above

These definitions may seem overly technical, however their technical definition is necessary to capture the, at times, complicated behaviour of finite-element methods for the Helmholtz equation. Whilst stating the definitions, we give some examples of what they mean for finite-element discretisations of the Helmholtz equation. These definitions are based on, and developed from, [55, Definition 2.3]. We define all of these properties for Problem 2.12 only, although one could define them completely analogously for Problem 2.10.

In the following definitions, we consider Problem 2.12, and its discretisation Problem 2.20, as parameterised by the wavenumber  $k$ , that is, we write  $u(k)$  for the solution of Problem 2.12 with wavenumber  $k$ , and  $u_h(k)$  for its finite-element approximation, the solution of Problem 2.20.

**Definition 2.23** ( $hk^a$ -quasi-optimal). *Given  $a > 0$ , we say that the  $h$ -finite-element method for the*

<sup>9</sup>If  $f$  and  $g_D$  are independent of  $k$ , and  $D$  is smooth enough, or a convex polygon, combining Theorem 2.6 and standard elliptic regularity results, gives the fact that  $\|u\|_{H^2(D)} \lesssim k$ , see, e.g., [86, Lemma 2.12].

Helmholtz equation is  $hk^a$ -quasi-optimal if, given  $k_0 > 0$ , there exist  $C_1(k_0), C_{\text{qo}}(k_0) > 0$  such that if

$$hk^a \leq C_1,$$

then the Galerkin solutions  $u_h(k)$  exist, are unique (for each  $k$ ), and satisfy

$$\|u(k) - u_h(k)\|_{H_k^1(D)} \leq C_{\text{qo}} \inf_{v_h \in V_{h,p}} \|u(k) - v_h\|_{H_k^1(D)},$$

for all  $k \geq k_0$ .

**Definition 2.24** ( $(hk^a, hk^b)$ -data-accurate). Given  $a, b > 0$ , we say that the  $h$ -finite-element method for Problem 2.12 is  $(hk^a, hk^b)$ -data-accurate if, given  $0 < \varepsilon < 1$ , and  $k_0 > 0$ , there exist  $C_1(k_0), C_2(\varepsilon, k_0) > 0$  such that if

$$hk^a \leq C_1 \tag{2.22}$$

and

$$hk^b \leq C_2, \tag{2.23}$$

then the Galerkin solutions  $u_h(k)$  exist, are unique (for each  $k$ ), and satisfy

$$\frac{\|u(k) - u_h(k)\|_{H_k^1(D)}}{\|f\|_{L^2(D)} + \|g_I\|_{H^{1/2}(\Gamma_I)}} \leq \varepsilon \quad \text{or} \quad \frac{\|u(k) - u_h(k)\|_{H_k^1(D)}}{\|f\|_{L^2(D)} + \|g_I\|_{L^2(\Gamma_I)}} \leq \varepsilon \tag{2.24}$$

for all  $k \geq k_0$ .

If  $a = b$  we say the  $h$ -finite-element method is  $hk^a$ -data-accurate.

To aid understanding of the above definition, we give the following illustrative example.

**Remark 2.25** (Example achieving  $(hk^a, hk^b)$ -data-accuracy). In [222, Corollary 4.2], Zhu and Wu proved that if  $u$  is the solution of Problem 2.12,  $u_h$  is the solution of Problem 2.20,  $k \geq k_0$ , and  $h^{p+1}k^{p+2} \leq C_1$ , then

$$\|u(k) - u_h(k)\|_{H_k^1(D)} \leq \tilde{C}_2 (h + h^p k^p + h^{2p} k^{2p+1}) (\|f\|_{L^2(D)} + \|g_I\|_{H^{1/2}(\Gamma_I)}) \tag{2.25}$$

for some  $k_0, C_1, \tilde{C}_2 > 0$ .

For  $0 < \varepsilon < 1$ , if we take

$$C_2 = \min \left\{ \frac{\varepsilon}{3\tilde{C}_2} k_0^{\frac{2p+1}{2p}}, \left( \frac{\varepsilon}{3\tilde{C}_2} k_0^{\frac{1}{2}} \right)^{\frac{1}{p}}, \left( \frac{\varepsilon}{3\tilde{C}_2} \right)^{\frac{1}{2p}} \right\} \tag{2.26}$$

then the  $h$ -finite-element method is  $(hk^{(p+2)/(p+1)}, hk^{(2p+1)/(2p)})$ -data-accurate.

When  $hk^{(2p+1)/2p} \leq C_2$ , the first term in the minimum (2.26) ensures  $h \leq \varepsilon / (3\tilde{C}_2)$ , the second term in (2.26) ensures  $h^p k^p \leq \varepsilon / (3\tilde{C}_2)$ , and the third term in (2.26) ensures  $h^{2p} k^{2p+1} \leq \varepsilon / (3\tilde{C}_2)$ . Therefore by (2.25) the bound (2.24) holds.

We choose this example to illustrate:

- Many results in the literature have  $a \neq b$ ; here  $a > b$ .
- Often (as in this case)  $a > b$ . I.e., whilst the mesh constraint ' $hk^b$  sufficiently small' makes the finite-element error small, the additional constraint ' $hk^a$  sufficiently small' is more restrictive.

**Definition 2.26** ( $(hk^a, hk^b)$ -accurate). Given  $a, b > 0$ , we say that the  $h$ -finite-element method for the Helmholtz equation is  $(hk^a, hk^b)$ -accurate if, given  $0 < \varepsilon < 1$ , and  $k_0 > 0$ , there exist  $C_1(k_0)$ ,  $C_2(\varepsilon, k_0) > 0$  such that if

$$hk^a \leq C_1 \tag{2.27}$$

and

$$hk^b \leq C_2, \tag{2.28}$$

then the Galerkin solutions  $u_b(k)$  exist, are unique (for each  $k$ ), and satisfy

$$\frac{\|u(k) - u_b(k)\|_{H_k^1(D)}}{\|u(k)\|_{H_k^1(D)}} \leq \varepsilon \tag{2.29}$$

for all  $k \geq k_0$ .

If  $a = b$  we say the  $h$ -finite-element method is  $hk^a$ -accurate.

**Remark 2.27** (Why have two mesh conditions in Definitions 2.24 and 2.26?). Frequently a more restrictive mesh condition is needed to show existence and uniqueness of a finite-element solution, but once one has shown existence and uniqueness, the finite-element error can be bounded under a less restrictive mesh condition. Therefore, as will be made clear in Section 2.3.4 below, the first mesh conditions (2.22) and (2.27) are needed to show existence and uniqueness of the finite-element solution, and the second mesh conditions (2.23) and (2.28) are needed to show that the finite-element error is bounded.

To give some intuition behind the definition of  $hk^a$ -accuracy, we refer to Figure 2.3, where we assume we have a  $hk^a$ -accurate finite-element method. Observe that if  $h$  is chosen to depend on  $k$  so that  $hk^a$  is constant, then (after a pre-asymptotic phase), the relative finite-element error is constant. However, if  $h$  is chosen so that  $hk^{a_2}$  is constant, where  $a < a_2$ , then the relative finite-element error *decreases* after a pre-asymptotic phase. This decrease is because the finite-element mesh is being over-refined, and therefore all the terms in a finite-element error bound such as (2.29) decrease with respect to  $k$ . Similarly, if  $h$  is chosen so that  $hk^{a_1}$  is constant, where  $a_1 < a$ , then the relative finite-element error *increases* after a pre-asymptotic phase, because the finite-element mesh is being under-refined<sup>10</sup>. We note that intuition for  $(hk^a, hk^b)$ -accuracy is more complex, since one must also consider the criterion (2.27) for the existence of the finite-element solution.

<sup>10</sup>Observe that Figure 2.3 does not tell the whole story, since under  $hk^a$ -accuracy, one also requires  $hk^a$  to be sufficiently small in order to guarantee that the finite-element solution  $u_b(k)$  exists. Therefore, if we only have that  $hk^{a_1}$  is sufficiently small, then we cannot guarantee the existence of  $u_b(k)$ . However, for the purposes of intuition, we ignore this detail here.

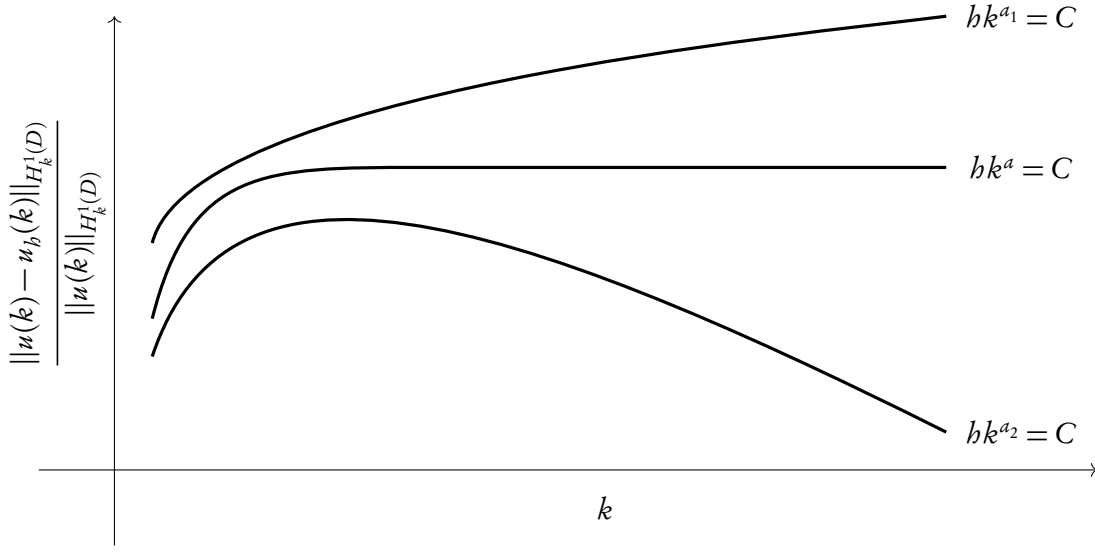


Figure 2.3: A schematic of the expected behaviour of an  $hk^a$ -accurate finite-element method, when  $hk^{a_1} = C$ ,  $hk^a = C$ , and  $hk^{a_2} = C$ , for  $C > 0$  chosen appropriately, where  $0 < a_1 < a < a_2$ .

**Remark 2.28** ( $hk^a$ -Quasi-optimality implies (better than)  $hk^a$ -data-accuracy). *Observe that, under standard assumptions on  $u$ , if the finite-element method is  $hk^a$ -quasi-optimal, then it is  $hk^a$ -data-accurate. We show this fact in the first order case, i.e., we take  $p = 1$  and  $a = 2$ . We assume  $\|u\|_{H^2(D)} \leq C_{H^2} \|f\|_{L^2(D)}$  for some  $k$ -independent constant  $C_{H^2} > 0$  (see, e.g., [86, Lemma 2.12] for a setting in which this bound holds), and refer to Table 2.3 below for references for the fact that the first-order finite-element method is usually  $hk^2$ -quasi-optimal.*

*We show, in fact, that the finite-element method is  $(hk^2, hk^1)$ -data-accurate. If  $hk^2 \leq C_1$  (where  $C_1$  is given in the definition of  $hk^2$ -quasi optimality), then*

$$\begin{aligned}
\|u - u_b\|_{H_k^1(D)} &\leq C_{\text{qo}} \|u - \tilde{u}_{b,p}\|_{H_k^1(D)} \text{ where } \tilde{u}_{b,p} \text{ is the approximation in Lemma 2.22,} \\
&\leq C_{\text{qo}} \left( \|u - \tilde{u}_{b,p}\|_{H^1(D)} + k \|u - \tilde{u}_{b,p}\|_{L^2(D)} \right) \\
&\leq C_{\text{qo}} C_{\text{BA},2} \left( h \|u\|_{H^2(D)} + h^2 k \|u\|_{H^2(D)} \right) \text{ by Lemma 2.22,} \\
&\leq C_{\text{qo}} C_{\text{BA},2} C_{H^2} (hk + h^2 k^2) \|f\|_{L^2(D)}. \tag{2.30}
\end{aligned}$$

Therefore, for  $\varepsilon \in (0, 1)$  we define

$$C_2(\varepsilon) = \frac{1}{C_{\text{qo}} C_{\text{BA},2} C_{H^2}} \min\{\varepsilon, \sqrt{\varepsilon}\}.$$

Hence if  $hk^2 \leq C_1$  and  $hk \leq C_2$ , then  $\|u - u_b\|_{H_k^1(D)} / \|f\|_{L^2(D)} \leq \varepsilon$ , i.e. the finite-element method is  $(hk^2, hk^1)$ -data-accurate (and therefore if  $k \geq k_0 > 0$  the finite-element method is  $hk^2$ -data-accurate, because  $hk \lesssim hk^2$ ).

Observe, however, that the finite-element method is actually *better than*  $hk^2$ -data-accurate, because the finite-element error *decreases* as  $k$  increases. With the above choices for  $C_1$ ,  $C_2$ , and  $h$ ,

from (2.30) we have

$$\|u - u_h\|_{H_k^1(D)} \leq C_{\text{qo}} C_{\text{BA},2} \left( \frac{C_1}{k} + \frac{C_1^2}{k^2} \right) \|f\|_{L^2(D)} \rightarrow 0 \quad \text{as } k \rightarrow \infty.$$

I.e., because we only need  $hk$  sufficiently small to bound the interpolation error, but have taken  $hk^2$  sufficiently small to ensure quasi-optimality, the finite-element error decreases as  $k \rightarrow \infty$ .

**Remark 2.29** (Relationship between  $(hk^a, hk^b)$ -accuracy and  $(hk^a, hk^b)$ -data-accuracy).

The only difference between the definitions of  $(hk^a, hk^b)$ -accuracy and  $(hk^a, hk^b)$ -data-accuracy is that in (2.29) the error is measured relative to the solution  $u$ , whereas in (2.24) the error is measured relative to the ‘data’  $f$  and  $g_I$ . However, if an a priori bound such as (2.11) holds, then  $(hk^a, hk^b)$ -accuracy implies  $(hk^a, hk^b)$ -data-accuracy, since  $\|u\|_{H_k^1(D)} \lesssim \|f\|_{L^2(D)} + \|g_I\|_{L^2(\Gamma_I)}$ .

On the other hand, to conclude  $(hk^a, hk^b)$ -accuracy from  $(hk^a, hk^b)$ -data-accuracy, one would need to show the physically realistic criterion

$$\|f\|_{L^2(D)} \lesssim \|u\|_{H_k^1(D)}. \quad (2.31)$$

However, in general (2.31) does not hold (as one can see by taking the unphysical solution  $u = e^{ik^2x} \chi$ , where  $\chi$  is a smooth cut-off function; in this case  $f = -(\Delta u + k^2 u) \sim k^4$ , but  $\|u\|_{H_k^1(D)} \sim k^2$ ). Therefore one cannot, in general, conclude  $(hk^a, hk^b)$ -accuracy from  $(hk^a, hk^b)$ -data-accuracy.

**Remark 2.30** (The above conditions for heterogeneous problems). For a heterogeneous problem, the constants  $C_1$ ,  $C_2$ , and  $C_{\text{qo}}$  in Definitions 2.23, 2.24, and 2.26 will all depend on the coefficients  $A$  and  $n$ . In general, this dependence is unknown. However, in Section 2.4 below we prove that the  $h$ -finite-element method with polynomial degree  $p$  is  $hk^{(2p+1)/(2p)}$ -data-accurate, where the dependence of  $C_1$  and  $C_2$  on  $n$  is completely known.

**Remark 2.31** (Generalisations of  $(hk^a, hk^b)$ -data-accuracy). For discretisations of other Helmholtz problems (e.g., full-space problems with no truncation boundary  $\Gamma_I$ , or problems truncated with a perfectly matched layer (PML)), then the denominator in (2.24) should be adapted appropriately, e.g., the denominator should equal  $\|f\|_{L^2(D)}$  for full-space problems and PML problems. In our discussion of the literature below (which includes such problems), we will make such an adaption without comment.

### Optimal mesh conditions for the finite-element method

We now provide a brief overview of the optimal values of  $a$  and  $b$  for  $(hk^a, hk^b)$ -accuracy,  $(hk^a, hk^b)$ -data-accuracy, and  $hk^a$ -quasi-optimality. By ‘optimal’, we mean the values of  $a$  and  $b$  that are smallest, corresponding to the least restrictive conditions on the mesh size  $h$ . We also, where possible, comment on whether these optimal conditions have been shown to be sharp. The literature reviews in this, and the next, section draw heavily on the literature reviews in [101, pp. 182–183] and [55, p. 112]. Unless otherwise stated, all the problems treated in this literature review are nontrapping (in the wider sense discussed under ‘Techniques for proving a priori bounds’ in Section 2.2.3), have constant coefficients, and have a impedance boundary condition on at least part of the boundary.

**Remark 2.32** (*hp*-methods for the Helmholtz equation). We briefly mention the mesh conditions one obtains for *hp*-finite-element methods, even though they are not the focus of this thesis. The landmark results on *hp*-finite-element methods for the Helmholtz equation were achieved by Melenk and Sauter [148, 149] who used a novel splitting of the solution of the Helmholtz equation (see the comments at the start of Section 2.4.2 below) to show that the *hp*-finite-element method is quasi-optimal if  $hk/p \leq c_1$  and  $p \geq c_2 \ln(k)$  (for some constants  $c_1, c_2 > 0$ ). Melenk and Sauter proved this result for the full-space problem in [148] and for: (i) the exterior Dirichlet problem, and (ii) the interior impedance problem in an analytic domain or a 2-d convex polygon in [149]. These results were generalised to an arbitrary 2-d Lipschitz polygon by Esterhazy and Melenk in [73, Theorem 4.2]. Other results in the literature for *hp*-finite-element methods are those of Zhu and Wu [222, Equation (1.7)] who showed the *hp*-finite-element method for the Helmholtz equation is data-accurate<sup>11</sup> provided

$$\frac{kh}{p} \leq C_0 \left( \frac{p}{k} \right)^{\frac{1}{p+1}},$$

where  $C_0 > 0$  is some constant.

( $hk^a, hk^b$ )-accuracy In 1-d the *h*-finite-element method has been proved to be  $hk^{3/2}$ -accurate for first-order finite elements by Ihlenburg and Babuška in [119, Theorem 5 and Equation (3.25)] and [118, Equation (4.5.15)]. For higher-order finite-elements (still in 1-d) they proved the *h*-finite-element method is  $hk^{(2p+1)/(2p)}$ -accurate in [121, Corollary 3.2] and [118, Theorem 4.27 and Equation 4.7.41]. However, all the above results were only measured in the  $H^1$  seminorm (and so we are slightly abusing the notation of  $hk^a$ -accuracy here), and the higher-order results were proved under the assumptions that  $f \in H^{p-1}(D)$ ,  $u \in H^{p+1}(D)$ , and  $|u|_{H^{p+1}(D)} \sim k^p |u|_{H^1(D)}$ .

These  $hk^a$ -accuracy results were confirmed numerically (for  $p = 1$ ) to be sharp by [119, Figure 11] and [118, Figure 4.13], which showed that the relative error is bounded if  $h \sim k^{-3/2}$  and by [118, Figure 4.10], which showed that the relative error is not bounded if  $h \sim k^{-1}$ .

These results were proved using the properties of the ‘discrete Green’s function’ for the Helmholtz equation. This function was used to first prove an a priori bound on the finite-element approximation of the solution, and the error bounds were then concluded from this a priori bound. Furthermore, the higher-order proofs in [121, 118] showed the finite-element error was bounded by  $|u|_{H^{p+1}(D)}$ , and then used the fact that

$$\frac{|u|_{H^{p+1}(D)}}{|u|_{H^1(D)}} \sim k^p \tag{2.32}$$

to bound the finite-element error by  $|u|_{H^1(D)}$ , and hence conclude a bound on the relative error. The relation (2.32) follows because the solution of the Helmholtz equation in 1-d is given by

$$A \cos(kx) + B \sin(kx) \tag{2.33}$$

<sup>11</sup>We do not use the ( $hk^a, hk^b$ )-accurate, etc., terminology for *hp*-methods, as it does not represent the interplay between the polynomial degree  $p$  and the wavenumber  $k$ .

(for some constants  $A$  and  $B$ ).

As the discrete Green's function is only known explicitly in 1-d, and because (2.33) only holds in 1-d (and therefore one can, in general, only prove (2.32) in 1-d), the proofs of  $hk^a$ -accuracy have not been extended to higher dimensions, although we conjecture they are true. The only computational results for higher dimensions are those by Bayliss, Goldstein, and Turkel, who observed in [18, Section 3, Tables 1–3] that, for low wavenumbers  $k \in (4.16, 8.32)$  the relative error for first-order finite-elements is bounded if  $h \sim k^{-3/2}$ , but is not bounded if  $h \sim k^{-1}$ .

*$(hk^a, hk^b)$ -data-accuracy* The best results known to date are the same (in terms of  $a$  and  $b$ ) as those for  $(hk^a, hk^b)$ -accuracy, except results for  $(hk^a, hk^b)$ -data-accuracy hold in higher dimensions. In [59, Theorem 5.1, Corollary 5.2] Du and Wu essentially proved that the  $h$ -finite-element method for the IIP is  $hk^{(2p+1)/(2p)}$ -data-accurate for arbitrary (fixed) polynomial degree  $p$  and  $d = 2$  or  $3$ , provided  $u \in H^{p+1}(D)$  (although this result can be shown for lower-regularity solutions by combining [59, Theorem 5.1] with [59, Lemma 3.5]). This result was proved for the IIP in  $d = 2$  or  $3$  for  $p = 1$  by Wu in [216]. We recall from Remark 2.29 that in physically realistic cases where (2.31) holds,  $(hk^a, hk^b)$ -data-accuracy implies  $(hk^a, hk^b)$ -accuracy.

*$hk^a$ -quasi-optimality* The best known result for  $hk^a$ -quasi-optimality for the Helmholtz equation is that the  $h$ -finite-element method is  $hk^{(p+1)/p}$ -quasi-optimal. This was first proved for  $p = 1$  in 1-d by Aziz, Kellogg, and Stevens in [6, Theorem 3.1] and Ihlenburg and Babuška in [119, Theorem 3] and [118, Theorems 4.9 and 4.13], with [119, Figures 7-9] and [118, Section 4.5.4 and Figures 4.11-4.12] showing this result is sharp in 1-d (although Ihlenburg and Babuška only work in the  $H^1$ -semi norm). The result was proved for  $p = 1$  and  $d = 2$  for the IIP by Melenk [147, Proposition 8.2.7], and for higher-order finite elements (for the full-space problem, IIP, and EDP) by Melenk and Sauter in [148, Corollary 5.6] and [149, Theorem 5.8] respectively. This result was shown for a PML problem by Chaumont-Frelet, Gallistl, Nicaise, and Tomezyk in [46, Theorem 5.1], and extended to a class of time-harmonic wave propagation problems by Chaumont-Frelet and Nicaise in [45, Theorem 2.15], who showed that if the constant in the a priori bound (2.13) grows like  $k^\alpha$ , then the  $h$ -finite-element method is  $hk^{(p+\alpha+1)/p}$ -accurate. They give numerical experiments in [45, Figure 8] showing the sharpness of  $hk^{(p+1)/p}$ -quasi-optimality for  $p = 1, 2$  for a heterogeneous IIP.

*Link with dispersion error* When computing on regular grids, one can mathematically analyse the 'dispersion error' of the finite-element method for the Helmholtz equation; i.e., the difference between the wavenumber of the true solution  $u$  (i.e.,  $k$ ) and the wavenumber of the approximation  $u_h$ . We mention briefly that Ainsworth analysed the dispersion error for the  $h$ - $p$ -finite-element method for the Helmholtz equation and proved the dispersion error is of the order  $h^{2p}k^{2p+1}$  [1, Equation (3.5)] (cited in [59, Remark 5.3(a)]). Observe that this is the same order as the best-available results for  $hk^a$ -accuracy and  $hk^a$ -data-accuracy discussed above. Therefore, results from finite-element error analysis and dispersion analysis both suggest that the error (measured in a suitable sense in each case) is bounded if  $h^{2p}k^{2p+1}$  is sufficiently small. See, e.g., [59, Remark

5.3(a)] for more references on dispersion error analysis for the Helmholtz equation.

**Remark 2.33** (Comparison with  $hp$ -finite-element methods). *Observe that the optimal results for higher-order finite elements become less stringent as  $p$  increases, i.e., the finite-element method is  $hk^{(2p+1)/(2p)}$ -data-accurate, and  $(2p+1)/(2p) \downarrow 1$  as  $p \rightarrow \infty$ . Therefore, in the  $p \rightarrow \infty$  limit, we recover the mesh condition ‘ $hk$  is sufficiently small’; that is, a fixed number of points per wavelength (recall the discussion on page 43 above).*

*Observe that the mesh condition ‘ $hk$  is sufficiently small’ implies that the number of degrees of freedom in the resulting linear systems is of the order  $k^d$ . This same scaling for the number of degrees of freedom is obtained for  $hp$ -methods for the Helmholtz equation, see [149, Remark 5.9]. Therefore, in the  $p \rightarrow \infty$  limit, the number of degrees of freedom required for the  $h$ -finite-element method scales optimally, as for the  $hp$ -finite-element method.*

### Complete summary of results in the literature

In Tables 2.1–2.3 we list all of the mathematical (as opposed to computational) results in the literature for  $(hk^a, hk^b)$ -accuracy,  $(hk^a, hk^b)$ -data-accuracy, and  $hk^a$ -quasi-optimality. We list these in chronological order, with any relevant comments in the ‘Notes’ column. The ‘Proof technique’ column details the method used in the proof; see Section 2.3.4 below for an extended discussion of these techniques. However, we now make a few general comments on the history of these results.

*Lack of coercivity* Recall that proving quasi-optimality (or an error bound) for the finite-element method for the Helmholtz equation is more difficult than for the stationary diffusion equation (2.12). For the stationary diffusion equation, one immediately obtains quasi-optimality for *any* mesh by Céa’s Lemma (and one then obtains that the relative error is bounded by Lemma 2.22). We emphasise again that this result holds for any shape-regular mesh, with no restriction on  $h$ .

However, Céa’s Lemma relies on the coercivity of the sesquilinear form, and the sesquilinear forms arising from standard discretisations of the Helmholtz equation are not coercive for large  $k$ . Therefore, to prove quasi-optimality, one instead uses the so-called Schatz argument, a modification of the standard Aubin–Nitsche duality argument<sup>12</sup> However, using the Schatz argument, one only obtains quasi-optimality under some  $k$ -dependent restriction on the mesh size  $h$ .

*Quasi-optimality using the Schatz argument* The Schatz argument was first used for problems satisfying a Gårding inequality by Schatz [192] and first used for Helmholtz problems by Aziz, Kellogg, and Stephens [6]. In [6] they proved that in 1-d the finite-element method for the Helmholtz equation is  $hk^2$ -quasi-optimal, and this result was extended to  $d = 2$  by Melenk [147] in his PhD thesis. However, the Schatz argument was first presented in the framework we use below by Sauter [188, Section 2]. Observe that for large values of  $k$ ,  $hk^2$ -quasi-optimality and the related mesh restriction ‘ $hk^2$  is sufficiently small’ is computationally prohibitive—it would result in linear systems of size, e.g.,  $10^{12}$ , for the Helmholtz equation with  $k = 100$  in 3-d.

<sup>12</sup>First introduced by Aubin [5] and Nitsche [159] for coercive problems.

*Error bounds using elliptic projections* Because of the severe mesh restrictions required for quasi-optimality for Helmholtz problems, recent research efforts have been focused on directly proving error bounds for the finite-element method. The key proof techniques are so-called elliptic projection ideas; these ideas are at the heart of our results in Section 2.4 below, and are discussed in more detail in Section 2.3.4 below. To our knowledge, the first use of elliptic projections to prove error estimates for Galerkin approximations was by Wheeler<sup>13</sup> [214, Theorem 3.1 ff.], who proved bounds on the error for a nonlinear parabolic problem by splitting the error into the error from an elliptic projection and the remaining error between the elliptic projection and the Galerkin approximation. The idea of using elliptic projections for Helmholtz problems was first introduced to prove error bounds for discontinuous Galerkin methods for Problem 2.2 by Feng and Wu [78, 79], and then used for standard finite-element methods (or closely-related continuous-interior-penalty methods) beginning with the work of Wu and Zhu [222, 216].

### 2.3.4 Extended discussion of proof techniques for finite-element errors for the Helmholtz equation

We now discuss in some detail the proof techniques used to obtain either  $(hk^a, hk^b)$ -accuracy,  $(hk^a, hk^b)$ -data-accuracy, or  $hk^a$ -quasi-optimality. We note that frequently quasi-optimality results are referred to as *asymptotic* error estimates, and accuracy or data-accuracy results (when they are proved under weaker mesh constraints than asymptotic estimates) are referred to as *pre-asymptotic* error estimates. This terminology is used because the mesh conditions to ensure quasi-optimality are more restrictive than those for bounded error, and therefore they hold only for smaller values of  $h$ .

For simplicity, our exposition below will assume that we are treating Problem 2.12 with homogeneous coefficients (i.e.,  $A = I$  and  $n = 1$ ) with  $g_I = 0$ , and that the problem is nontrapping<sup>14</sup>, and therefore in particular the solution  $u$  of Problem 2.12 is unique. Also, we suppress all of the constants involved, instead opting to use  $\lesssim$  notation, where  $a \lesssim b$  if  $a \leq Cb$ , with  $C$  independent of  $k$  and  $h$ . The new results we present in Theorem 2.39 consider heterogeneous problems that may be trapping, and state all of the constants involved explicitly, at the price of being more technical to state.

### Comparison of the different classes of argument

We first briefly discuss the positive and negative points for each of the two classes of argument we will outline below: (modified) duality arguments and error-splitting arguments.

The merits of duality arguments are their simplicity—we will state the arguments in their entirety in this overview section. Moreover, the mesh conditions and error bounds one obtains are completely  $h$ -,  $p$ -, and  $k$ -explicit. However, the main drawback of duality arguments (when used to prove data-accuracy) is their lack of sharpness in the conditions imposed on  $h$ , see, e.g., the discussion in Remark 2.25 above.

---

<sup>13</sup>Mentioned in [145].

<sup>14</sup>Recall that we say that the problem is nontrapping if  $C$  in (2.13) is bounded independent of  $k$ , for  $k \geq k_0$ .

|                                           | $(hk^a, hk^b)$ -accuracy       | Notes                                                                                                    | Proof technique                                                 |
|-------------------------------------------|--------------------------------|----------------------------------------------------------------------------------------------------------|-----------------------------------------------------------------|
| [119, Equation (3.25)]                    | $(hk^1, hk^{\frac{3}{2}})$     | $d = 1$ , unit interval,<br>$u(0) = 0$ ,<br>impedance boundary condition at 1<br>$H^1$ seminorm          | Discrete Green's function<br>(specific to $d = 1$ )             |
| [120, Theorem 4]                          | $(hk^1, hk^{\frac{3}{2}})$     | $d = 1$ , unit interval,<br>$u(0) = 0$ ,<br>impedance boundary condition at 1<br>$L^2$ norm <sup>1</sup> | Discrete Green's function<br>error splitting using interpolant  |
| [121, Corollary 3.2]                      | $(hk^1, hk^{\frac{2p+1}{2p}})$ | $d = 1$ , unit interval,<br>$u(0) = 0$ ,<br>impedance boundary condition at 1<br>$H^1$ seminorm          | Discrete Green's function,<br>error splitting using interpolant |
| [118, Theorem 4.13 and equation (4.5.15)] | $(hk^1, hk^{\frac{3}{2}})$     | $d = 1$ , unit interval,<br>$u(0) = 0$ ,<br>impedance boundary condition at 1<br>$H^1$ seminorm          | Discrete Green's function                                       |
| [118, Theorem 4.27 and equation (4.7.41)] | $(hk^1, hk^{\frac{2p+1}{2p}})$ | $d = 1$ , unit interval,<br>$u(0) = 0$ ,<br>impedance boundary condition at 1<br>$H^1$ seminorm          | Discrete Green's function,<br>error splitting using interpolant |

<sup>1</sup> Actually, [120, Theorem 4] only proves a bound on the  $L^2$ -norm of the error in terms of the  $H^2$ -seminorm of the solution. However, when  $d = 1$ ,  $|u|_{H^2(D)} \sim k^2 \|u\|_{L^2(D)}$  and so one can conclude  $(hk^a, hk^b)$ -accuracy.

Table 2.1: All the results in the literature on  $(hk^a, hk^b)$ -accuracy for  $h$ -finite-element discretisations of the Helmholtz equation.

|                                          | $(hk^a, hk^b)$ -data-accuracy                  | Notes                                                                                                                                                                  | Proof technique                                                |
|------------------------------------------|------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----------------------------------------------------------------|
| [58, Lemma 2.6]                          | $hk^{\frac{3}{2}}$                             | $d = 1$ , unit interval,<br>impedance boundary condition at both endpoints                                                                                             | Modified Schatz                                                |
| [119, Theorem 5]                         | $(hk^1, hk^{\frac{3}{2}})$                     | $d = 1$ , unit interval, $u(0) = 0$ ,<br>impedance boundary condition at 1                                                                                             | Discrete Green's function<br>error splitting using interpolant |
| [222, Corollary 4.2]                     | $(hk^{\frac{p+2}{p+1}}, hk^{\frac{2p+1}{2p}})$ | $d = 2, 3$ , IIP, $D$ smooth,<br>bounds obtained for $hp$ -finite-element method,<br>so fully $p$ -explicit                                                            | Modified Schatz                                                |
| [216, Theorem 5.1]                       | $hk^{3/2}$                                     | $d = 2, 3$ , IIP,<br>$D$ a star-shaped polygon/polyhedron                                                                                                              | Error splitting                                                |
| [59, Corollary 5.2]                      | $hk^{\frac{2p+1}{2p}}$                         | $d = 2, 3$ , IIP, $D$ smooth, star-shaped                                                                                                                              | Error splitting                                                |
| [44, Theorem 5.5]                        | $(hk^{\frac{3+\sigma}{1+\alpha}}, hk^{3/2})$   | $d = 2$ , TEDP with re-entrant corners,<br>a priori bound grows like $k^\sigma$ , $\alpha \in (1/2, 1)$<br>related to strength of corner singularities                 | Modified Schatz                                                |
| [139, Theorem 4.4<br>and Remark 4.5(iv)] | $hk^{3/2}$                                     | $d = 1, 2, 3$ , full-space problem truncated with PML,<br>posed in a ball                                                                                              | Modified Schatz                                                |
| [217, Lemma 3.3]                         | $hk^{3/2}$                                     | IIP, $D$ convex, $n$ heterogeneous with<br>$\ n - 1\ _{L^\infty(D; \mathbb{R})} \lesssim 1/k$ , part of an argument for a<br>nonlinear heterogeneous Helmholtz problem | Error splitting                                                |
| [46, Theorem 5.4]                        | $(hk^{\frac{p+2}{p+1}}, hk^{\frac{2p+1}{2p}})$ | $d = 2$ , EDP truncated with PML,<br>posed in a ball                                                                                                                   | Modified Schatz                                                |

Table 2.2: All the results in the literature on  $(hk^a, hk^b)$ -data-accuracy for  $h$ -finite-element discretisations of the Helmholtz equation.

|                                               | $hk^a$ -quasi-optimality    | Notes                                                                                                    | Proof technique                                                              |
|-----------------------------------------------|-----------------------------|----------------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------|
| [6, Theorem 3.1]                              | $hk^2$                      | $d = 1$                                                                                                  | Schatz                                                                       |
| [119, Theorem 3]                              | $hk^2$                      | $d = 1, H^1$ seminorm                                                                                    | Schatz                                                                       |
| [119, Corollary 2]                            | $hk^2$                      | $d = 1, H^1$ seminorm                                                                                    | Discrete Green's function,<br>error splitting using interpolant <sup>1</sup> |
| [118, Theorems 4.9<br>and 4.13]               | $hk^2$                      | $d = 1, H^1$ seminorm,<br>[118, Theorem 4.9] is [119, Theorem 3]                                         | Schatz and error splitting<br>using interpolant, respectively                |
| [147, Proposition 8.2.7]                      | $hk^2$                      | $d = 2$ , IIP, $D$ smooth and star-shaped or convex                                                      | Schatz                                                                       |
| [148, Corollary 5.6]                          | $hk^{\frac{p+1}{p}}$        | Full-space problem                                                                                       | Schatz                                                                       |
| [44, Theorem 5.3]                             | $hk^{2+\sigma}$             | $d = 2$ , TEDP with re-entrant corners<br>a priori bound grows like $k^\sigma$                           | Schatz                                                                       |
| [46, Theorem 5.1]                             | $hk^{\frac{p+1}{p}}$        | $d = 2$ , EDP truncated with PML, posed in a ball                                                        | Schatz                                                                       |
| [45, Theorem 2.15]                            | $hk^{\frac{p+\alpha+1}{p}}$ | Class of time-harmonic wave problems,<br>a priori bound grows at rate $k^\alpha$                         | Schatz                                                                       |
| [99, Theorems 4.2 and<br>4.5, Remark 4.6(ii)] | $hk^2$                      | IIP, $D$ Lipschitz, star-shaped w.r.t. a ball,<br>$n$ heterogenous, constants fully explicit in $n$      | Schatz                                                                       |
| [83, Theorem 3]                               | $hk^2$                      | EDP, $A$ and $n$ heterogeneous, $D_-, A$ , and $n C^\infty$ ,<br>constants fully explicit in $A$ and $n$ | Schatz                                                                       |

<sup>1</sup> Actually, [119, Corollary 2] proves quasi-optimality with constant proportional to  $k$  if  $hk$  is sufficiently small.

Table 2.3: All the results in the literature on  $hk^a$ -quasi-optimality for  $h$ -finite-element discretisations of the Helmholtz equation.

In contrast, the merit of error-splitting arguments is that they can give mesh conditions that are sharp in their  $h$ -dependence; in [59], Du and Wu proved that the finite-element method is  $hk^{(2p+1)/2p}$ -data-accurate, i.e., the mesh restriction needed for existence and uniqueness of  $u_b$  is of the same order as the mesh restriction needed to bound the finite-element error uniformly in  $k$ .

However, the drawback of error-splitting arguments is their complexity. They involve proving bounds on the solution of discrete Helmholtz problems; such bounds are complicated to prove, especially in the higher-order cases, and the constants involved depend on the polynomial degree  $p$  in highly complicated ways. These drawbacks limit the likelihood that such bounds can be used for  $hp$ -finite-element methods, where the dependence on the polynomial degree must be known explicitly.

### (Modified) duality arguments

These arguments are used to prove quasi-optimality and data-accuracy results; they include the more commonly known Schatz argument for Helmholtz problems. We first give the Schatz argument (for proving quasi-optimality), before going on to outline modified Schatz arguments (for proving data-accuracy).

Before we proceed we establish some notation that will enable us to discuss best approximation errors for solutions of Helmholtz problems. We let  $\mathcal{S} : L^2(D) \rightarrow H^1(D)$  denote the solution operator for Problem 2.12 with zero impedance boundary condition, and we let  $\mathcal{S}^\dagger$  denote the solution operator for the corresponding adjoint problem. That is, for any  $\tilde{f} \in L^2(D)$  and for all  $v \in H_{0,D}^1(D)$ ,

$$a_T(\mathcal{S}(\tilde{f}), v) = (\tilde{f}, v)_{L^2(D)}$$

and

$$a_T(v, \mathcal{S}^\dagger(\tilde{f})) = (v, \tilde{f})_{L^2(D)}.$$

We next define the approximability constants:

$$\eta := \sup_{\tilde{f} \in L^2(D)} \inf_{v_b \in V_{b,p}} \frac{\|\mathcal{S}(\tilde{f}) - v_b\|_{H_k^1(D)}}{\|\tilde{f}\|_{L^2(D)}} \quad (2.34)$$

and

$$\eta^\dagger := \sup_{\tilde{f} \in L^2(D)} \inf_{v_b \in V_{b,p}} \frac{\|\mathcal{S}^\dagger(\tilde{f}) - v_b\|_{H_k^1(D)}}{\|\tilde{f}\|_{L^2(D)}}.$$

Observe that the definitions of  $\eta$  and  $\eta^\dagger$  imply that for all  $\tilde{f} \in L^2(D)$

$$\inf_{v_b \in V_{b,p}} \|\mathcal{S}(\tilde{f}) - v_b\|_{H_k^1(D)} \leq \eta \|\tilde{f}\|_{L^2(D)} \quad (2.35)$$

and

$$\inf_{v_b \in V_{b,p}} \left\| \mathcal{S}^\dagger(\tilde{f}) - v_b \right\|_{H_k^1(D)} \leq \eta^\dagger \left\| \tilde{f} \right\|_{L^2(D)}. \quad (2.36)$$

We will use the bounds (2.35) and (2.36) in the Schatz and modified Schatz arguments below.

When dealing purely with the Schatz argument for quasi-optimality (as in [188, Section 2.2]) one only needs to consider the approximation of adjoint problems in the duality-argument step, and hence one only needs  $\eta^\dagger$ . However, in our exposition of elliptic-projection-based arguments below, we will also need to consider the approximation of standard Helmholtz problems, and hence we introduce notation for both  $\eta$  and  $\eta^\dagger$ .

*The Schatz argument for quasi-optimality* The first step is to use the Gårding inequality (2.15) satisfied by the Helmholtz equation to show that the error in the weighted  $H^1$  norm is bounded by the best approximation error, plus the error in the  $L^2$  norm<sup>15</sup>. We assume  $u_b$  exists, and observe that by the Gårding inequality (2.15) (with  $a_T$  as in (2.20))

$$\begin{aligned} \|u - u_b\|_{H_k^1(D)}^2 &\leq \Re a_T(u - u_b, u - u_b) + k^2 \|u - u_b\|_{L^2(D)}^2 \\ &= \Re a_T(u - u_b, u - v_b) + k^2 \|u - u_b\|_{L^2(D)}^2 \text{ by Galerkin orthogonality,} \\ &\lesssim \|u - u_b\|_{H_k^1(D)} \|u - v_b\|_{H_k^1(D)} + k^2 \|u - u_b\|_{L^2(D)}^2, \end{aligned} \quad (2.37)$$

for any  $v_b \in V_{b,p}$ .

We now use a modified version of the standard Aubin–Nitsche duality argument to bound  $\|u - u_b\|_{L^2(D)}$  by  $\eta^\dagger \|u - u_b\|_{H_k^1(D)}$  (and recall that  $\eta^\dagger$  can be made small). Let  $\xi \in H_{0,D}^1(D)$  solve the adjoint Helmholtz problem

$$a_T(v, \xi) = (v, u - u_b)_{L^2(D)} \text{ for all } v \in H_{0,D}^1(D). \quad (2.38)$$

Then, taking  $v = u - u_b$ , we have

$$\begin{aligned} \|u - u_b\|_{L^2(D)}^2 &= a_T(u - u_b, \xi) \\ &= a_T(u - u_b, \xi - v_b) \text{ by Galerkin orthogonality for } u - u_b, \text{ for any } v_b \in V_{b,p} \\ &\lesssim \|u - u_b\|_{H_k^1(D)} \|\xi - v_b\|_{H_k^1(D)} \\ &\lesssim \|u - u_b\|_{H_k^1(D)} \eta^\dagger \|u - u_b\|_{L^2(D)} \text{ by (2.36).} \end{aligned} \quad (2.39)$$

Cancelling a factor of  $\|u - u_b\|_{L^2(D)}$ , we obtain

$$\|u - u_b\|_{L^2(D)} \lesssim \eta^\dagger \|u - u_b\|_{H_k^1(D)}. \quad (2.40)$$

---

<sup>15</sup>Recall that for the stationary diffusion equation, one can show the finite-element error is bounded by the approximation error simply using coercivity and boundedness of the bilinear form—this is Cea’s Lemma.

Combining (2.37) and (2.40), we have

$$\|u - u_b\|_{H_k^1(D)}^2 \lesssim \|u - u_b\|_{H_k^1(D)} \|u - v_b\|_{H_k^1(D)} + (k\eta^\dagger)^2 \|u - u_b\|_{H_k^1(D)}^2,$$

and hence by cancelling a factor  $\|u - u_b\|_{H_k^1(D)}$  and taking the final term on to the left-hand side, we obtain

$$\|u - u_b\|_{H_k^1(D)} \lesssim \inf_{v_b \in V_{b,p}} \|u - v_b\|_{H_k^1(D)} \text{ if } k\eta^\dagger \text{ is sufficiently small.} \quad (2.41)$$

All results showing quasi-optimality for the Helmholtz equation (for different finite-element spaces and different domains) can then be seen as simply obtaining estimates on  $\eta^\dagger$  in these different scenarios; this literature is summarised in Table 2.3. For example, if  $p = 1$ , then one can show  $\eta^\dagger \sim hk$ , and hence we can conclude the first-order finite-element method is  $hk^2$ -quasi-optimal.

We have just shown quasi-optimality under the assumption that  $u_b$  exists. For proof of existence, we follow the proof of [200, Theorem 5.21]. Via the rank-nullity theorem (since (2.21) is equivalent to a finite-dimensional linear system) existence and uniqueness of  $u_b$  are equivalent. We now show uniqueness of  $u_b$ . By linearity, uniqueness of  $u_b$  is equivalent to showing uniqueness of the solution of

$$a_T(\tilde{u}_b, v_b) = 0 \text{ for all } v_b \in V_{b,p},$$

i.e., showing  $\tilde{u}_b = 0$ . Clearly  $\tilde{u}_b$  is a finite-element approximation of  $\tilde{u} \in H_{0,D}^1(D)$ , where  $\tilde{u}$  solves

$$a_T(\tilde{u}, v) = 0 \text{ for all } v_b \in H_{0,D}^1(D).$$

Since our problem is assumed nontrapping, it follows that  $\tilde{u}$  is unique, and hence  $\tilde{u} = 0$ . Therefore (2.41) becomes

$$\|\tilde{u}_b\|_{H_k^1(D)} \lesssim \inf_{v_b \in V_{b,p}} \|v_b\|_{H_k^1(D)} \text{ if } k\eta^\dagger \text{ is sufficiently small.} \quad (2.42)$$

The right-hand side of (2.42) is 0, and therefore  $\tilde{u}_b = 0$ . As outlined above, uniqueness of  $u_b$  follows and therefore  $u_b$  exists and is unique.

*Modified Schatz arguments for data-accuracy* The main difference between the Schatz argument and modified Schatz arguments is that modified Schatz arguments only prove a *bound* on the finite-element error, rather than proving quasi-optimality. Modified Schatz arguments (and error-splitting arguments, that we outline later) use the elliptic projection of a function. Therefore, we first outline the definition and some properties of elliptic projections.

The elliptic projection of a function  $w \in H_{0,D}^1(D)$  is the finite-element function  $\mathcal{P}_b w$  that has the same action as  $w$  on the finite-element space  $V_{b,p}$  with respect to a coercive and continuous

sesquilinear form<sup>16</sup>; i.e.,  $\mathcal{P}_h w$  is defined by

$$a_*(v_h, \mathcal{P}_h w) = a_*(v_h, w) \text{ for all } v_h \in V_{h,p},$$

for some coercive and continuous sesquilinear form  $a_*$ . For Problem 2.12, with sesquilinear form given by (2.20), choices for  $a_*(v_1, v_2)$  used in the literature are either

$$a_*(v_1, v_2) = (\nabla v_1, \nabla v_2)_{L^2(D)}, \quad (2.43)$$

$$a_*(v_1, v_2) = (\nabla v_1, \nabla v_2)_{L^2(D)} + (v_1, v_2)_{L^2(D)}, \quad (2.44)$$

or

$$a_*(v_1, v_2) = (\nabla v_1, \nabla v_2)_{L^2(D)} - ik(v_1, v_2)_{L^2(\Gamma_I)}. \quad (2.45)$$

These elliptic projections correspond to finding finite-element approximations of the solution of the PDEs

$$\Delta w = F \text{ in } D, \quad (2.46a)$$

$$w = 0 \text{ on } \Gamma_D, \text{ and} \quad (2.46b)$$

$$\partial_\nu w = 0 \text{ on } \Gamma_I; \quad (2.46c)$$

$$\Delta w + w = F \text{ in } D, \quad (2.47a)$$

$$w = 0 \text{ on } \Gamma_D, \text{ and} \quad (2.47b)$$

$$\partial_\nu w = 0 \text{ on } \Gamma_I; \quad (2.47c)$$

or

$$\Delta w = F \text{ in } D, \quad (2.48a)$$

$$w = 0 \text{ on } \Gamma_D, \text{ and} \quad (2.48b)$$

$$\partial_\nu w - ikw = 0 \text{ on } \Gamma_I \quad (2.48c)$$

respectively, where  $F$  is an appropriately chosen function. In the following exposition, we will assume  $a_*$  is given by (2.45), and to ease the exposition, we assume the solution of (2.48) is in  $H^2(D)$ . One can show in this case that the energy norm  $\|\cdot\|_*$  corresponding to (2.45) is equivalent

---

<sup>16</sup>The definition of the elliptic projection depends on the exact sesquilinear form used in the discretisation of the Helmholtz equation, and on the norm one is using to measure the error. For example, elliptic projection arguments for the Helmholtz equation originated in the study of discontinuous Galerkin methods for the Helmholtz equation (in [78, 79]); therefore the sesquilinear forms associated with the discretisation included penalty terms. These penalty terms were incorporated into the sesquilinear form  $a_*$  and the norms used to measure the error also included these penalty terms. In the following exposition, we will work with standard finite-element discretisations of the Helmholtz equation and standard Sobolev norms, and so the elliptic projections we used will be based on this setting.

to the weighted  $H^1$ -norm  $\|\cdot\|_{H_k^1(D)}$ , with equivalence constants independent of  $k$ , i.e.,

$$\|v\|_{\star} \sim \|v\|_{H_k^1(D)} \text{ for all } v \in H_{0,D}^1(D), \quad (2.49)$$

by the multiplicative trace inequality (Theorem 2.57) and the Poincaré–Friedrich’s inequality (Lemma 2.58). Since  $\mathcal{P}_b$  is a Galerkin projection, one can show that in its energy norm,  $a_{\star}$  is coercive and continuous<sup>17</sup>, and hence  $\mathcal{P}_b$  is quasi-optimal:

$$\|w - \mathcal{P}_b w\|_{\star} \lesssim \inf_{v_b \in V_{b,p}} \|w - v_b\|_{\star}. \quad (2.50)$$

Also, by the Aubin–Nitsche duality argument (since  $a_{\star}$  is coercive),

$$\|w - \mathcal{P}_b w\|_{L^2(D)} \lesssim b \|w - \mathcal{P}_b w\|_{\star}. \quad (2.51)$$

Combining (2.50) and (2.51) is a modification of the Schatz argument above. We first adopt the notation that, for  $v \in H_{0,D}^1(D)$ , the function  $\mathcal{I}_b v \in V_{b,p}$  is the function achieving the infimum in (2.35). We assume  $u_b$  exists, and start from (2.39), but instead of introducing an arbitrary  $v_b \in V_{b,p}$  into the second argument, we instead introduce the elliptic projection  $\mathcal{P}_b \xi$  by Galerkin orthogonality for  $u - u_b$ :

$$\begin{aligned} \|u - u_b\|_{L^2(D)}^2 &= a_T(u - u_b, \xi - \mathcal{P}_b \xi) \\ &= a_{\star}(u - u_b, \xi - \mathcal{P}_b \xi) - k^2(u - u_b, \xi - \mathcal{P}_b \xi)_{L^2(D)} \text{ by (2.45),} \\ &= a_{\star}(u - \mathcal{I}_b u, \xi - \mathcal{P}_b \xi) - k^2(u - u_b, \xi - \mathcal{P}_b \xi)_{L^2(D)} \\ &\quad \text{by Galerkin orthogonality for } \xi - \mathcal{P}_b \xi, \\ &\lesssim \|u - \mathcal{I}_b u\|_{\star} \|\xi - \mathcal{P}_b \xi\|_{\star} - k^2(u - u_b, \xi - \mathcal{P}_b \xi)_{L^2(D)} \\ &\lesssim \eta \|f\|_{L^2(D)} \|\xi - \mathcal{P}_b \xi\|_{\star} + k^2 \|u - u_b\|_{L^2(D)} \|\xi - \mathcal{P}_b \xi\|_{L^2(D)} \text{ by (2.49) and (2.35),} \\ &\lesssim \eta \|f\|_{L^2(D)} \eta^{\dagger} \|u - u_b\|_{L^2(D)} + k^2 \|u - u_b\|_{L^2(D)} \|\xi - \mathcal{P}_b \xi\|_{L^2(D)} \\ &\quad \text{by (2.50), (2.49), (2.38), and (2.36)} \\ &\lesssim \eta \eta^{\dagger} \|f\|_{L^2(D)} \|u - u_b\|_{L^2(D)} + k^2 \|u - u_b\|_{L^2(D)} b \eta \|u - u_b\|_{L^2(D)} \\ &\quad \text{by (2.51), (2.50), (2.49), and (2.35).} \end{aligned} \quad (2.52)$$

Therefore if  $b k^2 \eta$  is sufficiently small, the second term on the right-hand side of (2.52) can be absorbed into the left-hand side, and cancelling a factor of  $\|u - u_b\|_{L^2(D)}$ , we obtain

$$k \|u - u_b\|_{L^2(D)} \lesssim k \eta \eta^{\dagger} \|f\|_{L^2(D)} \text{ if } b k^2 \eta \text{ is sufficiently small.} \quad (2.53)$$

To obtain a bound on the error in the weighted  $H^1$  norm, we put (2.53) into (2.37) to get

$$\|u - u_b\|_{H_k^1(D)}^2 \lesssim \|u - u_b\|_{H_k^1(D)} \|u - v_b\|_{H_k^1(D)} + (k \eta \eta^{\dagger})^2 \|f\|_{L^2(D)}^2,$$

<sup>17</sup>For comment on showing coercivity and continuity for the other definitions of  $a_{\star}$ , see Remark 2.62 below.

and taking  $v_b = \mathcal{I}_b v$ , by (2.35) we have

$$\|u - u_b\|_{H_k^1(D)}^2 \lesssim \|u - u_b\|_{H_k^1(D)} \eta \|f\|_{L^2(D)} + (k\eta\eta^\dagger)^2 \|f\|_{L^2(D)}^2,$$

We then obtain for any  $\varepsilon > 0$  by Cauchy's inequality (2.85)

$$\|u - u_b\|_{H_k^1(D)}^2 \lesssim \varepsilon \|u - u_b\|_{H_k^1(D)}^2 + \frac{1}{\varepsilon} \eta^2 \|f\|_{L^2(D)}^2 + (k\eta\eta^\dagger)^2 \|f\|_{L^2(D)}^2. \quad (2.54)$$

Taking  $\varepsilon$  sufficiently small, moving the first term on the right-hand side of (2.54) to the left-hand side, and taking a square root, we obtain

$$\|u - u_b\|_{H_k^1(D)} \lesssim (\eta + k\eta\eta^\dagger) \|f\|_{L^2(D)} \text{ if } bk^2\eta \text{ is sufficiently small.} \quad (2.55)$$

Finally, we conclude existence and uniqueness of  $u_b$  using a similar argument to the one used for the Schatz argument above, except now we use the bound (2.55) with  $f = 0$  to conclude uniqueness of  $\tilde{u}_b$ .

By definition of  $\eta$  (2.35), the term  $\eta \|f\|_{L^2(D)}$  on the right-hand side of (2.55) is, up to a constant, the best-approximation error for  $u$  (i.e., the error when one interpolates/quasi-interpolates  $u$ ), and the term  $k\eta\eta^\dagger$  is the *pollution* term arising from the numerical method.

## Error-splitting arguments

The second class of arguments used in the literature are error-splitting arguments, used to prove accuracy and data-accuracy results. In these arguments the finite-element error is split using the elliptic projection of the solution  $u$ . To begin, we assume  $u_b$  exists and make the observation that

$$u - u_b = (u - \mathcal{P}_b u) + (\mathcal{P}_b u - u_b)$$

and therefore

$$\|u - u_b\|_{H_k^1(D)} \leq \|u - \mathcal{P}_b u\|_{H_k^1(D)} + \|\mathcal{P}_b u - u_b\|_{H_k^1(D)}.$$

An error bound for  $u - u_b$  can therefore be obtained by proving an error bound for the elliptic projection error  $u - \mathcal{P}_b u$  and proving a bound on the difference  $\mathcal{P}_b u - u_b$ . The former can either be accomplished by showing quasi-optimality of the elliptic projection (as above) or by proving such an error bound directly. The first approach is taken in [59, 46, 139], and the second approach is taken in [78, Lemma 5.2]<sup>18</sup>. In [78, 79, 216] the bound on the elliptic projection error is proved by observing that the sesquilinear form  $(\nabla v_1, \nabla v_2)_{L^2(D)}$  is coercive on  $H_{0,D}^1(D)$ , and then also controlling the additional term arising from the impedance boundary condition.

To bound the difference  $\mathcal{P}_b u - u_b$ , one first shows that it solves a deterministic Helmholtz problem: For any  $v_b \in V_{b,p}$  we have  $a_T(u - u_b, v_b) = 0$  and  $a_*(u - \mathcal{P}_b u, v_b) = 0$  by Galerkin

<sup>18</sup>Although the proof is only contained in [77, Lemma 5.2], [79, Lemma 4.3], and [216, Lemma 4.2].

orthogonality for  $u - u_b$  and  $u - \mathcal{P}_b u$  respectively<sup>19</sup>. Therefore

$$\begin{aligned} a_T(\mathcal{P}_b u - u_b, v_b) &= a_T(\mathcal{P}_b u - u, v_b) + a_T(u - u_b, v_b) \\ &= a_T(\mathcal{P}_b u - u, v_b) \\ &= a_*(\mathcal{P}_b u - u, v_b) - k^2(\mathcal{P}_b u - u, v_b)_{L^2(D)} \\ &= -k^2(\mathcal{P}_b u - u, v_b)_{L^2(D)}, \end{aligned}$$

that is,  $\mathcal{P}_b u - u_b$  solves the *discrete* Helmholtz problem

$$a_T(\mathcal{P}_b u - u_b, v_b) = (\tilde{f}, v_b)_{L^2(D)} \text{ for all } v_b \in V_{b,p}, \quad (2.56)$$

where  $\tilde{f} = -k^2(u - \mathcal{P}_b u)$ . One then uses this fact that  $\mathcal{P}_b u - u_b$  satisfies a discrete Helmholtz problem to prove a bound on the difference  $\mathcal{P}_b u - u_b$  directly.

All that remains to be discussed is how to prove the bound on  $\mathcal{P}_b u - u_b$ , using the fact that it solves the discrete Helmholtz problem (2.56). In [78, 79, 216] a discrete multiplier argument is used to prove a bound on  $\mathcal{P}_b u - u_b$ , reminiscent of the multiplier arguments used to prove a priori bounds on Helmholtz problems in Section 2.2. In [59] an argument using higher-order (discrete) norms is used in an argument conceptually similar to the modified duality arguments above; this argument is the heart of the proof in Section 2.4 below. In essence the argument in [59] reduces to showing the bounds

$$\|\mathcal{P}_b u - u_b\|_{L^2(D)} \lesssim (h + (hk)^p) \|u - \mathcal{P}_b u\|_{H_k^1(D)} + (h^{p+1}k^2 + b^{2p}k^{2p+2}) \|\mathcal{P}_b u - u_b\|_{p-1,b} \quad (2.57)$$

and

$$\|\mathcal{P}_b u - u_b\|_{p-1,b} \lesssim b^{2-p} \|u - \mathcal{P}_b u\|_{H_k^1(D)} + k^{p-1} \|\mathcal{P}_b u - u_b\|_{L^2(D)}, \quad (2.58)$$

where  $\|\cdot\|_{p-1,b}$  is a discrete norm analogous to the Sobolev norm of order  $p-1$ . The bounds (2.57) and (2.58) are then combined (under the assumption that  $hk \lesssim 1$ ) to show

$$\|\mathcal{P}_b u - u_b\|_{L^2(D)} \lesssim (h + (hk)^p) \|u - \mathcal{P}_b u\|_{H_k^1(D)} + ((hk)^{p+1} + b^{2p}k^{2p+1}) \|\mathcal{P}_b u - u_b\|_{L^2(D)},$$

and the final term can be absorbed into the left-hand side if  $b^{2p}k^{2p+1}$  is sufficiently small.

When using the elliptic projection in such an error-splitting argument, there are therefore two differences compared to the use of an elliptic projection in modified duality arguments:

1. One does not need the elliptic projection to be quasi-optimal (2.50), rather, one only needs to bound the error  $\|u - \mathcal{P}_b u\|_{H_k^1(D)}$ , where  $u$  solves a Helmholtz problem, in terms of  $\|f\|_{L^2(D)}$ .
2. The elliptic projection should be defined in the first argument, not the second, i.e.

$$a_T(\mathcal{P}_b u, v_b) = a_T(u, v_b) \text{ for all } v_b \in V_{b,p}. \quad (2.59)$$

<sup>19</sup>For the arguments in this section, we define the elliptic projection in the first argument, see (2.59) below.

## 2.4 NEW FINITE-ELEMENT-ERROR BOUNDS FOR THE HETEROGENEOUS HELMHOLTZ EQUATION

We now prove our new error bounds for higher-order finite-element approximations of the solution of the Helmholtz equation in heterogeneous media (Theorem 2.39 below)—we show that (for nontrapping media) the finite-element method applied to Problem 2.2 is  $hk^{(2p+1)/(2p)}$ -data-accurate, with constants depending on the medium. Our results are a generalisation of results proved by Du and Wu [59] for higher-order finite-element approximations of the Helmholtz equation in homogeneous media; our results and proofs broadly follow those in [59], with the main differences that:

1. We modify the proofs to cater for the heterogeneity of the coefficients, and
2. The dependence of our results on  $n$  is explicit.

In particular our results are explicit in  $n$  and  $k$  and are (in principle) explicit in  $A$ —see Remark 2.43 below for a discussion of why the results are not fully explicit in  $A$ . The only other result on data-accuracy for *heterogeneous* Helmholtz problems are by Wu and Zou [217, Lemma 3.3] (see Table 2.2), where they prove the  $h$ -finite-element method (with  $p = 1$ ) is  $hk^{3/2}$ -data-accurate in the case  $A = I$  and  $\|n - 1\|_{L^\infty(D; \mathbb{R})} \lesssim 1/k$ . However, the mesh conditions and error bounds in [217, Lemma 3.3] are *not* explicit in  $n$ .

The proofs of our results have many parts, and appear technical, largely due to the burden of explicitly keeping track of all of the constants involved. However, in essence, the proof consists of three ideas:

1. Decompose the error  $u - u_b = (u - \mathcal{P}_b u) + (\mathcal{P}_b u - u_b)$ , where  $\mathcal{P}_b u$  is an elliptic projection of  $u$ .
2. Bound the error  $u - \mathcal{P}_b u$  using the fact that  $\mathcal{P}_b u$  is a Galerkin projection.
3. Bound the error  $\mathcal{P}_b u - u_b$  in higher-order ‘discrete’ norms, using the fact that  $\mathcal{P}_b u - u_b$  solves a discrete Helmholtz problem.

The structure of the remainder of this section is as follows. In Section 2.4.1 we state our new finite-element-error bounds in Theorem 2.39. In Section 2.4.2 we prove a decomposition of the solution that allows us to prove a higher-order analogue of Lemma 2.22. In Section 2.4.3 we collect together some routine analysis results needed for our proofs. In Section 2.4.4 we prove error bounds for a number of different Galerkin projections (including the elliptic projection) that we use in subsequent proofs. In Section 2.4.5 we develop a notion of discrete derivatives and discrete norms, and prove properties of these norms. These norms will allow us to define higher-order discrete norms of functions in our finite-element space. Finally in Section 2.4.6 we prove our main finite-element error bounds.

### 2.4.1 Main result: new finite-element-error bounds

Recall that throughout this thesis we work under Assumption 2.18, i.e., we assume the spatial domain  $D$  is triangulated exactly, and the finite-element space  $V_{h,p}$  is conforming; see the discussion in Remark 2.17 above. Recall that under Assumption 2.18, if our mesh family  $(\mathcal{T}_h)_{h>0}$  is shape-regular, then Lemma 2.22 on the existence of a best approximation in  $V_{h,p}$  holds.

Since we are using higher-order finite elements (which we assume are of degree  $p$ ), the arguments we use to obtain full convergence rates will require smoothness assumptions on  $A$ ,  $n$ , and the boundaries  $\Gamma_D$  and  $\Gamma_I$ . We also make simple assumptions on  $k$ ,  $n_{\min}$ , and  $\|n\|_{L^\infty(D;\mathbb{R})}$  in order to simplify our calculations. We first recall the definition of quasi-uniformity (c.f. [29, Definition 4.4.13]):

**Definition 2.34** (Quasi-uniform). *Let  $\{\mathcal{T}_h\}_{h>0}$  be a set of triangulations of  $D$  indexed by their mesh size  $h$ . For each  $S \in \mathcal{T}_h$ , let  $B_S$  denote the largest ball contained in  $S$ . If there exists  $\rho > 0$  such that*

$$\min\{\text{diam } B_S : S \in \mathcal{T}\} \geq \rho h,$$

*then  $\{\mathcal{T}_h\}_{h>0}$  is said to be quasi-uniform.*

**Assumption 2.35** (Assumptions for higher- $p$  finite-element method bounds). *Assume*

- $\Gamma_D \neq \emptyset$ ,
- $\Gamma_D$  and  $\Gamma_I$  are  $C^{p,1}$ ,
- $A_{i,j} \in C^{p-1,1}(\overline{D})$  for all  $i, j$ , and
- $n \in H^{\max\{p-1, [d/2]+1\}}(D)$ .
- $V_{h,p}$  is defined on a quasi-uniform family of meshes.

The assumption  $\Gamma_D \neq \emptyset$  allows us to show the elliptic projection operator defined in (2.87) below is well-defined (see Remark 2.62 below for a more detailed discussion). The regularity assumptions on  $\Gamma_D$ ,  $\Gamma_I$ , and  $A$  ensure that we can apply a shift theorem for a related stationary diffusion equation up to order  $p-1$  (see Theorem 2.51 below). The assumption on  $n$  ensures that for all  $m \in [0, p-1]$  and for all  $v \in H^m(D)$ , the product  $nv \in H^m(D)$  (see Theorem 2.55 below). However, observe that we make no additional assumptions on the data  $f$  and  $g_I$ , and so the solution  $u$  may not be smoother than  $H^2$ .

We make the following assumption on the solution of Problem 2.12. Recall that the adjoint boundary-value problem of Problem 2.12 is the same as Problem 2.12, except (2.19) is replaced by

$$a_T(v, u) = \overline{L_T(v)} \text{ for all } v \in H_{0,D}^1(D), \quad (2.60)$$

see, e.g., [200, Section 4.2].

**Assumption 2.36.** For any  $f \in L^2(D)$  and  $g_I \in L^2(\Gamma_I)$ , Problem 2.12 (and its adjoint) has a unique solution  $u$  in  $H^2(D)$ , and there exists  $C_{\text{stab}} > 0$  (possibly dependent on  $A$ ,  $n$ , and  $k$ ) such that

$$k\|u\|_{L^2(D)} + |u|_{H^1(D)} + \frac{1}{k}|u|_{H^2(D)} \leq C_{\text{stab}} C_{f,g_I}, \quad (2.61)$$

where

$$C_{f,g_I} := \|f\|_{L^2(D)} + \|g_I\|_{L^2(\Gamma_I)}.$$

**Remark 2.37** (One can derive Assumption 2.36 for the adjoint from Assumption 2.36 for Problem 2.12). Suppose Assumption 2.36 only holds for solutions of Problem 2.12, not its adjoint. Observe that by taking complex conjugates of both sides of (2.60), we can write the adjoint problem (2.60) as: Find  $\tilde{u} \in H_{0,D}^1(D)$  such that

$$a_T^\dagger(\tilde{u}, v) = L_T(v) \text{ for all } v \in H_{0,D}^1(D),$$

where

$$a_T^\dagger(w, v) = \int_D ((A\nabla w) \cdot \nabla \bar{v} - k^2 n w \bar{v}) + ik \int_{\Gamma_I} \gamma_I w \gamma_I \bar{v}.$$

Therefore, by Corollary 4.23 in Chapter 4 below, we see  $\tilde{u}$  satisfies Problem 2.12 with right-hand side  $\overline{L_T(\bar{v})} = \int_D \bar{f} \bar{v} + \int_{\Gamma_I} \bar{g}_I \gamma_I \bar{v}$ . That is,  $\tilde{u}$  satisfies Problem 2.12 with right-hand sides  $\bar{f}$  and  $\bar{g}_I$ . Therefore, by Assumption 2.36,  $\tilde{u}$  exists, is unique, is in  $H^2(D)$  and satisfies (2.61), and hence all these properties also hold for  $\tilde{u}$ , i.e., for the adjoint problem.

Finally, we make the following assumption to simplify the proofs in this section. Assumption 2.38 is by no means necessary, but greatly simplifies the proofs.

**Assumption 2.38** (Assumptions for convenience of proofs). Assume:  $n_{\text{max}} := \|n\|_{L^\infty(D;\mathbb{R})} \geq 1$ ,  $k \geq 1$ ,  $n_{\text{min}} \leq 1$ ,  $hk \leq 1$ , and  $C_{\text{stab}} \geq 1$ .

Throughout this section, we adopt the following piece of notation.

$$n_{\text{var}} = \frac{n_{\text{max}}}{n_{\text{min}}}.$$

Our main result is the following theorem.

**Theorem 2.39** (Error bound for higher-order finite-element approximation of the heterogeneous Helmholtz equation). Let  $u$  be the solution of Problem 2.12. Under Assumptions 2.35, 2.36, and 2.38, there exist constants  $C_{\text{FEM},L^2}$ ,  $C_{\text{FEM},H^1}$ ,  $C_{\text{cond}} > 0$ , independent of  $h$ ,  $k$ , and  $n$ , such that if

$$h \leq C_{\text{cond}} \mathcal{C}(n) C_{\text{stab}}^{-\frac{1}{2p}} k^{-1-\frac{1}{2p}}, \quad (2.62)$$

then the finite-element solution  $u_h$  of Problem 2.20 exists, is unique, and satisfies the error bounds

$$\|u - u_h\|_{L^2(D)} \leq (h^2 + C_{\text{stab}} h (hk)^p + C_{\text{stab}}^2 (hk)^{2p}) C_{\text{FEM},L^2} \mathcal{C}_{L^2}(n) C_{f,g_I} \text{ and} \quad (2.63)$$

$$\|u - u_h\|_{H_k^1(D)} \leq (h + C_{\text{stab}}(hk)^p + C_{\text{stab}}^2 k(hk)^{2p}) C_{\text{FEM}, H^1} \mathcal{C}_{H^1}(n) C_{f, g_I}, \quad (2.64)$$

where

$$\begin{aligned} \mathcal{C}(n) &= \left( \left( (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\frac{1}{2}(\lfloor \frac{p-1}{2} \rfloor + 1)} n_{\max} n_{\min}^{-\frac{p}{2}} \right)^{p-1} (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\frac{1}{2}(\lfloor \frac{p-1}{2} \rfloor + 1)} n_{\max}^2 P_{p-2}(n) \right)^{-\frac{1}{2p}}, \\ \mathcal{C}_{L^2}(n) &= \left( (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\frac{1}{2}(\lfloor \frac{p-1}{2} \rfloor + 1)} n_{\max} n_{\min}^{-\frac{p}{2}} \right)^{p-1} n_{\text{var}}^5 n_{\min}^{-(p+1)} (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\frac{1}{2}(\lfloor \frac{p-1}{2} \rfloor + 1)} P_{p-2}(n)^2 \\ &\quad + P_{p-2}(n) \end{aligned} \quad (2.65)$$

and

$$\begin{aligned} \mathcal{C}_{H^1}(n) &= \max \left\{ P_{p-2}(n), \right. \\ &\quad \left[ \left( (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\frac{1}{2}(\lfloor \frac{p-1}{2} \rfloor + 1)} n_{\max} n_{\min}^{-\frac{p}{2}} \right) \left( \frac{\mathcal{C}_{L^2}(n)}{P_{p-2}(n)} - 1 \right) \right. \\ &\quad \left. \left. + \left( (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\frac{1}{2}(\lfloor \frac{p-1}{2} \rfloor + 1)} n_{\max} n_{\min}^{-\frac{p}{2}} \right) n_{\text{var}}^2 n_{\max} \right] P_{p-2}(n), \right. \\ &\quad \left. \mathcal{C}_{L^2}(n) - P_{p-2}(n) \right\}, \end{aligned}$$

where the function  $P_j(x)$  is defined in (2.69) below, and the  $n$ -dependent constant  $\mathcal{C}_{\text{err}}(n)$  is defined in (2.95) below.

The proof of Theorem 2.39 is on page 99 below.

**Remark 2.40** (Theorem 2.39 is a higher-order result). *At first glance, Theorem 2.39 appears not to be a higher-order result, because the lowest-order terms in (2.63) and (2.64) are  $h^2$  and  $h$ , respectively. However, we make two observations:*

1. *In general the solution  $u$  is only in  $H^2(D)$  (as in Assumption 2.36), and so we do not expect a typical higher-order error bound in  $h$  (involving  $h^{p+1}$  for  $\|u - u_h\|_{L^2(D)}$  or  $h^p$  for  $\|u - u_h\|_{H_k^1(D)}$ ).*
2. *The bounds (2.63) and (2.64) are higher-order bounds in  $h$ , in the sense that the lower-order terms in  $h$  do not dictate the rate of convergence. For example, if one takes  $k(hk)^{2p} \sim 1$  (so that (2.62) is satisfied and the final term in (2.64) is bounded) then for  $k$  large,  $h \sim k^{-(2p+1)/2p} \ll 1$  and  $(hk)^p \sim k^{-1/2} \ll 1$ . Therefore the dominant term in (2.64) (with regards to the magnitude of the finite-element error) is the final term. (Moreover, if  $k(hk)^{2p} \sim 1$ , then the magnitude of the other terms decreases as  $k \rightarrow \infty$ .) One could perform a similar analysis for (2.63).*

**Corollary 2.41** ( $hk^{(2p+1)/(2p)}$ -data-accuracy). *Under the assumptions of Theorem 2.39, if  $C_{\text{stab}} \lesssim 1$  (i.e., Problem 2.2 is nontrapping), then the finite-element method is  $hk^{(2p+1)/(2p)}$ -data-accurate (where the constants  $C_1$  and  $C_2$  in Definition 2.24 depend on  $A$  and  $n$ ).*

Whilst the calculations in this section are explicit in all the constants involved, these dependencies are complicated and, to a large extent, unnecessary to understand the flow of the arguments. Therefore, for ease of reference, the definition of all the constants in this section (which are many-layered and interdependent) are given in Section 2.4.7; i.e., any constant introduced or defined in Section 2.4 will be listed in Section 2.4.7. Also, for ease of reference, if a constant is only used inside a proof (and not in the statement of a theorem, or similar) it will typically be numbered using the equation number of its first appearance, similar to the approach in [42].

### Discussion of new finite-element-error bounds in Theorem 2.39

**Remark 2.42** (Relationship of new bounds to the work of Du and Wu). *In [59] Du and Wu proved that the  $h$ -finite-element method for the homogeneous Helmholtz interior impedance problem is  $hk^{(2p+1)/(2p)}$ -data-accurate. Our proof follows theirs, and achieves analogous results, see Corollary 2.41. We note as above that our results and proof have the following modifications to those of Du and Wu (the modifications listed in order of their impact upon the proof).*

1. *We prove bounds for heterogeneous coefficients, whereas [59] only has bounds for homogeneous coefficients. In particular, [59] uses the splitting argument of Melenk and Sauter (developed in [148, 149]) to prove a bound on the interpolation error that is higher-order in  $h$ . However, such a splitting argument only works in the homogeneous case, and so we instead use the recent work of Chaumont-Frelet and Nicaise [45], who provide a similar splitting in the heterogeneous case.*
2. *Because we work with heterogeneous coefficients, in several places in the proof we must work with  $n$ -weighted inner products and norms; see Remark 2.76 for more details on why  $n$ -weighted inner products and norms are required.*
3. *We explicitly track all of the constants involved in the proof—our results are completely explicit in  $n$ , and are in theory explicit in  $A$  (see Remark 2.43 below for information on the explicitness in  $A$ ).*
4. *We allow for the possibility that the Helmholtz problem is trapping—the constant  $C_{\text{stab}}$  appearing in (2.62) may depend on  $k$  (as well as  $A$  and  $n$ ). If the Helmholtz problem was nontrapping,  $C_{\text{stab}}$  would be independent of  $k$ .*
5. *We assume the existence of a Dirichlet scatterer  $D_-$ , as opposed to only considering the interior impedance problem, where  $D_- = \emptyset$ .*

**Remark 2.43** (Why are the results not fully explicit in  $A$ ?). *The condition (2.62) and the bounds (2.63) and (2.64) are not fully explicit in  $A$ , i.e. the constants  $C_{\text{cond}}$ ,  $C_{\text{FEM},L^2}$ , and  $C_{\text{FEM},H^1}$  may depend on  $A$ . This dependence is because the constants in the shift theorem for the stationary diffusion equation (Theorem 2.51 below) are not explicit in their  $A$ -dependence. In principle one can determine this dependence (it is determined for a right-hand side in  $L^2(D)$  and a solution in  $H^2(D)$  in [42, Appendix A]).*

**Remark 2.44** (Why the appearance of  $n_{\text{var}}$ ?). *The quantity  $n_{\text{var}} = n_{\text{max}}/n_{\text{min}}$  appears in multiple places in the definition of the  $n$ -dependent constants  $\mathcal{C}(n)$ ,  $\mathcal{C}_{L^2}(n)$ , and  $\mathcal{C}_{H^1}(n)$ . This appearance is*

mainly due to the fact that in multiple places in the proof of Theorem 2.39, we must convert from working in an  $n$ -weighted norm to a standard norm, and then later convert back to an  $n$ -weighted norm again.

This conversion is usually necessary because certain results (such as Theorems 2.51, 2.54, and 2.57 and Lemma 2.58) are only available in the literature in standard (non- $n$ -weighted) norms, and so to apply these results we must first transfer to non- $n$ -weighted norms, apply the results, and then transfer back. If one could prove these results for  $n$ -weighted norms (under sufficient smoothness conditions on  $n$ ) with constants that were completely explicit in  $n$ , then many of the instances of  $n_{\text{var}}$  could be removed.

**Remark 2.45** (Bounds not sharp in  $n$ ). *The  $n$ -dependence of the three constants  $\mathcal{C}(n)$ ,  $\mathcal{C}_{L^2}(n)$ , and  $\mathcal{C}_{H^1}(n)$  is almost certainly not sharp, because*

1. *The proof of Theorem 2.39 is complicated, and involves recursively applying bounds on finite-element functions (as in, e.g., the proof of Lemma 2.75, see (2.113) and (2.120)). These arguments may well result in non-sharp  $n$ -dependence.*
2. *As described in Remark 2.44 above, many of the appearances of  $n_{\text{var}}$  in the constants  $\mathcal{C}(n)$ ,  $\mathcal{C}_{L^2}(n)$ , and  $\mathcal{C}_{H^1}(n)$  are purely due to changing to non- $n$ -weighted norms, using results such as Theorem 2.51, then changing back, and so it is possible that the dependence of these constants on  $n_{\text{var}}$  (and therefore  $n$ ) is not sharp. One could mitigate these appearances of  $n_{\text{var}}$  by proving versions of, e.g., Theorem 2.51 in  $n$ -weighted norms, but it is not clear what  $n$ -dependence would result.*

Nevertheless, Theorem 2.39 is the first finite-element error bound for the heterogeneous Helmholtz equation that is completely explicit in  $n$ .

**Remark 2.46** ( $p$ -dependence). *To prove error bounds for  $h$   $p$ -finite-element methods, one must know explicitly how the error bounds (2.63) and (2.64) depend on  $p$  as well as  $h$  and  $k$ . Establishing the  $p$ -dependence for our finite-element error bounds would be challenging, and moreover, the  $p$ -dependence may not be sharp. It would be challenging because several of the constants in our argument depend on  $p$  in unknown ways (although this dependence could, in principle, be determined); e.g., the constants in Theorems 2.49, 2.51, 2.54, and 2.55 and Lemmas 2.56, 2.60, and 2.61. Moreover, even if one knew the  $p$ -dependence of all these constants explicitly, the  $p$ -dependence may still not be sharp, for similar reasons that the  $n$ -dependence of our results may not be sharp, as outlined in Remark 2.45 above.*

**Remark 2.47** (Special cases of  $n$ ). *In order to better understand the constants  $\mathcal{C}(n)$ ,  $\mathcal{C}_{L^2}(n)$ , and  $\mathcal{C}_{H^1}(n)$  it may be instructive to note their behaviour in the following three cases:*

- *If  $n = 1$ , then  $\mathcal{C}(n) = 1$ ,  $\mathcal{C}_{L^2}(n) = 1$ , and  $\mathcal{C}_{H^1}(n) = 2$ .*
- *If either  $n_{\text{min}}$  is fixed, and  $n_{\text{max}} \rightarrow \infty$ , or  $n_{\text{max}}$  is fixed, and  $n_{\text{min}} \rightarrow 0$ , then  $\mathcal{C}(n) \rightarrow 0$ , (i.e., the condition (2.62) becomes more restrictive),  $\mathcal{C}_{L^2}(n) \rightarrow \infty$ , and  $\mathcal{C}_{H^1}(n) \rightarrow \infty$  (i.e., the right-hand sides of the error bounds (2.63) and (2.64) become larger).*

**Remark 2.48** (Extension to different boundary conditions on  $\Gamma_I$ ). As in [59, Remark 5.3(e)], we remark that it is not obvious how to extend the proof of Theorem 2.39 to a Helmholtz problem with an exact Dirichlet-to-Neumann (DtN) boundary condition on  $\Gamma_I$ . In Lemmas 2.79 and 2.81 below one must bound terms involving  $\|\tilde{T}\theta_b\|_{L^2(\Gamma_I)} = k\|\theta_b\|_{L^2(\Gamma_I)}$ , where  $\tilde{T}$  is the impedance approximation to the DtN map  $T_R$ , i.e.,  $\tilde{T} = ik$ . These bounds are achieved using Lemma 2.78, where we bound  $\|\theta_b\|_{L^2(\Gamma_I)}$  by higher-order discrete norms of  $\theta_b$  and by the  $k$ -weighted  $H^1$ -norm of  $\rho$ . However, a crucial part of the proof of Lemma 2.78 is the fact that one has the equality  $(\tilde{T}\theta_b, \theta_b)_{L^2(\Gamma_I)} = \Im(\tilde{T}\theta_b, \theta_b)_{L^2(\Gamma_I)}$  (see (2.130) below).

To replicate the proofs of Lemmas 2.79 and 2.81 for an exact DtN boundary condition, one would need to bound terms involving  $\|T_R\theta_b\|_{L^2(\Gamma_I)}$ . However, repeating the proof of Lemma 2.78 for an exact DtN boundary condition only gives a bound on  $\Im(T_R\theta_b, \theta_b)_{L^2(\Gamma_I)}$ . Therefore, it is at this stage not clear how one can bound the terms on  $\Gamma_I$  for an exact DtN boundary condition.

## 2.4.2 Decomposition of solution and best approximation bound

For the first part of the proof of Theorem 2.39, we prove a best approximation bound (Lemma 2.56 below) in  $V_{b,p}$  for the solution of the Helmholtz equation, via a decomposition of the solution into functions of increasing regularity (Theorem 2.49 below). This technique was developed by Chaumont-Frelet and Nicaise in [45], and we follow their presentation (although we explicitly keep track of the constants involved at each point). Chaumont-Frelet and Nicaise were motivated by the work of Melenk and Sauter (see [45, Section 7]) in [148, 149] who showed for the homogeneous Helmholtz equation that the solution  $u$  can be decomposed as  $u = u_{H^2} + u_{\mathcal{A}}$ , where  $u_{H^2} \in H^2(D)$  but is not oscillatory ( $\|u_{H^2}\|_{H^2(D)} \lesssim 1$ ), and  $u_{\mathcal{A}}$  is analytic but is oscillatory ( $\|u_{\mathcal{A}}\|_{H^m(D)} \lesssim k^{m+\beta}$  for all  $m \geq 0$  and some  $\beta \in \mathbb{R}$ , where  $\beta$  depends on the problem being considered)<sup>20</sup>. Melenk and Sauter use their decomposition in [148, 149] to prove convergence results for  $hp$ -finite element methods for the Helmholtz equation. Whilst the work of Melenk and Sauter is very powerful, it is only valid for homogeneous media, and so Chaumont-Frelet and Nicaise developed their technique to handle heterogeneous media.

**Theorem 2.49** (Expansion of the solution of the Helmholtz equation). *Under Assumptions 2.35 and 2.36, if  $u$  is the solution of Problem 2.2 or its adjoint then there exists  $u_{\text{osc}} \in H^{p+1}(D)$  and a sequence  $u_j \in H^{j+2}(D)$ ,  $j = 0, \dots, p-2$  such that*

$$u = u_{\text{osc}} + \sum_{j=0}^{p-2} u_j. \quad (2.66)$$

Furthermore,

$$\|u_j\|_{H^{j+2}(D)} \leq C_{\text{expansion},j} P_j(n) k^j C_{f,g_t}, \quad (2.67)$$

<sup>20</sup>These results are proved with no obstacle and  $f$  given by a Dirac delta function in [148, Lemma 3.5] and for: (i) the IIP with a bounded Lipschitz boundary that is either a 2-d polygon or analytic, or (ii) the EDP with an analytic scatterer, in [149, Theorems 4.10, 4.20] respectively. In the former case the proof is under an assumption of a polynomial growth of the a priori bound [149, Assumption 4.8], i.e.,  $C_{\text{stab}}$  is a polynomial in  $k$ .

and

$$\|u_{\text{osc}}\|_{H^{p+1}(D)} \leq C_{\text{osc}} C_{\text{stab}} k^p C_{f, g_I}, \quad (2.68)$$

where

$$P_j(n) = \begin{cases} 1 & j = 0, 1 \\ \|n\|_{H^{\max\{p-1, \lceil d/2 \rceil + 1\}}(D)} P_{j-2}(n) + P_{j-1}(n) & 2 \leq j \leq p-2, \end{cases} \quad (2.69)$$

where

$$\|n\|_{H^{\max\{p-1, \lceil d/2 \rceil + 1\}}(D)} = \max\{\|n\|_{H^{p-1}(D)}, \|n\|_{H^{\lceil d/2 \rceil + 1}(D)}\}.$$

The proof of Theorem 2.49 is on page 73 below.

Recall that  $C_{\text{stab}}$  and  $C_{f, g_I}$  are defined in Assumption 2.36. The constants  $C_{\text{expansion}, j}$  and  $C_{\text{osc}}$  are defined in Section 2.4.7. Theorem 2.49 is essentially just [45, Theorem 1] in the particular case of a Helmholtz problem, but with the dependence on all the constants kept track of. The results in [45] are stated for a wider class of time-harmonic wave propagation problems, but the dependence on all of the constants is not made explicit. The main advantage of Theorem 2.49 is that it enables us to prove a higher-order best-approximation bound (Lemma 2.56 below) for solutions of the Helmholtz equation, *even though* the solutions do not have high regularity.

**Remark 2.50** (How oscillatory are the functions in Theorem 2.49?). *Recall that a higher power of  $k$  appearing in an a priori bound indicates that a function is more oscillatory. In (2.67) below, the  $(j+2)$ th-order norm of  $u_j$  is of order  $k^j$  (i.e., the power of  $k$  is two orders of magnitude less than the order of the norm) whereas in (2.68) the  $(p+1)$ st-order norm of  $u_{\text{osc}}$  is of order  $k^p$  (the power of  $k$  is one order of magnitude less than the order of the norm). Therefore, in this sense,  $u_{\text{osc}}$  is ‘more oscillatory’ than  $u_j$ .*

In order to obtain bounds for high  $p$ , we require the following shift theorem:

**Theorem 2.51** (Shift theorem). *Under Assumption 2.35, for all integers  $l \in [0, p-1]$  there exists a constant  $C_{A, l} > 0$  (depending on  $A$ ) such that if  $\tilde{f} \in H^l(D)$  and  $\tilde{g}_I \in H^{l+1/2}(\Gamma_I)$ , then there exists a unique  $\tilde{u} \in H^{l+2}(D)$  such that  $\tilde{u}$  solves*

$$\nabla \cdot (A \nabla \tilde{u}) = -\tilde{f}, \quad (2.70)$$

$$\partial_\nu \tilde{u} = \tilde{g}_I, \text{ and} \quad (2.71)$$

$$\gamma_D \tilde{u} = 0 \quad (2.72)$$

and  $\tilde{u}$  satisfies the bound

$$\|\tilde{u}\|_{H^{l+2}(D)} \leq C_{A, l} \left( \|\tilde{f}\|_{H^l(D)} + \|\tilde{g}_I\|_{H^{l+1/2}(\Gamma_I)} \right). \quad (2.73)$$

*Proof of Theorem 2.51.* The uniqueness and existence of  $\tilde{u}$  (in  $H_{0, D}^1(D)$ ) follows from the Lax-Milgram theorem, as the variational formulation of (2.70)–(2.72) is bounded and coercive. The proof of the higher regularity bounds uses standard elliptic regularity estimates in the interior and near the boundaries  $\Gamma_D$  and  $\Gamma_I$ , and the work of the proof is combining these estimates. As a

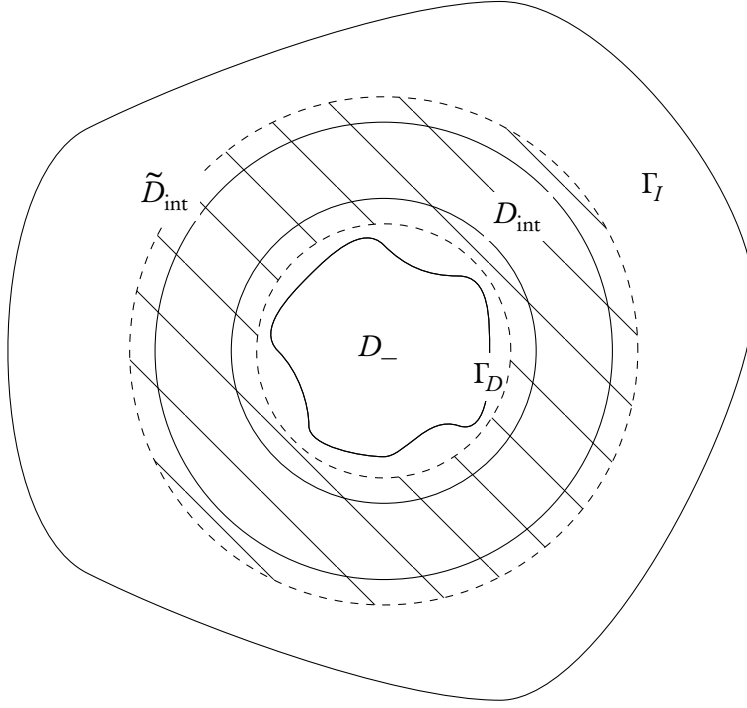


Figure 2.4: A schematic of the sets  $D_{\text{int}}$  (between solid circles only) and  $\tilde{D}_{\text{int}}$  (between dotted circles) from the proof of Theorem 2.51.

reference for these estimates we use [146, pp. 137-138]. By Assumption 2.35 we can apply these estimates, as we have the necessary higher regularity of the coefficients and the boundaries  $\Gamma_D$  and  $\Gamma_I$ .

To deal with the interior regularity and regularity near the boundary separately, we define the following subsets of  $D$ :  $D_{\text{int}}$ ,  $\tilde{D}_{\text{int}}$ ,  $D_{\text{scat}}$ , and  $D_{\text{trunc}}$  (see Figures 2.4–2.6 for a schematic) with the following properties:

- $D_{\text{int}} \subset\subset \tilde{D}_{\text{int}} \subset\subset D$ ,
- $\Gamma_D \subset \overline{D_{\text{scat}}}$
- $\text{dist}(D_{\text{scat}}, \Gamma_I) > 0$
- $\Gamma_I \subset \overline{D_{\text{trunc}}}$
- $\text{dist}(D_{\text{trunc}}, \Gamma_D) > 0$

First applying interior regularity [146, Theorem 4.16] in  $\tilde{D}_{\text{int}}$ , we obtain the bound

$$\|\tilde{u}\|_{H^{l+2}(D_{\text{int}})} \leq C_{\text{int},A,l} \left( \|\tilde{u}\|_{H^1(\tilde{D}_{\text{int}})} + \|\tilde{f}\|_{H^l(\tilde{D}_{\text{int}})} \right). \quad (2.74)$$

Applying regularity up to the boundary for Dirichlet data [146, Theorem 4.18 (i)] in  $D_{\text{scat}}$ , we obtain (as  $\gamma_D \tilde{u} = 0$ )

$$\|\tilde{u}\|_{H^{l+2}(D_{\text{scat}})} \leq C_{\text{scat},A,l} \left( \|\tilde{u}\|_{H^1(D)} + \|\tilde{f}\|_{H^l(D)} \right) \quad (2.75)$$

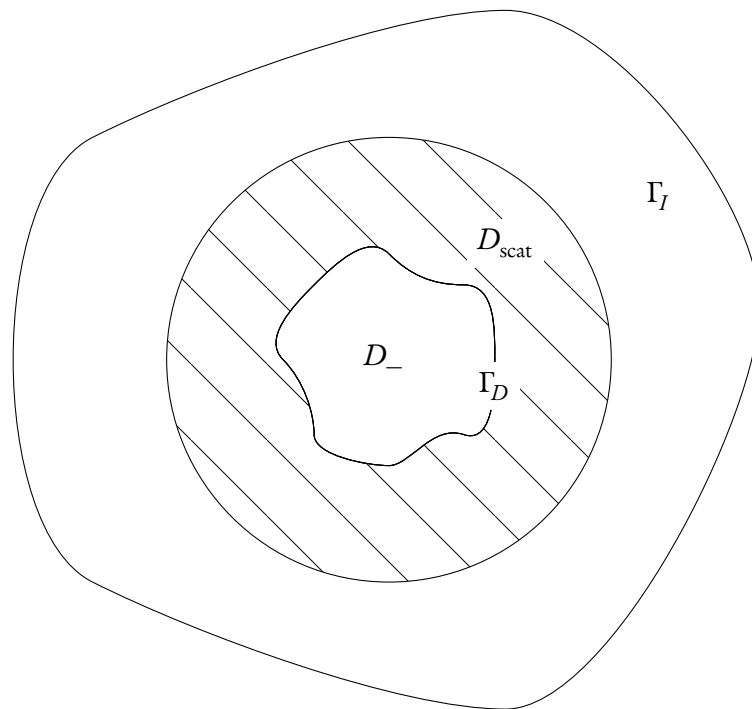


Figure 2.5: A schematic of the set  $D_{\text{scat}}$  from the proof of Theorem 2.51.

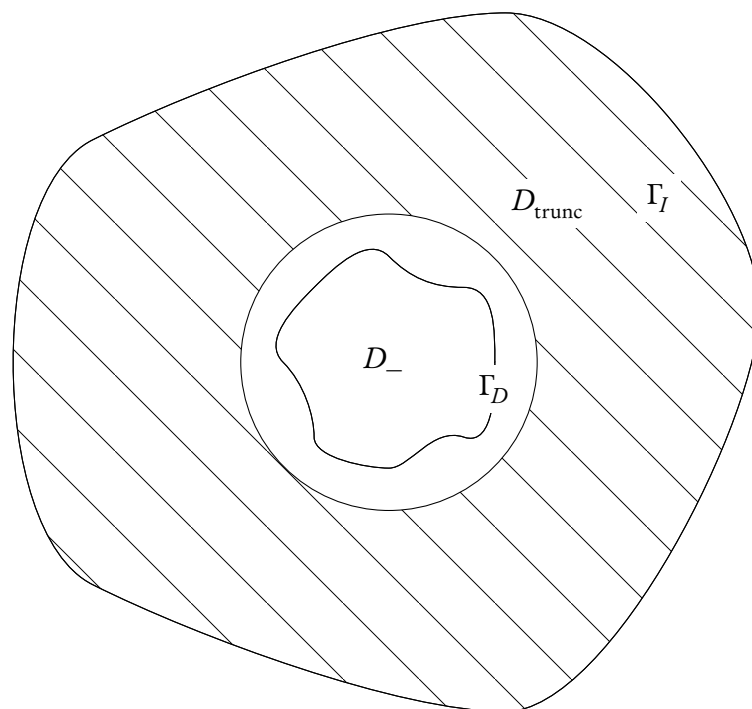


Figure 2.6: A schematic of the set  $D_{\text{trunc}}$  from the proof of Theorem 2.51.

and similarly for Neumann data [146, Theorem 4.18 (ii)] in  $D_{\text{trunc}}$ , we obtain

$$\|\tilde{u}\|_{H^{l+2}(D_{\text{trunc}})} \leq C_{\text{trunc},A,l} \left( \|\tilde{u}\|_{H^1(D)} + \|\partial_\nu \tilde{u}\|_{H^{l+1/2}(\Gamma_I)} + \|\tilde{f}\|_{H^1(D)} \right). \quad (2.76)$$

Combining (2.74)–(2.76), we obtain (2.73).  $\square$

**Remark 2.52** (Relaxing Assumption 2.35). *Observe that one can relax the assumption that  $\Gamma_D$  and  $\Gamma_I$  are  $C^{p,1}$  to piecewise- $C^{p,1}$  in certain scenarios, see, e.g., [45, Section 2.1].*

The following corollary follows from Theorem 2.51.

**Corollary 2.53.** *Under Assumption 2.35, let  $\tilde{f} \in H^l(D)$  and  $\tilde{g} \in H^{l+1}(D)$ , for  $0 \leq l \leq p-1$ . If  $\tilde{u} \in H^{l+2}(D)$  solves*

$$\begin{aligned} \nabla \cdot (A \nabla \tilde{u}) &= -\tilde{f}, \\ \gamma_D \tilde{u} &= 0, \end{aligned}$$

and

$$\partial_\nu \tilde{u} = \gamma_I \tilde{g},$$

then

$$\|\tilde{u}\|_{H^{l+2}(D)} \leq C_{A,l} (1 + C_{\text{Tr},l+1}) \left( \|\tilde{f}\|_{H^l(D)} + \|\gamma_I \tilde{g}\|_{H^{l+1}(D)} \right).$$

The proof of Corollary 2.53 requires the Trace theorem.

**Theorem 2.54** (Trace Theorem). *If  $v \in H^m(D)$ , for  $1/2 < m \leq p+1$ , then there exists  $C_{\text{Tr},m} > 0$  independent of  $v$  such that*

$$\|\gamma_I v\|_{H^{m-1/2}(\Gamma_I)} \leq C_{\text{Tr},m} \|v\|_{H^m(D)}.$$

For a proof of Theorem 2.54, see [146, Theorem 3.37]

*Proof of Corollary 2.53.* By Theorems 2.51 and 2.54

$$\|\tilde{u}\|_{H^{l+2}(D)} \leq C_{A,l} \left( \|\tilde{f}\|_{H^l(D)} + \|\gamma_I \tilde{g}\|_{H^{l+1/2}(\Gamma_I)} \right) \leq C_{A,l} \left( \|\tilde{f}\|_{H^l(D)} + C_{\text{Tr},l} \|\tilde{g}\|_{H^{l+1}(D)} \right),$$

and the result follows.  $\square$

The proof of Theorem 2.49 requires the following result on the multiplication of functions in Sobolev spaces.

**Theorem 2.55** (Multiplication in  $H^m(D)$ ). *For  $m \in \mathbb{N}$  let  $\tilde{m} \geq m$  and  $\tilde{m} > d/2$ . For all  $v_1 \in H^m(D)$ ,  $v_2 \in H^{\tilde{m}}(D)$ , the product  $v_1 v_2 \in H^m(D)$  and there exists a constant  $C_{\text{mult},m,\tilde{m}} > 0$  independent of  $v_1$  and  $v_2$  such that*

$$\|v_1 v_2\|_{H^m(D)} \leq C_{\text{mult},m,\tilde{m}} \|v_1\|_{H^m(D)} \|v_2\|_{H^{\tilde{m}}(D)}.$$

*Proof of Theorem 2.55.* The proof is immediate from the more general result on the multiplication of functions in Sobolev spaces defined on Lipschitz domains given in [20, Theorem 6.1, Corollary 6.3].  $\square$

*Proof of Theorem 2.49.* We give the proof only for Problem 2.2, as the proof for its adjoint is essentially identical. The idea of the proof is as follows. We write  $u$  as a formal series expansion

$$u = \sum_{j=0}^{\infty} u_j, \quad (2.77)$$

and then substitute this series into the PDE (2.7) and the boundary condition (2.8). Equating powers of  $k$ , we derive a recursive sequence of stationary diffusion equations for the functions  $u_j$ , with right-hand sides dependent on  $u_{j-1}$  and  $u_{j-2}$ . We use this recursive sequence and Corollary 2.53 to prove the a priori bounds (2.67).

We then define the  $l$ th remainder  $r_l = u - \sum_{j=0}^{l-1} u_j$ , and by applying the operator  $\nabla \cdot (A\nabla \cdot)$  with Neumann boundary conditions to  $r_l$ , we obtain a recursive sequence for the remainders  $r_l$ , and can similarly prove a priori bounds for the functions  $r_l$ . The oscillatory function  $u_{\text{osc}}$  is then just  $r_{p-1}$ . The format of this proof is identical to that in [45, Theorem 1], except we keep track of all of the constants involved.

For the purposes of the proof, it is more convenient to define  $v_j = u_j/k^j$ , so that the series expansion (2.77) becomes<sup>21</sup>

$$u = \sum_{j=0}^{\infty} k^j v_j \quad (2.78)$$

as in [45]. Also, in this proof, all the boundary-value problems involved included a zero Dirichlet condition on the scatterer  $\Gamma_D$ ; we omit this boundary condition throughout the proof for brevity.

By applying the Helmholtz operator to the formal series (2.78) and equating powers of  $k$  we obtain the following equations for  $v_j \in H^{j+2}(D)$ ,  $j \geq 1$ :

$$\nabla \cdot (A\nabla v_0) = -f \quad \text{and} \quad \partial_\nu v_0 = g_I,$$

$$\nabla \cdot (A\nabla v_1) = 0 \quad \text{and} \quad \partial_\nu v_1 = i\gamma_I v_0,$$

and

$$\nabla \cdot (A\nabla v_j) = -n v_{j-2} \quad \text{and} \quad \partial_\nu v_j = i\gamma_I v_{j-1} \quad \text{for } j \in [2, p-2]. \quad (2.79)$$

By Theorem 2.51 we immediately conclude the bound

$$\|v_0\|_{H^2(D)} \leq C_{A,0} C_{f,g_I} \leq C_{\text{expansion},0} C_{f,g_I}, \quad (2.80)$$

<sup>21</sup>In [45] the notation is changed slightly, and the series expansion is defined as  $u = \sum_{j=0}^{\infty} k^j u_j$ , i.e., the functions  $v_j$  in our proof are denoted  $u_j$  in [45].

i.e., (2.67) for  $j = 0$ . By Corollary 2.53 and (2.80) we can conclude the bound

$$\begin{aligned} \|v_1\|_{H^3(D)} &\leq C_{A,1}(1 + C_{\text{Tr},2})\|i v_0\|_{H^2(D)} \\ &\leq \max\{1, C_{A,1}\}(1 + C_{\text{Tr},2})C_{\text{expansion},0}C_{f,g_I} \\ &= C_{\text{expansion},1}C_{f,g_I}, \end{aligned}$$

i.e., (2.67) for  $j = 1$ .

We prove the bound (2.67) for higher  $j$  by induction. First observe that by Theorem 2.55, for any  $j \in \{0, 1, \dots, p-2\}$  and any  $v \in H^j(D)$

$$\|n v\|_{H^j(D)} \leq C_{\text{mult}}\|n\|_{H^{\max\{p-1, \lceil d/2 \rceil + 1\}}(D)}\|v\|_{H^j(D)}.$$

Let  $j \in [2, p-2]$  and suppose (2.67) holds for all  $s \in [0, j-1]$ . Using Corollary 2.53, we conclude that

$$\begin{aligned} \|v_j\|_{H^{j+2}(D)} &\leq C_{A,j}(1 + C_{\text{Tr},j+1})\left(\|n v_{j-2}\|_{H^j(D)} + \|v_{j-1}\|_{H^{j+1}(D)}\right) \\ &\leq C_{A,j}(1 + C_{\text{Tr},j+1})\left(C_{\text{mult}}\|n\|_{H^{\max\{p-1, \lceil d/2 \rceil + 1\}}(D)}\|v_{j-2}\|_{H^j(D)} + \|v_{j-1}\|_{H^{j+1}(D)}\right) \\ &\leq C_{A,j}(1 + C_{\text{Tr},j+1})\left(C_{\text{mult}}C_{\text{expansion},j-2}\|n\|_{H^{\max\{p-1, \lceil d/2 \rceil + 1\}}(D)}P_{j-2}(n)\right. \\ &\quad \left.+ C_{\text{expansion},j-1}P_{j-1}(n)\right)C_{f,g_I}, \text{ by induction,} \\ &= C_{\text{expansion},j}P_j(n)C_{f,g_I} \end{aligned}$$

by definition of  $C_{\text{expansion},j}$  and  $P_j(n)$ .

We will now define the remainders  $r_l$ , and proceed similarly. Let  $r_1 \in H^3(D)$  solve

$$\nabla \cdot (A \nabla r_1) = -k^2 u \text{ and } \partial_\nu r_1 = i k \gamma_I u.$$

Then by Corollary 2.53

$$\begin{aligned} \|r_1\|_{H^3(D)} &\leq C_{A,1}(1 + C_{\text{Tr},2})\left(k^2\|u\|_{H^1(D)} + k\|u\|_{H^2(D)}\right) \\ &\leq C_{A,1}(1 + C_{\text{Tr},2})k^2C_{\text{stab}}C_{f,g_I} \\ &= C_{\text{rem},1}k^2C_{\text{stab}}C_{f,g_I} \end{aligned}$$

by definition of  $C_{\text{rem},1}$ . Let  $r_2 \in H^4(D)$  solve

$$\nabla \cdot (A \nabla r_2) = -k^2 u \text{ and } \partial_\nu r_2 = i k \gamma_I r_1.$$

Then by Corollary 2.53

$$\begin{aligned} \|r_2\|_{H^4(D)} &\leq C_{A,2}(1 + C_{\text{Tr},3})\left(k^2\|u\|_{H^2(D)} + k\|r_1\|_{H^3(D)}\right) \\ &\leq C_{A,2}(1 + C_{\text{Tr},3})(1 + C_{\text{rem},1})C_{\text{stab}}k^3C_{f,g_I} \\ &= C_{\text{rem},2}C_{\text{stab}}k^3C_{f,g_I}. \end{aligned}$$

Then for  $j \geq 3$ , let  $r_j \in H^{j+2}(D)$  solve

$$\nabla \cdot (A \nabla r_j) = -k^2 r_{j-2} \text{ and } \partial_\nu r_j = ik\gamma_I r_{j-1}.$$

By induction and Corollary 2.53 again, letting  $u_{\text{osc}} = r_{p-1}$ , we have (2.68). It is straightforward to see that  $r_{p-1} + \sum_{j=1}^{p-2} u^{(j)}$  solves Problem 2.2, and therefore (2.66) holds, since  $u$  is unique.  $\square$

Using the expansion in Theorem 2.49, we can prove the following error bound for the best approximation of  $u$  in  $V_{b,p}$ :

**Lemma 2.56** (Best approximation error bound). *If Assumptions 2.35 and 2.36 hold, there exist constants  $C_{\text{FEM},1}, C_{\text{FEM},2} > 0$  independent of  $k$  and  $n$  (although dependent on  $A$  and  $p$ ) such that if  $u$  solves Problem 2.12 or its adjoint, then there exists  $\hat{u}_b \in V_{b,p}$  such that*

$$\|u - \hat{u}_b\|_{L^2(D)} \leq P_{p-2}(n)(C_{\text{FEM},1}h^2 + C_{\text{FEM},2}C_{\text{stab}}h(hk)^p)C_{f,g_I}, \quad (2.81)$$

$$\|u - \hat{u}_b\|_{H_k^1(D)} \leq 2P_{p-2}(n)(C_{\text{FEM},1}h + C_{\text{FEM},2}C_{\text{stab}}(hk)^p)C_{f,g_I}. \quad (2.82)$$

*Proof of Lemma 2.56.* We apply Lemma 2.22 to all the  $u_j$  and  $u_{\text{osc}}$  in Theorem 2.49, and obtain that there exist  $u_{j,b} \in V_{b,p}$   $j = 0, \dots, p-2$  and  $u_{\text{osc},b} \in V_{b,p}$  such that

$$\|u_j - u_{j,b}\|_{L^2(D)} + h\|u_j - u_{j,b}\|_{H^1(D)} \leq C_{\text{BA},j+2}C_{\text{expansion},j}P_j(n)h^{j+2}k^jC_{f,g_I}$$

and

$$\|u_{\text{osc}} - u_{\text{osc},b}\|_{L^2(D)} + h\|u_{\text{osc}} - u_{\text{osc},b}\|_{H^1(D)} \leq C_{\text{BA},p+1}C_{\text{osc}}C_{\text{stab}}h^{p+1}k^pC_{f,g_I}.$$

Therefore, by letting  $\hat{u}_b = u_{\text{osc},b} + \sum_{j=0}^{p-2} u_{j,b}$ , we have (2.81) and (2.82) (using the facts that  $hk \leq 1$  and  $P_{p-2}(n) \geq P_j(n) \geq 1$  for all  $j \leq p-2$  since  $\|n\|_{H^{\max\{p-1, \lfloor d/2 \rfloor + 1\}}(D)} \geq 1$ ).  $\square$

Observe that the bounds (2.81) and (2.82) are not fully  $p$ -explicit, as the constants  $C_{\text{BA},j}$  are dependent on  $p$  in an unknown way.

### 2.4.3 Routine analysis results

In this section we collect together routine results that we use throughout the following proofs.

- For  $N \in \mathbb{N}$  and  $a_1, \dots, a_N > 0$

$$\sqrt{\sum_{j=1}^N a_j^2} \leq \sum_{j=1}^N a_j. \quad (2.83)$$

- (Young's inequality) If  $s, q \in (1, \infty)$  and  $1/s + 1/q = 1$  then for all  $a, b > 0$

$$ab \leq \frac{a^s}{s} + \frac{b^q}{q}. \quad (2.84)$$

- (Cauchy's inequality) For all  $\varepsilon, a, b > 0$

$$ab \leq \frac{a^2}{2\varepsilon} + \frac{\varepsilon b^2}{2}. \quad (2.85)$$

**Theorem 2.57** (Multiplicative Trace Inequality). *There exists a constant  $C_{\text{MT}} > 0$  such that for all  $v \in H^1(D)$*

$$\|v\|_{L^2(\partial D)} \leq C_{\text{MT}} \|v\|_{L^2(D)}^{\frac{1}{2}} \|v\|_{H^1(D)}^{\frac{1}{2}}.$$

A proof of Theorem 2.57 can be found in [107, Last formula on p. 41].

**Lemma 2.58** (Poincaré–Friedrichs Inequality). *Let  $\Gamma \subseteq \partial D$  have nonvanishing  $d - 1$ -dimensional measure. There exist constants  $C_P, \tilde{C} > 0$  depending only on  $D$  and  $\Gamma$  such that for all  $v \in H^1(D)$*

$$\|v\|_{L^2(D)}^2 \leq C_P \|v\|_{H^1(D)}^2 + \tilde{C} \|v\|_{L^2(\Gamma)}^2.$$

In particular, taking  $\Gamma = \Gamma_D$ , for all  $v \in H_{0,D}^1(D)$

$$\|v\|_{L^2(D)} \leq C_P \|v\|_{H^1(D)}. \quad (2.86)$$

For a proof of Lemma 2.58 see [208, Lemma A.14].

#### 2.4.4 Error bounds for Galerkin projections

In this section we state a sequence of error bounds in negative Sobolev norms for two different projection operators. The proofs of these error bounds are all simple modifications of the standard duality-argument proofs of finite-element errors in negative Sobolev norms, as in, e.g., [29, Theorem 5.8.3]. We first define the projections we use.

Given  $w \in H_{0,D}^1(D)$ , define the elliptic projection  $\mathcal{P}_b w$  as the solution of the variational problem: Find  $\mathcal{P}_b w \in V_{b,p}$  such that.

$$a_*(\mathcal{P}_b w, v_b) = a_*(w, v_b) \text{ for all } v_b \in V_{b,p}, \quad (2.87)$$

where

$$a_*(v_1, v_2) = (A/\nabla v_1, \nabla v_2)_{L^2(D)}. \quad (2.88)$$

Observe that this construction defines the map  $\mathcal{P}_b : H_{0,D}^1(D) \rightarrow V_{b,p}$ . Also, observe that  $\mathcal{P}_b w$  is the finite-element approximation of the solution of the stationary diffusion problem with ‘diffusion coefficient’  $A$  and right-hand side in  $(H_{0,D}^1(D))^*$  given by  $a_*(w, \cdot)$ .

Define the  $L^2(D)$  projection in the  $n$ -weighted norm,  $Q_{b,n} : H_{0,D}^1(D) \rightarrow V_{b,p}$  by, for  $w \in H_{0,D}^1(D)$

$$(Q_{b,n} w, v_b)_{L^2(D),n} = (w, v_b)_{L^2(D),n} \text{ for all } v_b \in V_{b,p},$$

where, for  $v, w \in L^2(D)$ ,  $(v, w)_{L^2(D),n}$  is the  $n$ -weighted inner product

$$(v, w)_{L^2(D),n} := \int_D n v \bar{w}.$$

We also define the corresponding  $n$ -weighted  $L^2(D)$  norm  $\|v\|_{L^2(D),n} = \sqrt{(v, v)_{L^2(D),n}}$  and for  $m \in \mathbb{N}$ , the  $n$ -weighted  $H^m(D)$  norms

$$\|v\|_{H^m(D),n}^2 := \sum_{\alpha: |\alpha| \leq m} \|D^\alpha v\|_{L^2(D),n}^2,$$

and the negative  $n$ -weighted Sobolev norms

$$\|v\|_{H^{-m}(D),n} := \sup_{w \in H^m(D)} \frac{(v, w)_{L^2(D),n}}{\|w\|_{H^m(D),n}}. \quad (2.89)$$

Observe that, for  $v \in H^m(D)$ ,

$$n_{\min} \|v\|_{H^m(D)} \leq \|v\|_{H^m(D),n} \leq n_{\max} \|v\|_{H^m(D)}. \quad (2.90)$$

The proofs of the error bounds for the Galerkin projections will use the following notation for the solution operator of a particular stationary diffusion problem. We let  $\mathcal{S}_n : L^2(D) \rightarrow H^2(D)$  denote the solution operator for the following stationary diffusion equation: given  $\tilde{f} \in L^2(D)$  find  $\tilde{u} \in H^2(D)$  such that

$$\nabla \cdot (A \nabla \tilde{u}) = -n \tilde{f} \text{ in } D, \quad (2.91a)$$

$$\gamma_D \tilde{u} = 0 \text{ on } \Gamma_D, \text{ and} \quad (2.91b)$$

$$\partial_\nu \tilde{u} = 0 \text{ on } \Gamma_I. \quad (2.91c)$$

For the reason why there is a factor  $n$  on the right-hand side of (2.91a), see Remark 2.76 below. Observe that  $\mathcal{S}_n$  is well-defined by Theorem 2.51 as  $n \tilde{f} \in L^2(D)$ . Also, observe that  $\mathcal{S}_n^m$  is well-defined for any  $m \in \mathbb{N}$ , as  $H^2(D) \subseteq L^2(D)$ , and so one can place  $\mathcal{S}_n \tilde{f}$  on the right-hand side of (2.91a). For any  $\tilde{f} \in L^2(D)$  and for any  $v \in H_{0,D}^1(D)$ , we have, by Green's identity,

$$\int_D (A \nabla (\mathcal{S}_n \tilde{f})) \cdot \nabla \bar{v} = \int_D n \tilde{f} \bar{v},$$

i.e.,

$$(A \nabla (\mathcal{S}_n \tilde{f}), \nabla v)_{L^2(D)} = (\tilde{f}, v)_{L^2(D),n}. \quad (2.92)$$

We now state and prove error bounds for the two projections given above. As stated above,

the proofs below are all modifications of the standard proof (in, e.g., [29, Theorem 5.8.3]). We can, in essence, use the standard proof because all the projections defined above are Galerkin projections defined in terms of coercive and bounded sesquilinear forms on  $H^1(D)$  (for  $\mathcal{P}_h$ ) or  $L^2(D)$  (for  $Q_{h,n}$ ).

**Lemma 2.59** (Existence and uniqueness of Galerkin projections). *For any  $w \in H_{0,D}^1(D)$  the elliptic projection  $\mathcal{P}_h w$  and the  $n$ -weighted  $L^2$  projection  $Q_{h,n} w$  exist and are unique.*

*Proof of Lemma 2.59.* The existence and uniqueness of  $\mathcal{P}_h w$  and  $Q_{h,n} w$  follows from the Lax–Milgram Theorem (see, e.g., [29, Theorem 2.7.7]) applied in  $V_{h,p}$  (as in, e.g., [29, Corollary 2.7.13]), because  $\mathcal{P}_h w$  and  $Q_{h,n} w$  are defined by sesquilinear forms that are continuous and coercive on  $H_{0,D}^1(D)$  (equipped with the  $k$ -weighted  $H^1$  norm  $\|\cdot\|_{H_k^1(D)}$ ) and  $L^2(D)$  respectively.

Continuity and coercivity are immediate in the case of the  $n$ -weighted  $L^2$  projection  $Q_{h,n}$ . For the elliptic projection  $\mathcal{P}_h$ , continuity is immediate, and coercivity follows from the Poincaré inequality (2.86), because we work in  $H_{0,D}^1(D)$ , and so all the functions we consider vanish on  $\Gamma_D$ .  $\square$

**Lemma 2.60** (Error bounds for the elliptic projection). *Under Assumption 2.35, for any integer  $m \in [-1, p-1]$ , there exists a constant  $C_{\text{proj},-m} > 0$  such that for all  $w \in H_{0,D}^1(D)$*

$$\|w - \mathcal{P}_h w\|_{H^{-m}(D)} \leq C_{\text{proj},-m} h^{m+1} \inf_{w_b \in V_{h,p}} \|w - w_b\|_{H^1(D)}. \quad (2.93)$$

Because  $\mathcal{P}_h$  is defined in terms of a coercive and bounded sesquilinear form, the proof of Lemma 2.60 is completely standard, see, e.g., [29, Theorem 5.8.3].

**Lemma 2.61** (Error bounds for the elliptic projection in  $n$ -weighted norms).

*Under Assumption 2.35, for any integer  $m \in [-1, p-1]$ , there exists a constant  $C_{\text{weight},-m} > 0$  such that for all  $w \in H_{0,D}^1(D)$*

$$\|w - \mathcal{P}_h w\|_{H^{-m}(D),n} \leq C_{\text{weight},-m} \mathcal{C}_{\text{err},m}(n) h^{m+1} \inf_{w_b \in V_{h,p}} \|w - w_b\|_{H^1(D),n},$$

where

$$\mathcal{C}_{\text{err},m}(n) = \begin{cases} \frac{\|n\|_{H^{\max\{p-1, \lfloor d/2 \rfloor + 1\}}(D)}}{n_{\min}^2} n_{\text{var}} & \text{if } p > 1 \text{ and } m \in [1, p-1] \\ n_{\text{var}} & \text{if } m = -1, 0. \end{cases} \quad (2.94)$$

We define

$$\mathcal{C}_{\text{err}}(n) := \max_{m=-1, \dots, p-1} \{\mathcal{C}_{\text{err},m}(n)\}. \quad (2.95)$$

*Proof of Lemma 2.61.* For the case  $m = 0$ , using Lemma 2.60 and then converting to the  $n$ -weighted  $L^2$  norm yields (2.94). For the other cases, first observe that if we apply Céa’s Lemma to  $\mathcal{P}_h$  in the  $n$ -weighted  $H^1$  norm, we find

$$\|w - \mathcal{P}_h w\|_{H^1(D),n} \leq \frac{2\|A\|_{L^\infty(D;\text{op})}}{\min\{1, 1/C_P^2\}A_{\min}} n_{\text{var}} \inf_{w_b \in V_{h,p}} \|w - w_b\|_{H^1(D),n}, \quad (2.96)$$

since  $\mathcal{P}_b$  corresponds to a sesquilinear form with continuity constant  $\|A\|_{L^\infty(D;\text{op})}/n_{\min}$  and coercivity constant

$$\frac{A_{\min} \min\{1, 1/C_P^2\}}{2n_{\max}}.$$

For the case  $m = -1$ , (2.96) immediately gives (2.94).

If  $p = 1$ , then the proof is finished. If  $p > 1$ , let  $m \in [1, p - 1]$ . Let  $\phi \in H^m(D)$ , and observe that by Assumption 2.35 and Theorem 2.55, the product  $n\phi \in H^m(D)$ . Let  $\tilde{v} = \mathcal{S}_n(\phi)$  and observe  $\tilde{v} \in H^{m+2}(D)$  by Theorem 2.51. Observe that for all  $v \in H_{0,D}^1(D)$ , by the definition of  $\tilde{v}$  and the  $n$ -weighted inner product  $(\cdot, \cdot)_{L^2(D),n}$ , we have

$$(A\nabla v, \nabla \tilde{v})_{L^2(D)} = (nv, \phi)_{L^2(D)} = (v, \phi)_{L^2(D),n} \quad (2.97)$$

(where we multiply the complex conjugate of (2.91a) by  $v$ , and integrate by parts). Taking  $v = w - \mathcal{P}_b w$  in (2.97), we have (with  $\mathcal{I}_b v$  as defined in Lemma 2.22)

$$\begin{aligned} (v, \phi)_{L^2(D),n} &= (A\nabla(w - \mathcal{P}_b w), \nabla(\tilde{v} - \mathcal{I}_b v))_{L^2(D)} \text{ by Galerkin orthogonality for } \mathcal{P}_b, \\ &\leq C_{\text{BA},m+2} \|A\|_{L^\infty(D;\text{op})} \|\tilde{v}\|_{H^{m+2}(D)} h^{m+1} |w - \mathcal{P}_b w|_{H^1(D)} \text{ by Lemma 2.22,} \\ &\leq C_{\text{BA},m+2} C_{A,m} \|A\|_{L^\infty(D;\text{op})} \|n\phi\|_{H^m(D)} h^{m+1} |w - \mathcal{P}_b w|_{H^1(D)} \text{ by Theorem 2.51.} \end{aligned} \quad (2.98)$$

To bound the term  $\|n\phi\|_{H^m(D)}$  in (2.98) we use Theorem 2.55 and (2.90) to bound (2.98) above by

$$\begin{aligned} &C_{\text{BA},m+2} C_{A,m} C_{\text{mult},m,\max\{p-1, \lceil d/2 \rceil + 1\}} \\ &\|A\|_{L^\infty(D;\text{op})} \frac{\|n\|_{H^{\max\{p-1, \lceil d/2 \rceil + 1\}}(D)}}{n_{\min}} \|\phi\|_{H^m(D),n} h^{m+1} |w - \mathcal{P}_b w|_{H^1(D)}. \end{aligned} \quad (2.99)$$

Therefore we have, by definition of  $\|\cdot\|_{H^{-m}(D),n}$  and  $\|\cdot\|_{H^1(D),n}$ ,

$$\begin{aligned} &\|w - \mathcal{P}_b w\|_{H^{-m}(D),n} \leq \\ &C_{\text{BA},m+2} C_{A,m} C_{\text{mult},m,\max\{p-1, \lceil d/2 \rceil + 1\}} \|A\|_{L^\infty(D;\text{op})} \frac{\|n\|_{H^{\max\{p-1, \lceil d/2 \rceil + 1\}}(D)}}{n_{\min}^2} h^{m+1} \|w - \mathcal{P}_b w\|_{H^1(D),n}. \end{aligned} \quad (2.100)$$

Combining (2.96) and (2.100), we obtain (2.94).  $\square$

**Remark 2.62** (Subtleties regarding  $\mathcal{P}_b$  for the IIP). *If we consider the IIP (i.e., we remove the assumption  $\Gamma_D \neq \emptyset$  in Assumption 2.35), then we can no longer prove Lemmas 2.59, 2.60, or 2.61, because the sesquilinear form  $a_\star$  is no longer coercive on  $H_{0,D}^1(D)$ . (Note  $H_{0,D}^1(D) = H^1(D)$  in this case.)*

*This lack of coercivity stems from the fact that we cannot apply the Poincaré inequality (2.86) to*

functions in  $H^1(D)$ , because such functions are no longer zero on a portion of  $\partial D$  with non-zero  $d-1$ -dimensional measure. (Recall Lemma 2.58 requires such a property.) One can alternatively view this problem as arising from the fact that the stationary diffusion equation with Neumann boundary conditions does not have a unique solution (as one can simply add a constant to any solution and get another solution). In the case  $\Gamma_D = \emptyset$ , the PDE corresponding to the elliptic projection is precisely such a stationary diffusion equation: (2.46).

The remedy for this lack of coercivity/lack of uniqueness is to change the sesquilinear form  $a_*$  from (2.43) to either (2.44) or (2.45) (both of which are coercive in the  $k$ -weighted  $H^1$  norm, and so the proof of Lemma 2.59 goes through as before). This change corresponds to changing the PDE underlying the elliptic projection from (2.46) to either (2.47) or (2.48) respectively.

However, if one uses (2.45) to define  $a_*$  (i.e., one incorporates the impedance boundary condition into the sesquilinear form), then one cannot obtain results with the same (sharp)  $k$ -dependence as we do in Theorem 2.39.

The reason for this lack of sharp  $k$ -dependence is that there does not exist a higher-order  $k$ -independent shift theorem (analogous to Theorem 2.51) for the underlying PDE (2.48). Such a shift theorem is used in the proof of Lemmas 2.60 and 2.61 to prove bounds in negative-order norms. If one rewrote (2.48) as

$$\begin{aligned} -\Delta w &= F \text{ in } D \text{ and} \\ \partial_\nu w &= ikw \text{ on } \Gamma_I \end{aligned}$$

and then used Theorem 2.51 (which is a higher-order shift theorem) to obtain results analogous to Theorem 2.51, the constants in the resulting bounds (2.73) would now be  $k$ -dependent. Whilst a  $k$ -independent first-order shift theorem for (2.48) has been proved by Chaumont-Frelet, Nicaise, and Tomezyk [47, Theorems 3.1, 4.3, and 5.1], this is only a lowest-order shift theorem (i.e., the analogue of Theorem 2.51 for  $l = 0$ ) and no mention is made in [47] of an extension to higher order.

In summary, for the IIP, one cannot use (2.43) to define  $a_*$ , because  $a_*$  is now not coercive, and one cannot use (2.45) to define  $a_*$ , because it does not have a  $k$ -independent higher-order shift theorem. Therefore, for proving higher-order results, one must define  $a_*$  using (2.44). In this case one can then repeat the proof of Theorem 2.51 almost verbatim (as the results from [146] used in the proof of Theorem 2.51 also hold for the PDE (2.47)), and the proofs of Lemmas 2.60 and 2.61 proceed as before.

**Lemma 2.63** (Error bounds for  $n$ -weighted  $L^2(D)$  projection). *Under Assumption 2.35, for any integer  $m \in [0, p-1]$ , for all  $w \in H_{0,D}^1(D)$*

$$\left\| w - Q_{b,n} w \right\|_{H^{-m}(D),n} \leq C_{BA,m} n_{\text{var}} b^m \inf_{w_b \in V_{b,p}} \|w - w_b\|_{L^2(D),n}. \quad (2.101)$$

*Proof of Lemma 2.63.* Fix  $\tilde{v} \in H^m(D)$ . Then using Galerkin orthogonality for  $w - Q_{b,n} w$  we

have

$$\begin{aligned} (\omega - Q_{b,n}\omega, \tilde{v})_{L^2(D),n} &\leq \left\| \omega - Q_{b,n}\omega \right\|_{L^2(D),n} \|\tilde{v} - \mathcal{I}_b \tilde{v}\|_{L^2(D),n} \\ &\leq C_{\text{BA},m} n_{\text{var}} \left\| \omega - Q_{b,n}\omega \right\|_{L^2(D),n} b^m \|\tilde{v}\|_{H^m(D),n} \end{aligned}$$

by Lemma 2.22. Taking the supremum over  $\tilde{v}$ , we have

$$\left\| \omega - Q_{b,n}\omega \right\|_{H^{-m}(D),n} \leq C_{\text{BA},m} n_{\text{var}} b^m \left\| \omega - Q_{b,n}\omega \right\|_{L^2(D),n},$$

and hence by Céa's Lemma (since the inner product  $(\cdot, \cdot)_{L^2(D),n}$  is clearly bounded and coercive (with continuity and coercivity constants equal to 1) in the  $n$ -weighted  $L^2$ -norm  $\|\cdot\|_{L^2(D),n}$ ) the result follows.  $\square$

### 2.4.5 Discrete Sobolev spaces

When we analyse the high-order finite-element method, we need to measure higher-order norms of functions in the finite-element space  $V_{b,p}$ . However, as these functions do not have higher-order weak derivatives, we must first define a notion of higher-order discrete derivatives, and then develop some theory of so-called discrete Sobolev spaces. We follow the presentation in [59], albeit working in the heterogeneous case, and with some changes of notation. The main result of this section is Lemma 2.75 below giving the relationship between negative-order discrete Sobolev norms and negative-order continuous Sobolev norms.

**Definition 2.64** (Discrete derivative operator). *Define the  $A$ -weighted discrete second derivative operator  $\Delta_b : V_{b,p} \rightarrow V_{b,p}$  by, for  $z_b \in V_{b,p}$ ,*

$$(\Delta_b z_b, v_b)_{L^2(D),n} = (A \nabla z_b, \nabla v_b)_{L^2(D)} \text{ for all } v_b \in V_{b,p}. \quad (2.102)$$

**Lemma 2.65** (Discrete derivative operator is well-defined). *For any  $w_b \in V_{b,p}$ ,  $\Delta_b w_b$  exists and is unique.*

*Proof of Lemma 2.65.* Choose an orthonormal (in the  $n$ -weighted inner product) basis  $(\phi_j)_j$  for  $V_{b,p}$ . Write  $\Delta w_b = \sum_j w_j \phi_j$ , and take in turn  $v_b = \phi_j$  for each  $j$ ; then (2.102) is equivalent to the linear system  $lw = b$ , where  $b_j = (A \nabla w_b, \nabla \phi_j)_{L^2(D),n}$ . The solution of this linear system clearly exists and is unique.  $\square$

Since  $A$  is real and symmetric, it is self-adjoint. Hence it follows that  $\Delta_b$  is self-adjoint in the  $n$ -weighted inner product, since

$$\begin{aligned} (\Delta_b w_b, v_b)_{L^2(D),n} &= (A \nabla w_b, \nabla v_b)_{L^2(D)} \\ &= (\nabla w_b, A \nabla v_b)_{L^2(D)} = \overline{(\Delta_b v_b, w_b)_{L^2(D),n}} = (w_b, \Delta_b v_b)_{L^2(D),n}. \end{aligned}$$

Therefore  $\Delta_b$  is diagonalisable, i.e., there exists a set of eigenfunctions  $\phi_{1,b}, \dots, \phi_{\dim(V_{b,p}),b}$  with

corresponding real eigenvalues  $\lambda_{1,b}, \dots, \lambda_{\dim(V_{b,p}),b}$  such that the  $\phi_{m,b}$  form an orthonormal (in the  $n$ -weighted inner product) basis of  $V_{b,p}$ . The diagonalisability of  $\Delta_b$  allows us to define arbitrary powers of  $\Delta_b$ .

**Definition 2.66** (Higher-order discrete derivative operators).

For  $v_b \in V_{b,p}$ , if  $v_b = \sum_{m=1}^{\dim(V_{b,p})} a_m \phi_{m,b}$ , then for  $j \in \mathbb{R}$  define

$$\Delta_b^j v_b = \sum_{m=1}^{\dim(V_{b,p})} \lambda_{m,b}^j a_m \phi_{m,b}.$$

Observe that for  $w_b \in V_{b,p}$ ,  $\Delta_b^{-1}(\Delta_b w_b) = w_b$ , i.e., one can think of  $\Delta_b^{-1}$  as being in some sense a ‘discrete solution operator’ for the stationary diffusion equation (2.91a)–(2.91c). I.e.,  $\Delta_b^{-1}$  is a discrete counterpart to  $\mathcal{S}_n$ . We can use the higher-order derivative operators to define discrete higher-order norms:

**Definition 2.67** (Discrete higher-order norm). For  $v_b \in V_{b,p}$  and  $m \in \mathbb{R}$ , define

$$\|v_b\|_{m,b,n} = \left\| \Delta_b^{m/2} v_b \right\|_{L^2(D),n}.$$

**Remark 2.68** (Related literature for higher-order discrete norms). *To our knowledge, [59] is the only place in the literature where the above construction of higher-order discrete norms appears, although Thomée defines these norms for negative integers  $m$  in [206, Equation above Lemma 1]. However, the idea of using a self-adjoint, coercive operator to define a norm can be found in, e.g., [29, Section 6.2], [28, p. 238 ff.], where mesh-dependent norms are used to analyse multigrid methods, and [140, Section 2.1], where [140, Text at the bottom of page 9] observes that one can use a spectral decomposition to define arbitrary powers of such operators (analogous to Definition 2.66). See [10, Section 2.1] for a simpler exposition of defining a (fractional-order) Sobolev norm via an operator.*

*We observe in passing that Definition 2.66 is analogous to the spectral definition of the fractional Laplacian  $(-\Delta)^m$ ; see the recent review article [141, Section 2.5.1] for an overview of this idea.*

We will use the following lemma to bound the inner product of two discrete functions by their negative- and positive-higher-order discrete norms, or to transfer discrete derivatives from one argument of the inner product to the other.

**Lemma 2.69** (Introduction of derivatives into inner product). For  $v_b, w_b \in V_{b,p}$ , and  $m \in \mathbb{R}$  we have

$$(v_b, w_b)_{L^2(D),n} = \left( \Delta_b^{-m/2} v_b, \Delta_b^{m/2} w_b \right)_{L^2(D),n} \quad (2.103)$$

and

$$\left( \Delta_b^m v_b, v_b \right)_{L^2(D),n} = \left( \Delta_b^{m/2} v_b, \Delta_b^{m/2} v_b \right)_{L^2(D),n}. \quad (2.104)$$

*Proof of Lemma 2.69.* We only prove (2.103), as the proof of (2.104) is analogous. Since  $v_b, w_b \in V_{b,p}$ , there exist sequences  $(a_j)_{j=1, \dots, \dim(V_{b,p})}, (b_l)_{l=1, \dots, \dim(V_{b,p})}$  such that  $v_b = \sum_{j=1}^{\dim(V_{b,p})} a_j \phi_{j,b}$

and  $w_b = \sum_{l=1}^{\dim(V_{b,p})} b_l \phi_{l,b}$ . Then we have

$$\begin{aligned}
(\Delta_b^{-m/2} v_b, \Delta_b^{m/2} w_b)_{L^2(D),n} &= \int_D n \left( \sum_{j=1}^{\dim(V_{b,p})} \lambda_j^{-m/2} a_j \phi_{j,b} \right) \overline{\left( \sum_{l=1}^{\dim(V_{b,p})} \lambda_l^{m/2} b_l \phi_{l,b} \right)} \\
&= \sum_{j,l=1}^{\dim(V_{b,p})} \lambda_j^{-m/2} \lambda_l^{m/2} a_j b_l \int_D n \phi_{j,b} \overline{\phi_{l,b}} \quad \text{as the } \lambda_j \text{ are real,} \\
&= \sum_j^{\dim(V_{b,p})} a_j b_j \int_D n |\phi_{j,b}|^2 \quad \text{as the } \phi_{j,b} \text{ are orthonormal, (2.105)} \\
&= (v_b, w_b)_{L^2(D),n} \tag{2.106}
\end{aligned}$$

where (2.106) follows from (2.105) by repeating the above process in reverse without the factors  $\lambda_j^{-m/2}$  and  $\lambda_l^{m/2}$ .  $\square$

The next corollary follows from Lemma 2.69 and the Cauchy–Schwarz inequality.

**Corollary 2.70** (Inner product bounded by discrete norms). *If  $v_b, w_b \in V_{b,p}$ , then for all  $m \in \mathbb{R}$*

$$(v_b, w_b)_{L^2(D),n} \leq \|v_b\|_{-m,b,n} \|w_b\|_{m,b,n}.$$

We recall the standard inverse inequality for finite-element functions, so that we can prove an analogous inverse inequality for discrete norms.

**Lemma 2.71** (Standard inverse inequality). *Under Assumption 2.35 there exists  $C_{\text{inv},p} > 0$  such that for all  $v_b \in V_{b,p}$*

$$\|v_b\|_{H^1(D)} \leq C_{\text{inv},p} b^{-1} \|v_b\|_{L^2(D)}.$$

For a proof of Lemma 2.71 see, e.g., [29, Theorem 4.5.11 and Remark 4.5.20].

**Lemma 2.72** (Inverse inequality for discrete norms). *For all  $m \in \mathbb{R}$ , for all  $v_b \in V_{b,p}$*

$$\|v_b\|_{m,b,n} \leq C_{\text{disc,inv},p} \frac{1}{n_{\min}} b^{-1} \|v_b\|_{m-1,b,n}.$$

*Proof of Lemma 2.72.* We only prove the case  $m = 1$ , as the other cases will follow immediately. We have

$$\begin{aligned}
\|v_b\|_{1,b,n}^2 &= (\Delta_b^{1/2} v_b, \Delta_b^{1/2} v_b)_{L^2(D),n} \\
&= (\Delta_b v_b, v_b)_{L^2(D),n} \text{ by (2.104),} \\
&= (A \nabla v_b, \nabla v_b)_{L^2(D)} \text{ by definition of } \Delta_b, \\
&\leq \|A\|_{L^\infty(D;\text{op})} C_{\text{inv},p}^2 b^{-2} \frac{1}{n_{\min}^2} \|v_b\|_{L^2(D),n}^2
\end{aligned}$$

by the standard inverse estimate, and the result follows as  $\|\cdot\|_{0,b,n} = \|\cdot\|_{L^2(D),n}$ .

For  $m \neq 1$  we have

$$\begin{aligned} \|v_b\|_{m,b,n} &= \left\| \Delta_b^{\frac{m}{2}} v_b \right\|_{L^2(D),n} \\ &= \left\| \Delta_b^{\frac{1}{2}} \left( \Delta_b^{\frac{m-1}{2}} v_b \right) \right\|_{L^2(D),n} \\ &= \left\| \Delta_b^{\frac{m-1}{2}} v_b \right\|_{1,b,n} \\ &\leq C_{\text{disc,inv},p} \frac{1}{n_{\min}} b^{-1} \left\| \Delta_b^{\frac{m-1}{2}} v_b \right\|_{0,b,n} \end{aligned}$$

by the result for  $m = 1$ , and the result for  $m \neq 1$  follows.  $\square$

**Lemma 2.73** (Relationship between standard and discrete  $H^1$  norms). *Let  $v_b \in V_{b,p}$ . Then*

$$|v_b|_{H^1(D)} \leq A_{\min}^{-\frac{1}{2}} \|v_b\|_{1,b,n}.$$

*Proof of Lemma 2.73.* We have, using (2.104),

$$\begin{aligned} \|v_b\|_{1,b,n}^2 &= \left( \Delta_b^{1/2} v_b, \Delta_b^{1/2} v_b \right)_{L^2(D),n} \\ &= \left( \Delta_b v_b, v_b \right)_{L^2(D),n} = \left( A \nabla v_b, \nabla v_b \right)_{L^2(D)} \geq A_{\min} \|\nabla v_b\|_{L^2(D)}^2, \end{aligned}$$

and the result follows.  $\square$

To prove Lemma 2.75 below we require the following lemma giving the shift theorem in negative  $n$ -weighted norms. Recall that the  $n$ -weighted solution operator  $\mathcal{S}_n$  is defined as the solution operator of (2.91a)–(2.91c).

**Lemma 2.74** (Shift theorem in negative  $n$ -weighted norms). *Let  $\tilde{f} \in L^2(D)$  and  $m \in [-1, p-1]$  be an integer. Under Assumption 2.35 we have*

$$\left\| \mathcal{S}_n \tilde{f} \right\|_{H^{-m}(D),n} \leq C_{\text{shift},-m} \mathcal{C}_{\text{shift},m}(n) \left\| \tilde{f} \right\|_{H^{-(m+2)}(D),n}, \quad (2.107)$$

where

$$\mathcal{C}_{\text{shift},m}(n) := \begin{cases} n_{\max}^2 & \text{if } m = -1. \\ n_{\max} n_{\text{var}} & \text{if } m = 0 \\ \left\| n \right\|_{H^{\max\{p-1, \lceil d/2 \rceil + 1\}}(D)} n_{\text{var}} & \text{if } m \in [1, p-1] \end{cases}$$

We define

$$\mathcal{C}_{\text{shift}}(n) = \max_{m=-1, \dots, p-1} \{ \mathcal{C}_{\text{shift},m}(n) \}.$$

*Proof of Lemma 2.74.* We first observe that the operator  $\mathcal{S}_1$  is self-adjoint on  $L^2(D)$ . Let  $\Delta_A : H^2(D) \rightarrow L^2(D)$  denote the stationary diffusion operator  $\nabla \cdot (A \nabla \cdot)$  with no boundary conditions applied. Then for any  $v_1 \in L^2(D)$ ,

$$\Delta_A \circ \mathcal{S}_1 v_1 = v_1. \quad (2.108)$$

Moreover,  $\Delta_A$  is self-adjoint on the set  $\{v \in H^2(D) : v \text{ satisfies (2.91b) and (2.91c)}\}$  by Green's Theorem; this set contains the image of  $\mathcal{S}_1$ . Therefore, for any  $v_1, v_2 \in L^2(D)$ , we have

$$(\mathcal{S}_1 v_1, v_2)_{L^2(D)} = (\mathcal{S}_1 v_1, \Delta_A \circ \mathcal{S}_1 v_2)_{L^2(D)} = (\Delta_A \circ \mathcal{S}_1 v_1, \mathcal{S}_1 v_2)_{L^2(D)} = (v_1, \mathcal{S}_1 v_2)_{L^2(D)}, \quad (2.109)$$

using (2.108) and the fact that  $\Delta_A$  is self adjoint. I.e., (2.109) shows  $\mathcal{S}_1$  is self-adjoint on  $L^2(D)$ .

Observe that by Theorem 2.55, if  $v \in H^m(D)$  then  $nv \in H^m(D)$ , and therefore by Theorem 2.51,  $\mathcal{S}_1(nv) \in H^{m+2}(D)$ . With these facts in place we can compute

$$\begin{aligned} \|\mathcal{S}_n \tilde{f}\|_{H^{-m}(D),n} &= \sup_{v \in H^m(D)} \frac{(\mathcal{S}_n \tilde{f}, nv)_{L^2(D)}}{\|v\|_{H^m(D),n}} \\ &= \sup_{v \in H^m(D)} \frac{(\mathcal{S}_1(n\tilde{f}), nv)_{L^2(D)}}{\|v\|_{H^m(D),n}} \\ &= \sup_{v \in H^m(D)} \frac{(\tilde{f}, \mathcal{S}_1(nv))_{L^2(D),n}}{\|v\|_{H^m(D),n}} \quad \text{as } \mathcal{S}_1 \text{ is self-adjoint,} \\ &\leq \sup_{v \in H^m(D)} \frac{\|\tilde{f}\|_{H^{-(m+2)}(D),n} \|\mathcal{S}_1(nv)\|_{H^{m+2}(D),n}}{\|v\|_{H^m(D),n}} \\ &\leq \sup_{v \in H^m(D)} \frac{C_{A,m} n_{\max} \|nv\|_{H^m(D)} \|\tilde{f}\|_{H^{-(m+2)}(D),n}}{\|v\|_{H^m(D),n}} \end{aligned}$$

and by applying Theorem 2.55 to the term  $\|nv\|_{H^m(D)}$  (or, in the case  $m = 0$ , by observing that  $\|nv\|_{L^2(D)} \leq n_{\max} \|v\|_{L^2(D)}$ ), the result follows, except for  $m = -1$ .

For  $m = -1$ , we have, by the Lax–Milgram Theorem in non-weighted norms,

$$\|\mathcal{S}_n \tilde{f}\|_{H^1(D)} \leq \|n\tilde{f}\|_{H^{-1}(D)} / A_{\min}.$$

From (2.90) we have that

$$\|\mathcal{S}_n \tilde{f}\|_{H^1(D),n} \leq n_{\max} \|\mathcal{S}_n \tilde{f}\|_{H^1(D)}$$

and

$$\|n\tilde{f}\|_{H^{-1}(D)} \leq n_{\max} \|\tilde{f}\|_{H^{-1}(D),n},$$

and so the result follows.  $\square$

We can now prove the main result of this section.

**Lemma 2.75** (Relationship between discrete and continuous negative-order norms).

*Under Assumption 2.35, for any integer  $j \in [0, p + 1]$ , there exists a constant  $C_{2.110,j} > 0$  such that for all  $v_h \in V_{h,p}$ ,*

$$\|v_h\|_{-j,b,n} \leq C_{2.110,j} (\mathcal{C}_{\text{err}}(n)n_{\text{var}})^{\lfloor \frac{j}{2} \rfloor} n_{\text{max}} \sum_{m=0}^j h^m \|v_h\|_{H^{-(j-m)}(D),n}. \quad (2.110)$$

*Proof of Lemma 2.75.* Let  $w_b \in V_{h,p}$ , and define  $z_b = \Delta_b^{-1} w_b$  and  $z = \mathcal{S}_n w_b$  (observe  $z$  is well-defined since  $V_{h,p} \subseteq L^2(D)$ ). Then, for all  $v_h \in V_{h,p}$ , we have

$$\begin{aligned} (A\nabla z, \nabla v_h)_{L^2(D)} &= (A\nabla(\mathcal{S}_n w_b), \nabla v_h)_{L^2(D)} = (w_b, v_h)_{L^2(D),n}, \quad \text{and} \\ (A\nabla z_b, \nabla v_h)_{L^2(D)} &= (\Delta_b z_b, v_h)_{L^2(D),n} = (w_b, v_h)_{L^2(D),n}, \end{aligned}$$

where the equalities in the first line follow from the definition of  $z$  and (2.92), and the equalities in the second line follows from (2.102) and the definition of  $z_b$ . Therefore, for all  $v_h \in V_{h,p}$ ,  $(A\nabla z, \nabla v_h)_{L^2(D)} = (A\nabla z_b, \nabla v_h)_{L^2(D)}$ , i.e.,  $z_b = \mathcal{P}_b z$ .

We now have, for  $m \in [-1, p - 1]$

$$\begin{aligned} \left\| \Delta_b^{-1} w_b \right\|_{H^{-m}(D),n} &\leq \|z\|_{H^{-m}(D),n} + \|z - z_b\|_{H^{-m}(D),n} \\ &\leq \|z\|_{H^{-m}(D),n} + C_{\text{weight},-m} C_{A,0} C_{\text{BA},2} \mathcal{C}_{\text{err},m}(n) n_{\text{max}} h^{m+2} \|w_b\|_{L^2(D)} \\ &\quad \text{by Lemmas 2.22 and 2.61 and Theorem 2.51, since } z_b = \mathcal{P}_b z, \\ &= C_{\text{shift},-m} \mathcal{C}_{\text{shift},m}(n) \|w_b\|_{H^{-(m+2)}(D),n} \\ &\quad + C_{2.111,m} \mathcal{C}_{\text{err},m}(n) n_{\text{var}} h^{m+2} \|w_b\|_{L^2(D),n} \end{aligned} \quad (2.111)$$

by Lemma 2.74.

From (2.111), we can conclude that, for  $l \in \mathbb{N}$  and  $v_b \in V_{h,p}$ , writing  $w_b = \Delta_b^{-l+1} v_b$ ,

$$\begin{aligned} \left\| \Delta_b^{-l} v_b \right\|_{H^{-m}(D),n} &\leq C_{\text{shift},-m} \mathcal{C}_{\text{shift},m}(n) \left\| \Delta_b^{-l+1} v_b \right\|_{H^{-(m+2)}(D),n} \\ &\quad + C_{2.111,m} \mathcal{C}_{\text{err},m}(n) n_{\text{var}} h^{m+2} \left\| \Delta_b^{-l+1} v_b \right\|_{L^2(D),n} \end{aligned} \quad (2.112)$$

as  $\Delta_b^{-l} = \Delta_b^{-1} \Delta_b^{-l+1}$ . We now use (2.112) recursively to bound  $\|v_b\|_{j,b,n}$ .

If  $j = 2l$ , then one can show inductively using (2.112) that for any integer  $t \in [0, l]$

$$\|v_b\|_{-2l,b,n} \leq (\mathcal{C}_{\text{err}}(n)n_{\text{var}})^t \sum_{m=0}^t C_{2.113,m,t} h^{2m} \|v_b\|_{H^{-2(t-m)}(D)}, \quad (2.113)$$

where we define the constants  $C_{2.113,m,t}$  inductively by

$$C_{2.113,0,0} = 1, \quad (2.114)$$

$$C_{2.113,m,t} = C_{\text{shift},-2(t-1-m)} C_{2.113,m,t-1} \text{ for } 0 \leq m \leq t-1, \quad \text{and} \quad (2.115)$$

$$C_{2.113,t,t} = \sum_{m=0}^{t-1} C_{2.111,2(t-1-m)} C_{2.113,m,t-1}. \quad (2.116)$$

To see this recurrence, we prove the inductive step: suppose

$$\|v_b\|_{-2l,b,n} \leq (\mathcal{C}_{\text{err}}(n)n_{\text{var}})^{t-1} \sum_{m=0}^{t-1} C_{2.113,m,t-1} b^{2m} \|v_b\|_{H^{-2(t-1-m)}(D)}.$$

Then using (2.112), we have

$$\begin{aligned} \|v_b\|_{-2l,b,n} &\leq (\mathcal{C}_{\text{err}}(n)n_{\text{var}})^{t-1} \\ &\sum_{m=0}^{t-1} C_{2.113,m,t-1} b^{2m} \left( C_{\text{shift},-2(t-1-m)} \mathcal{C}_{\text{shift},2(t-1-m)}(n) \left\| \Delta_b^{-l+t} v_b \right\|_{H^{-(2(t-1-m)+2)}(D),n} \right. \\ &\quad \left. + C_{2.111,2(t-1-m)} \mathcal{C}_{\text{err},2(t-1-m)}(n)n_{\text{var}} b^{2(t-1-m)+2} \left\| \Delta_b^{-l+t} v_b \right\|_{L^2(D),n} \right), \end{aligned}$$

which upon rearranging, and using the fact that  $\mathcal{C}_{\text{shift},2(t-1-m)}(n) \leq \mathcal{C}_{\text{err},2(t-1-m)}(n)$  and  $n_{\text{var}} \geq 1$ , yields (2.113), with the recurrence (2.114)–(2.116).

If  $j = 2l + 1$ , then we first reduce  $\|v_b\|_{-j,b}$  to a point analogous to the even case, and then proceed as before. Let  $w_b$  and  $z_b$  be as at the beginning of the proof, and let  $z$  solve the variational formulation<sup>22</sup> of (2.91a)–(2.91c) (i.e. (2.92)) with  $\tilde{f} = w_b$ . Observe that we still have  $z_b = \mathcal{P}_b z$ , and

$$\|z\|_{H^1(D)} \leq \frac{1}{A_{\min}} \|nw_b\|_{H^{-1}(D)} \quad (2.117)$$

by the Lax–Milgram Theorem.

<sup>22</sup>We use the variational formulation here, as we will need to bound the  $H^1$ -norm of  $z$  by the  $H^{-1}$ -norm of  $w_b$ , which is immediate from the Lax–Milgram theorem.

Then

$$\begin{aligned} \left\| \Delta_b^{-1/2} \omega_b \right\|_{L^2(D),n} &= \left\| \Delta_b^{1/2} z_b \right\|_{L^2(D),n} \\ &= (\Delta_b z_b, z_b)_{L^2(D),n} \text{ by (2.104),} \end{aligned} \quad (2.118)$$

$$\begin{aligned} &= (A \nabla z_b, \nabla z_b)_{L^2(D)} \\ &\leq \|A\|_{L^\infty(D;\text{op})} \|\mathcal{P}_b z\|_{H^1(D)} \\ &\quad \text{by the Cauchy-Schwarz inequality and the definition of } z_b, \\ &\leq \|A\|_{L^\infty(D;\text{op})} (\|z\|_{H^1(D)} + C_{\text{proj},1} \|0 - z\|_{H^1(D)}) \text{ by Lemma 2.60,} \\ &\leq \frac{(1 + C_{\text{proj},1}) \|A\|_{L^\infty(D;\text{op})}}{A_{\min}} \|n \omega_b\|_{H^{-1}(D)} \text{ by (2.117)} \quad (2.119) \\ &\leq \frac{(1 + C_{\text{proj},1}) \|A\|_{L^\infty(D;\text{op})}}{A_{\min}} n_{\max} \|\omega_b\|_{H^{-1}(D),n} \end{aligned}$$

as in the proof of Lemma 2.74.

We now return to  $\|v_b\|_{-j,b,n}$ :

$$\begin{aligned} \|v_b\|_{-j,b,n} &= \left\| \Delta_b^{-l-1/2} v_b \right\|_{L^2(D),n} \\ &= \left\| \Delta_b^{-1/2} \Delta_b^{-l} v_b \right\|_{L^2(D),n} \\ &\leq \frac{(1 + C_{\text{proj},1}) \|A\|_{L^\infty(D;\text{op})}}{A_{\min}} n_{\max} \left\| \Delta_b^{-l} v_b \right\|_{H^{-1}(D),n} \end{aligned}$$

by (2.119).

Similarly to (2.113), one can use (2.112) recursively to show that, for any integer  $t \in [0, l]$

$$\begin{aligned} \left\| \Delta_b^{-l} v_b \right\|_{H^{-1}(D),n} &\leq (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^t \left( C_{2.120,0,t} \left\| \Delta_b^{-l+t} v_b \right\|_{H^{-(2t+1)}(D),n} \right. \\ &\quad \left. + \sum_{m=0}^t C_{2.120,m,t} h^{2m+1} \left\| \Delta_b^{-l+t} v_b \right\|_{H^{-2(t-m)}(D),n} \right), \end{aligned} \quad (2.120)$$

where we define the  $C_{2.120,m,t}$  inductively for  $t \in [0, l]$  by

$$C_{2.120,0,0} = 1, \quad (2.121)$$

$$C_{2.120,0,t} = C_{2.120,0,t-1} C_{\text{shift},-2(t-1)+1}, \quad (2.122)$$

$$C_{2.120,m,t} = C_{2.120,m,t-1} C_{\text{shift},-2(t-1-m)} \text{ for } m = 1, \dots, t-1, \quad \text{and} \quad (2.123)$$

$$C_{2.120,t,t} = C_{2.111,2(t-1)+1} + \sum_{m=0}^{t-1} C_{2.120,m,t-1} C_{2.111,2(t-1-m)}. \quad (2.124)$$

To show (2.120)–(2.124), observe that the case  $t = 0$ , including only (2.121) is immediate. We

show (2.120) for  $t \in [1, l]$  and (2.122)–(2.124) by induction. Suppose

$$\begin{aligned} \left\| \Delta_b^{-l} v_b \right\|_{H^{-1}(D),n} &\leq (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{t-1} \left( C_{2.120,0,t-1} \left\| \Delta_b^{-l+t-1} v_b \right\|_{H^{-(2(t-1)+1)}(D),n} \right. \\ &\quad \left. + \sum_{m=0}^{t-1} C_{2.120,m,t-1} b^{2m+1} \left\| \Delta_b^{-l+t-1} v_b \right\|_{H^{-2(t-1-m)}(D),n} \right). \end{aligned} \quad (2.125)$$

Then applying (2.112) to the terms in (2.125), with  $l$  in (2.112) given by  $l-t+1$  in (2.125), and  $m$  in (2.112) given by  $2(t-1)+1$  when applying (2.112) to the first term in (2.125), and  $2(t-1-m)$  for the other terms, we have

$$\begin{aligned} \left\| \Delta_b^{-l} v_b \right\|_{H^{-1}(D),n} &\leq \\ &(\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{t-1} C_{2.120,0,t-1} \left( C_{\text{shift},-(2(t-1)+1)} \mathcal{C}_{\text{shift},2(t-1)+1}(n) \left\| \Delta_b^{-l+t} v_b \right\|_{H^{-(2(t-1)+3)}(D),n} \right. \\ &\quad \left. + C_{2.111,2(t-1)t+1} \mathcal{C}_{\text{err},2(t-1)+1}(n) n_{\text{var}} b^{2t+1} \left\| \Delta_b^{-l+t} v_b \right\|_{L^2(D),n} \right) \\ &+ (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{t-1} \\ &\quad \sum_{m=1}^{t-1} C_{2.120,m,t-1} b^{2m+1} \left( C_{\text{shift},-2(t-1-m)} \mathcal{C}_{\text{shift},2(t-1-m)}(n) \left\| \Delta_b^{-l+t} v_b \right\|_{H^{-(2(t-1-m)+2)}(D),n} \right. \\ &\quad \left. + C_{2.111,2(t-1-m)} \mathcal{C}_{\text{err},2(t-1-m)}(n) n_{\text{var}} b^{2(t-m)} \left\| \Delta_b^{-l+t} v_b \right\|_{L^2(D),n} \right) \end{aligned} \quad (2.126)$$

and rearranging (2.126), and using the facts that  $\mathcal{C}_{\text{shift},m}(n) \leq \mathcal{C}_{\text{err},m}(n)$  for all  $m$  and  $n_{\text{var}} \geq 1$ , we obtain (2.120) with the constants  $C_{2.120,m,t}$  given by (2.121)–(2.124). Therefore, we conclude that if  $j = 2l + 1$  taking  $t = l$  in (2.120)

$$\begin{aligned} \|v_b\|_{-j,b,n} &\leq \frac{(1 + C_{\text{proj},1}) \|A\|_{L^\infty(D;\text{op})}}{A_{\min}} n_{\max} (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^l \\ &\quad \left( C_{2.120,0,l} \|v_b\|_{H^{-(2l+1)}(D),n} + \sum_{m=0}^l C_{2.120,m,l} b^{2m+1} \|v_b\|_{H^{-2(l-m)}(D),n} \right). \end{aligned} \quad (2.127)$$

The bound (2.113) gives the required bound for even  $j$ , and the bound (2.127) gives the required bound for odd  $j$ , and so summing the right-hand sides of (2.113) and (2.127) gives the result for any  $j$ , i.e., (2.110).  $\square$

**Remark 2.76** (Why is the factor  $n$  only on the right-hand side of (2.91a)?).

*The reason we define  $\mathcal{S}_n$  with a factor  $n$  on the right-hand side of (2.91a) but not on the left-hand side is somewhat buried in the proof of Theorem 2.39 and its associated lemmas. However, we give an overview of the reason here.*

*All the bounds in the proofs of Lemmas 2.78, 2.79, and 2.81 are in  $n$ -weighted discrete norms,*

because to prove bounds on the  $n$ -weighted  $L^2$  projection  $Q_{b,n}$  we work in  $n$ -weighted higher-order discrete norms (as in Lemma 2.63), rather than in non-weighted norms. Because we only work in  $n$ -weighted norms, we need Lemma 2.69 above to hold in the  $n$ -weighted inner product. To prove Lemma 2.69 in the  $n$ -weighted inner product, we therefore use the  $n$ -weighted inner product on the left-hand side of (2.102), where the  $\Delta_b$  operator appears. Because we use the  $n$ -weighted inner product on the left-hand side of (2.102), we must then use the  $n$ -weighted inner product on the right-hand side of (2.92). Using this inner product in (2.92) ensures the beginning of the proof of Lemma 2.75 works—see the proof for more details.

We do not, however, put a factor  $n$  on the left-hand side of (2.91a) or (2.92). If we did, when we apply Theorem 2.51 to  $\mathcal{S}_n \tilde{f}$  (as we do in Lemma 2.74), the resulting bounds would not be fully explicit in  $n$  (because in Theorem 2.51 we do not know explicitly how the constants depend on the ‘diffusion coefficient’). Not putting a factor  $n$  on the left-hand side of (2.91a) is the reason why there is a non-weighted inner product on the right-hand side of (2.102) (so that the beginning of the proof of Lemma 2.75, as mentioned above, works).

#### 2.4.6 Proof of Theorem 2.39

Having established the necessary preliminary results about discrete Sobolev spaces, we are now in a position to prove our main theorem, Theorem 2.39, which we do via a series of lemmas. The proof proceeds via an error-splitting argument, as discussed in Section 2.3.4. Recall that  $u$  solves Problem 2.12 and  $u_b$  solves Problem 2.20.

For ease of notation in the following lemmas, we follow the notation of [59] and let

$$\rho := u - \mathcal{P}_b u, \text{ and}$$

$$\theta_b := \mathcal{P}_b u - u_b = u - u_b - \rho.$$

The following lemma shows that  $\theta_b$  solves a discrete Helmholtz problem with data  $k^2 \rho$  (in  $D$ ) and  $ik\rho$  (on  $\Gamma_I$ ).

**Lemma 2.77** ( $\theta_b$  solves a discrete Helmholtz problem). *For any  $v_b \in V_{b,p}$ ,*

$$a_T(\theta_b, v_b) = k^2(Q_{b,n}\rho, v_b)_{L^2(D),n} + ik(\rho, v_b)_{L^2(\Gamma_I)}. \quad (2.128)$$

*Proof of Lemma 2.77.* Let  $v_b \in V_{b,p}$ . Then  $a_T(\theta_b, v_b) = a_T(u - u_b, v_b) - a_T(\rho, v_b) = -a_T(\rho, v_b)$  by Galerkin orthogonality. By definition of  $a_T$ , we have

$$-a_T(\rho, v_b) = -(A\nabla\rho, \nabla v_b)_{L^2(D)} + k^2(n\rho, v_b)_{L^2(D)} + ik(\rho, v_b)_{L^2(\Gamma_I)}.$$

By Galerkin orthogonality for  $\rho = u - \mathcal{P}_b u$  we have  $(A\nabla\rho, \nabla v_b)_{L^2(D)} = 0$ , and so by the definition of the  $n$ -weighted  $L^2$  inner product, and the  $n$ -weighted  $L^2$ -projection  $Q_{b,n}$ , the result follows.  $\square$

As mentioned in Remark 2.62, the definition of the elliptic projection  $\mathcal{P}_b$  ((2.87) and (2.88)

above) uses a Neumann boundary condition on  $\Gamma_I$ , rather than an impedance boundary condition, and therefore the right-hand side of (2.128) includes a term defined on the truncation boundary  $\Gamma_I$ . If the definition of the elliptic projection instead used an impedance boundary condition, this term would disappear. The presence of this term means that when we bound  $\|\theta_b\|_{L^2(D),n}$  and  $\|\theta_b\|_{p-1,b,n}$  in the proofs of Lemmas 2.79 and 2.81 below, we will encounter terms involving  $\|\theta_b\|_{L^2(\Gamma_I)}$ . Therefore, we first prove a bound on  $\|\theta_b\|_{L^2(\Gamma_I)}$ .

**Lemma 2.78** (Bound on  $\|\theta_b\|_{L^2(\Gamma_I)}$  by  $\|\theta_b\|_{p-1,b,n}$ ). *Under the assumptions of Theorem 2.39, we have*

$$\|\theta_b\|_{L^2(\Gamma_I)}^2 \leq C_{2.129,1}(\mathcal{C}_{\text{err}}(n)n_{\text{var}})^2 \left(\lfloor \frac{p-1}{2} \rfloor + 1\right) n_{\text{max}}^4 k^2 h^{2p-1} \|\theta_b\|_{p-1,b,n}^2 + C_{2.129,2} h \|\rho\|_{H_k^1(D)}^2, \quad (2.129)$$

*Proof of Lemma 2.78.* In (2.128), let  $v_b = \theta_b$ , and take the imaginary part to obtain

$$-k \|\theta_b\|_{L^2(\Gamma_I)}^2 = \Im k^2 (Q_{b,n} \rho, \theta_b)_{L^2(D),n} + \Re k (\rho, \theta_b)_{L^2(\Gamma_I)}. \quad (2.130)$$

Therefore by Corollary 2.70

$$\|\theta_b\|_{L^2(\Gamma_I)}^2 \leq k \left\| Q_{b,n} \rho \right\|_{-(p-1),b,n} \|\theta_b\|_{p-1,b,n} + \|\rho\|_{L^2(\Gamma_I)} \|\theta_b\|_{L^2(\Gamma_I)}. \quad (2.131)$$

We first bound the negative norm  $\left\| Q_{b,n} \rho \right\|_{-(p-1),b}$ , to do this we use Lemma 2.75. However, since we will apply Lemma 2.75 we need to estimate negative (standard) Sobolev norms of  $Q_{b,n} \rho$ ; for integers  $m \in [0, p-1]$  we have (observing that  $Q_{b,n} \mathcal{P}_b u = \mathcal{P}_b u$  since  $\mathcal{P}_b u \in V_{h,p}$ ).

$$\begin{aligned} \left\| Q_{b,n} \rho \right\|_{H^{-(p-1-m)}(D),n} &\leq \left\| Q_{b,n} u - u \right\|_{H^{-(p-1-m)}(D),n} + \|u - \mathcal{P}_b u\|_{H^{-(p-1-m)}(D),n} \\ &\leq C_{\text{BA},(p-1-m)} n_{\text{var}} h^{(p-1-m)} \|u - \mathcal{P}_b u\|_{L^2(D),n} \\ &\quad + C_{\text{weight},-(p-1-m)} \mathcal{C}_{\text{err},(p-1-m)}(n) h^{p-m} \|u - \mathcal{P}_b u\|_{H^1(D),n} \\ &\quad \text{by Lemmas 2.63 and 2.61,} \\ &\quad \text{taking } w_b = \mathcal{P}_b u \text{ in Lemma 2.61 and (2.101),} \\ &\leq \left( C_{\text{BA},(p-1-m)} C_{\text{weight},0} + C_{\text{weight},-(p-1-m)} \right) \\ &\quad \mathcal{C}_{\text{err}}(n) n_{\text{var}} n_{\text{max}} h^{p-m} \|\rho\|_{H_k^1(D)} \end{aligned} \quad (2.132)$$

by Lemma 2.61. By Lemma 2.75 and (2.132) we have

$$\begin{aligned} \left\| Q_{b,n} \rho \right\|_{-(p-1),b,n} &\leq C_{2.110,p-1} (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\lfloor \frac{p-1}{2} \rfloor} n_{\text{max}} \sum_{m=0}^{p-1} h^m \left\| Q_{b,n} \rho \right\|_{H^{-(p-1-m)}(D)} \\ &\leq C_{2.133} (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\lfloor \frac{p-1}{2} \rfloor + 1} n_{\text{max}}^2 h^p \|\rho\|_{H_k^1(D)}. \end{aligned} \quad (2.133)$$

To bound  $\|\rho\|_{L^2(\Gamma_I)}$  appearing in the second term on the right-hand side of (2.131) we use Theorem 2.57 and Lemma 2.60. We take  $w_b = \mathcal{P}_b u$  in (2.93) and use the fact that  $\|\cdot\|_{H^1(D)} \leq \|\cdot\|_{H_k^1(D)}$

to obtain

$$\|\rho\|_{L^2(\Gamma_I)} \leq C_{\text{MT}} \|\rho\|_{H^1(D)}^{1/2} \|\rho\|_{L^2(D)}^{1/2} \leq C_{\text{MT}} C_{\text{proj},0}^{\frac{1}{2}} b^{\frac{1}{2}} \|\rho\|_{H^1(D)} \leq C_{\text{MT}} C_{\text{proj},0}^{\frac{1}{2}} b^{\frac{1}{2}} \|\rho\|_{H_k^1(D)}. \quad (2.134)$$

Therefore by (2.134) and Young's inequality (2.84), we obtain

$$\|\rho\|_{L^2(\Gamma_I)} \|\theta_b\|_{L^2(\Gamma_I)} \leq \frac{1}{2} C_{\text{MT}}^2 C_{\text{proj},0} b \|\rho\|_{H_k^1(D)}^2 + \frac{1}{2} \|\theta_b\|_{L^2(\Gamma_I)}^2. \quad (2.135)$$

By combining (2.131), (2.133), and (2.135) we have

$$\begin{aligned} \|\theta_b\|_{L^2(\Gamma_I)}^2 &\leq k C_{2.133} (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\lfloor \frac{p-1}{2} \rfloor + 1} n_{\text{max}}^2 b^p \|\rho\|_{H_k^1(D)} \|\theta_b\|_{p-1,b,n} \\ &\quad + \frac{1}{2} C_{\text{MT}}^2 C_{\text{proj},0} b \|\rho\|_{H_k^1(D)}^2 + \frac{1}{2} \|\theta_b\|_{L^2(\Gamma_I)}^2. \end{aligned} \quad (2.136)$$

By using Young's inequality on the first term in (2.136), and moving the  $\|\theta_b\|_{L^2(\Gamma_I)}^2/2$  term onto the left-hand side, we obtain (2.129).  $\square$

We can now prove the main two lemmas in the proof of Theorem 2.39.

**Lemma 2.79** (Bound on higher-order discrete norms of  $\theta_b$  by  $\|\theta_b\|_{L^2(D)}$ ).

Under the assumptions of Theorem 2.39, for integer  $m \in [1, p-1]$  there exist constants  $C_{2.137,m,1}, C_{2.137,m,2} > 0$  such that

$$\begin{aligned} \|\theta_b\|_{m,b,n} &\leq C_{2.137,m,1} \left( (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\frac{1}{2}(\lfloor \frac{p-1}{2} \rfloor + 1)} n_{\text{max}} n_{\text{min}}^{-\frac{p}{2}} \right)^m k^m \|\theta_b\|_{L^2(D)} \\ &\quad + C_{2.137,m,2} \left( (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\frac{1}{2}(\lfloor \frac{p-1}{2} \rfloor + 1)} n_{\text{max}} n_{\text{min}}^{-\frac{p}{2}} \right)^m n_{\text{var}}^2 n_{\text{max}}^{1-m} n_{\text{min}}^{1-m} b^{1-m} \|\rho\|_{H_k^1(D)}. \end{aligned} \quad (2.137)$$

*Proof of Lemma 2.79.* By inserting the definitions of  $a_T$  and  $\Delta_b$  into (2.128) and rearranging, we have for any  $v_b \in V_{b,p}$

$$(\Delta_b \theta_b, v_b)_{L^2(D),n} = k^2 (\theta_b, v_b)_{L^2(D),n} + k^2 (Q_{b,n} \rho, v_b)_{L^2(D),n} + ik (\theta_b, v_b)_{L^2(\Gamma_I)} + ik (\rho, v_b)_{L^2(\Gamma_I)}.$$

Therefore, if we take  $v_b = \Delta_b^{m-1} \theta_b$ , by Lemma 2.69 we have

$$\begin{aligned} \|\theta_b\|_{m,b,n}^2 &= k^2 \|\theta_b\|_{m-1,b,n}^2 + k^2 \left( \Delta_b^{\frac{m-1}{2}} Q_{b,n} \rho, \Delta_b^{\frac{m-1}{2}} \theta_b \right)_{L^2(D),n} \\ &\quad + ik (\theta_b, \Delta_b^{m-1} \theta_b)_{L^2(\Gamma_I)} + ik (\rho, \Delta_b^{m-1} \theta_b)_{L^2(\Gamma_I)}. \end{aligned} \quad (2.138)$$

We now proceed to bound the two terms in (2.138) defined on the truncation boundary  $\Gamma_I$ . For

the first term, we have

$$\begin{aligned}
(\theta_b, \Delta_b^{m-1} \theta_b)_{L^2(\Gamma_I)} &\leq \|\theta_b\|_{L^2(\Gamma_I)} \|\Delta_b^{m-1} \theta_b\|_{L^2(\Gamma_I)} \\
&\leq C_{\text{MT}} C_{\text{inv},p}^{1/2} \|\theta_b\|_{L^2(\Gamma_I)} b^{-\frac{1}{2}} \|\Delta_b^{m-1} \theta_b\|_{L^2(D)} \\
&\quad \text{by Theorem 2.57 and Lemma 2.71,} \\
&= C_{\text{MT}} C_{\text{inv},p}^{1/2} b^{-\frac{1}{2}} \|\theta_b\|_{L^2(\Gamma_I)} n_{\min}^{-1} \|\theta_b\|_{2m-2,b,n} \\
&\quad \text{by the definition of } \|\cdot\|_{2m-2,b,n}, \\
&\leq C_{\text{MT}} C_{\text{inv},p}^{1/2} C_{\text{disc,inv},p}^{m-1} n_{\min}^{-m} b^{-m+\frac{1}{2}} \|\theta_b\|_{L^2(\Gamma_I)} \|\theta_b\|_{m-1,b,n} \\
&\quad \text{by Lemma 2.72 applied } m-1 \text{ times,} \tag{2.139}
\end{aligned}$$

$$\begin{aligned}
&\leq C_{\text{MT}} C_{\text{inv},p}^{1/2} C_{\text{disc,inv},p}^{m-1} b^{-m+\frac{1}{2}} n_{\min}^{-m} \\
&\quad \left( C_{2.129,1}^{\frac{1}{2}} (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\lfloor \frac{p-1}{2} \rfloor + 1} n_{\max}^2 k b^{p-\frac{1}{2}} \|\theta_b\|_{p-1,b,n} \right. \\
&\quad \left. + C_{2.129,2}^{\frac{1}{2}} b^{\frac{1}{2}} \|\rho\|_{H_k^1(D)} \right) \|\theta_b\|_{m-1,b,n} \\
&\quad \text{by Lemma 2.78 and (2.83),} \\
&\leq \left( C_{2.140,1} (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\lfloor \frac{p-1}{2} \rfloor + 1} n_{\max}^2 n_{\min}^{-p} k \|\theta_b\|_{m-1,b,n} \right. \\
&\quad \left. + C_{2.140,2} n_{\min}^{-m} b^{1-m} \|\rho\|_{H_k^1(D)} \right) \|\theta_b\|_{m-1,b,n} \tag{2.140}
\end{aligned}$$

by Lemma 2.72 applied  $p-m$  times.

To bound the second boundary term in (2.138), we have

$$(\rho, \Delta_b^{m-1} \theta_b)_{L^2(\Gamma_I)} \leq C_{\text{MT}} C_{\text{inv},p}^{1/2} C_{\text{disc,inv},p}^{m-1} n_{\min}^{-m} b^{\frac{1}{2}-m} \|\rho\|_{L^2(\Gamma_I)} \|\theta_b\|_{m-1,b} \tag{2.141}$$

using the same reasoning as we used to obtain (2.139) above. By Theorem 2.57 and Lemma 2.60 (with  $w_b = \mathcal{P}_b u$ ) we have

$$\|\rho\|_{L^2(\Gamma_I)} \leq C_{\text{MT}} C_{\text{proj},0}^{\frac{1}{2}} b^{\frac{1}{2}} \|\rho\|_{H_k^1(D)}. \tag{2.142}$$

Inserting (2.142) into (2.141) we obtain

$$(\rho, \Delta_b^{m-1} \theta_b)_{L^2(D)} \leq C_{2.143} n_{\min}^{-m} b^{1-m} \|\rho\|_{H_k^1(D)} \|\theta_b\|_{m-1,b}. \tag{2.143}$$

Therefore, from (2.138), (2.140), (2.143), and the Cauchy–Schwarz inequality, we have

$$\begin{aligned} \|\theta_b\|_{m,b,n}^2 &\leq k^2 \|\theta_b\|_{m-1,b,n}^2 + k^2 \left\| Q_{b,n} \rho \right\|_{m-1,b,n} \|\theta_b\|_{m-1,b,n} \\ &\quad + k \left( C_{2.140,1} (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\lfloor \frac{p-1}{2} \rfloor + 1} n_{\text{max}}^2 n_{\text{min}}^{-p} k \|\theta_b\|_{m-1,b,n} \right. \\ &\quad \left. + C_{2.140,2} n_{\text{min}}^{-m} b^{1-m} \|\rho\|_{H_k^1(D)} \right) \|\theta_b\|_{m-1,b,n} \\ &\quad + k C_{2.143} n_{\text{min}}^{-m} b^{1-m} \|\rho\|_{H_k^1(D)} \|\theta_b\|_{m-1,b} \end{aligned}$$

Therefore using Young’s inequality (2.84) with  $s = q = 2$  we have

$$\begin{aligned} \|\theta_b\|_{m,b,nXS}^2 &\leq k^2 \left( \frac{3}{2} + C_{2.140,1} (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\lfloor \frac{p-1}{2} \rfloor + 1} n_{\text{max}}^2 n_{\text{min}}^{-p} \right. \\ &\quad \left. + \frac{1}{2} C_{2.140,2}^2 n_{\text{min}}^{-2m} + \frac{1}{2} C_{2.143}^2 n_{\text{min}}^{-2m} \right) \|\theta_b\|_{m-1,b,n}^2 \\ &\quad + \frac{k^2}{2} \left\| Q_{b,n} \rho \right\|_{m-1,b,n}^2 + b^{2(1-m)} \|\rho\|_{H_k^1(D)}^2, \end{aligned}$$

and by (2.83)

$$\begin{aligned} \|\theta_b\|_{m,b,n} &\leq k \left( \sqrt{\frac{3}{2}} + C_{2.140,1}^{\frac{1}{2}} (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\frac{1}{2} (\lfloor \frac{p-1}{2} \rfloor + 1)} n_{\text{max}} n_{\text{min}}^{-\frac{p}{2}} \right. \\ &\quad \left. + \frac{1}{\sqrt{2}} C_{2.140,2} n_{\text{min}}^{-m} + \frac{1}{\sqrt{2}} C_{2.143} n_{\text{min}}^{-m} \right) \|\theta_b\|_{m-1,b,n} \\ &\quad + \frac{k}{\sqrt{2}} \left\| Q_{b,n} \rho \right\|_{m-1,b,n} + b^{1-m} \|\rho\|_{H_k^1(D)}. \end{aligned} \quad (2.144)$$

We now proceed to bound  $\left\| Q_{b,n} \rho \right\|_{m-1,b,n}$  : By Lemma 2.72 we have

$$\begin{aligned} \left\| Q_{b,n} \rho \right\|_{m-1,b,n} &\leq C_{\text{disc,inv},p}^{m-1} n_{\text{min}}^{1-m} b^{1-m} \left\| Q_{b,n} \rho \right\|_{L^2(D),n} \\ &\leq C_{\text{disc,inv},p}^{m-1} n_{\text{min}}^{1-m} b^{1-m} \left( \left\| Q_{b,n} u - u \right\|_{L^2(D),n} + \left\| u - \mathcal{P}_b u \right\|_{L^2(D),n} \right) \end{aligned} \quad (2.145)$$

as in the proof of Lemma 2.78, using the fact that  $Q_{b,n} \mathcal{P}_b u = \mathcal{P}_b u$ .

Using Lemma 2.63 we can bound the first of these terms by the second:

$$\left\| Q_{b,n} u - u \right\|_{L^2(D),n} \leq C_{\text{BA},0} n_{\text{var}} \left\| u - \mathcal{P}_b u \right\|_{L^2(D),n}. \quad (2.146)$$

We can also bound

$$\left\| u - \mathcal{P}_b u \right\|_{L^2(D),n} \leq C_{\text{weight},0} n_{\text{var}} b \|\rho\|_{H^1(D),n} \leq C_{\text{weight},0} n_{\text{var}} n_{\text{max}} b \|\rho\|_{H_k^1(D)} \quad (2.147)$$

by Lemma 2.61. Therefore by (2.145)–(2.147) we have (as  $kh \leq 1$ )

$$k \left\| Q_{b,n} \rho \right\|_{m-1,b,n} \leq C_{\text{disc,inv},p}^{m-1} C_{\text{weight},0} (1 + C_{\text{BA},0}) n_{\text{var}}^2 n_{\text{max}} n_{\text{min}}^{1-m} h^{1-m} \|\rho\|_{H_k^1(D)}. \quad (2.148)$$

Therefore using (2.144) and (2.148) (and the fact that  $n_{\text{max}}, n_{\text{min}}^{-1} \geq 1$ , and so  $\mathcal{C}_{\text{err}}(n), n_{\text{var}} \geq 1$  and  $n_{\text{min}}^{-p/2}, n_{\text{min}}^{-m} \leq n_{\text{min}}^{-p}$ ) we obtain

$$\begin{aligned} \|\theta_b\|_{m,b,n} &\leq C_{2.149,1} \left( 1 + (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\frac{1}{2}(\lfloor \frac{p-1}{2} \rfloor + 1)} n_{\text{max}} n_{\text{min}}^{-\frac{p}{2}} \right) k \|\theta_b\|_{m-1,b,n} \\ &\quad + C_{2.149,2,m} n_{\text{var}}^2 n_{\text{max}} n_{\text{min}}^{1-m} h^{1-m} \|\rho\|_{H_k^1(D)}. \end{aligned} \quad (2.149)$$

Using (2.149) recursively, and the facts that  $n_{\text{min}}^{-1} \geq 1$  and  $kh \leq 1$ , we obtain (2.137).  $\square$

The following lemma is straightforward to prove, and is used in the proof of Lemma 2.81 below.

**Lemma 2.80** (Continuity of  $a_T$ ). *For any  $v_1, v_2 \in H_{0,D}^1(D)$ ,*

$$|a_T(v_1, v_2)| \leq C_C n_{\text{max}} \|v_1\|_{H_k^1(D)} \|v_2\|_{H_k^1(D)}.$$

Note, we keep  $n_{\text{max}}$  out of the definition of the continuity constant  $C_C$  so that we can explicitly keep track of how all the constants in this section depend on  $n$ .

**Lemma 2.81** (Bound on  $\|\theta_b\|_{L^2(D)}$  by  $\|\theta_b\|_{p-1,b,n}$ ). *Under the assumptions of Theorem 2.39,*

$$\begin{aligned} \|\theta_b\|_{L^2(D)} &\leq \left( C_{2.150,1} \|\rho\|_{H_k^1(D)} + C_{2.150,2} k^2 h^p \|\theta_b\|_{p-1,b,n} \right) (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\lfloor \frac{p-1}{2} \rfloor + 1} n_{\text{max}}^2 \\ &\quad P_{p-2}(n) (C_{\text{FEM},1} h + C_{\text{FEM},2} C_{\text{stab}} (hk)^p). \end{aligned} \quad (2.150)$$

*Proof of Lemma 2.81.* The proof initially uses the standard duality technique, but then becomes more complex than standard proofs. This complexity is due to the facts that: (i) we are bounding  $\theta_b$ , not the finite-element error  $u - u_b$ , and (ii) we are bounding  $\theta_b$  by its higher-order-discrete norms, rather than by its  $H^1$ -norm as in the Aubin–Nitsche argument.

Consider the adjoint variational problem: Find  $w \in H_{0,D}^1(D)$  such that for all  $v \in H_{0,D}^1(D)$

$$a_T(v, w) = (v, \theta_b)_{L^2(D)}. \quad (2.151)$$

(I.e.,  $w$  solves the adjoint problem (2.60) with right-hand side given by  $\theta_b$ .)

Let  $e_b := u - u_b$  be the finite-element error, put  $v = e_b$  in (2.151) and take the complex conjugate<sup>23</sup> to obtain

$$\begin{aligned} (\theta_b, e_b)_{L^2(D)} &= \overline{a_T(e_b, w - \mathcal{P}_b w)} \\ &= a_*(w - \mathcal{P}_b w, e_b) - k^2(w - \mathcal{P}_b w, e_b)_{L^2(D),n} + ik(w - \mathcal{P}_b w, e_b)_{L^2(\Gamma_I)} \end{aligned}$$

(recalling the definition of  $a_*$  in (2.88)). By Galerkin orthogonality for  $w - \mathcal{P}_b w$ , we have (recalling  $e_b = \rho + \theta_b$ )

$$\begin{aligned} (\theta_b, e_b)_{L^2(D)} &= a_*(w - \mathcal{P}_b w, \rho) - k^2(w - \mathcal{P}_b w, e_b)_{L^2(D),n} + ik(w - \mathcal{P}_b w, e_b)_{L^2(\Gamma_I)} \\ &= \overline{a_T(\rho, w - \mathcal{P}_b w)} - k^2(w - \mathcal{P}_b w, \theta_b)_{L^2(D),n} - ik(w - \mathcal{P}_b w, \theta_b)_{L^2(\Gamma_I)}. \end{aligned} \quad (2.152)$$

Therefore since  $\theta_b = e_b - \rho$  we can rearrange (2.152) and use the Cauchy-Schwarz inequality to obtain

$$\begin{aligned} \|\theta_b\|_{L^2(D)}^2 &\leq C_C n_{\max} \|\rho\|_{H_k^1(D)} \|w - \mathcal{P}_b w\|_{H_k^1(D)} + k^2 \left| (w - \mathcal{P}_b w, \theta_b)_{L^2(D),n} \right| \\ &\quad + k \left| (w - \mathcal{P}_b w, \theta_b)_{L^2(\Gamma_I)} \right| + \|\rho\|_{L^2(D)} \|\theta_b\|_{L^2(D)} \end{aligned} \quad (2.153)$$

By combining Lemmas 2.56 and 2.60, we can show (since  $w$  satisfies an adjoint Helmholtz problem with right-hand side  $\theta_b$ )

$$\|w - \mathcal{P}_b w\|_{L^2(D)} \leq C_{\text{proj},0} P_{p-2}(n) (C_{\text{FEM},1} h^2 + C_{\text{FEM},2} C_{\text{stab}} h (hk)^p) \|\theta_b\|_{L^2(D)} \quad (2.154)$$

and

$$\|w - \mathcal{P}_b w\|_{H^1(D)} \leq 2C_{\text{proj},-1} P_{p-2}(n) (C_{\text{FEM},1} h + C_{\text{FEM},2} C_{\text{stab}} (hk)^p) \|\theta_b\|_{L^2(D)}. \quad (2.155)$$

We will be able to use (2.154) and (2.155) to bound the terms involving  $w - \mathcal{P}_b w$  in (2.153).

---

<sup>23</sup>The reason we take the complex conjugate is that to apply Galerkin orthogonality for  $w - \mathcal{P}_b w$ , the term  $w - \mathcal{P}_b w$  must be the first argument of  $a_T$ . Alternatively, one could define  $\widetilde{\mathcal{P}}_b$  to be the analogue of the elliptic projection but defined in the second argument, show analogues of the error bounds in Lemma 2.60 and proceed with the proof of the current lemma. However, for simplicity, we instead take the complex conjugate of (2.151).

We first estimate the inner product terms in (2.153):

$$\begin{aligned}
\left| (\mathcal{W} - \mathcal{P}_b \mathcal{W}, \theta_b)_{L^2(D),n} \right| &= \left| (Q_{b,n} \mathcal{W} - \mathcal{P}_b \mathcal{W}, \theta_b)_{L^2(D),n} \right| \\
&\leq \|\theta_b\|_{p-1,b,n} \left\| Q_{b,n} \mathcal{W} - \mathcal{P}_b \mathcal{W} \right\|_{-(p-1),b,n} \text{ by Lemma 2.69,} \\
&\leq \|\theta_b\|_{p-1,b,n} C_{2.110,p-1} (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\lfloor \frac{p-1}{2} \rfloor} n_{\text{max}} \\
&\quad \sum_{m=0}^{p-1} h^m \left( \left\| Q_{b,n} \mathcal{W} - \mathcal{W} \right\|_{H^{-(p-1-m)}(D),n} + \|\mathcal{W} - \mathcal{P}_b \mathcal{W}\|_{H^{-(p-1-m)}(D),n} \right) \\
&\quad \text{by Lemma 2.75,} \\
&\leq \|\theta_b\|_{p-1,b,n} C_{2.110,p-1} (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\lfloor \frac{p-1}{2} \rfloor} n_{\text{max}} h^{p-1} \\
&\quad \sum_{m=0}^{p-1} \left( C_{\text{BA},p-1-m} n_{\text{var}} \|\mathcal{W} - \mathcal{P}_b \mathcal{W}\|_{L^2(D),n} \right. \\
&\quad \quad \left. + C_{\text{weight},p-1-m} \mathcal{C}_{\text{err},p-1-m}(n) h \|\mathcal{W} - \mathcal{P}_b \mathcal{W}\|_{H^1(D),n} \right) \\
&\quad \text{by Lemmas 2.63 and 2.61,} \\
&\quad \text{taking } \mathcal{W}_b = \mathcal{P}_b u \text{ in Lemma 2.61 and (2.101),} \\
&\leq 2 \|\theta_b\|_{p-1,b,n} C_{2.110,p-1} \sum_{m=0}^{p-1} \left( C_{\text{BA},p-1-m} C_{\text{proj},0} + C_{\text{weight},p-1-m} C_{\text{proj},-1} \right) \\
&\quad (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\lfloor \frac{p-1}{2} \rfloor + 1} n_{\text{max}}^2 h^p P_{p-2}(n) \\
&\quad (C_{\text{FEM},1} h + C_{\text{FEM},2} C_{\text{stab}}(hk)^p) \|\theta_b\|_{L^2(D)} \text{ by (2.154) and (2.155),} \\
&= \|\theta_b\|_{p-1,b,n} C_{2.156} (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\lfloor \frac{p-1}{2} \rfloor + 1} n_{\text{max}}^2 h^p P_{p-2}(n) \\
&\quad (C_{\text{FEM},1} h + C_{\text{FEM},2} C_{\text{stab}}(hk)^p) \|\theta_b\|_{L^2(D)}. \tag{2.156}
\end{aligned}$$

We now estimate the other inner product term

$$\begin{aligned}
\left| (\theta_b, \mathcal{W} - \mathcal{P}_b \mathcal{W})_{L^2(\Gamma_I)} \right| &\leq C_{\text{MT}} C_{\text{inv},p} \left( C_{2.129,1}^{\frac{1}{2}} (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\lfloor \frac{p-1}{2} \rfloor + 1} n_{\text{max}}^2 k h^{p-\frac{1}{2}} \|\theta_b\|_{p-1,b,n} \right. \\
&\quad \left. + C_{2.129,2}^{\frac{1}{2}} h^{\frac{1}{2}} \|\rho\|_{H_k^1(D)} \right) h^{-\frac{1}{2}} \|\mathcal{W} - \mathcal{P}_b \mathcal{W}\|_{L^2(D)} \\
&\quad \text{by Lemma 2.78, (2.83), Theorem 2.57, and Lemma 2.71,} \\
&\leq C_{\text{MT}} C_{\text{inv},p} C_{\text{proj},0} \left( C_{2.129,1}^{\frac{1}{2}} (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\lfloor \frac{p-1}{2} \rfloor + 1} n_{\text{max}}^2 k h^p \|\theta_b\|_{p-1,b,n} \right. \\
&\quad \quad \left. + C_{2.129,2}^{\frac{1}{2}} h \|\rho\|_{H_k^1(D)} \right) P_{p-2}(n) \\
&\quad (C_{\text{FEM},1} h + C_{\text{FEM},2} C_{\text{stab}}(hk)^p) \|\theta_b\|_{L^2(D)} \tag{2.157}
\end{aligned}$$

by (2.154).

We now insert (2.154)–(2.157) into (2.153):

$$\begin{aligned}
\|\theta_b\|_{L^2(D)}^2 &\leq \left[ C_C n_{\max} \|\rho\|_{H_k^1(D)} (C_{\text{proj},0} + 2C_{\text{proj},1}) \right. \\
&\quad + k^2 \|\theta_b\|_{p-1,b,n} C_{2.156} (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\lfloor \frac{p-1}{2} \rfloor + 1} n_{\max}^2 h^p \\
&\quad + k C_{\text{MT}} C_{\text{inv},p} C_{\text{proj},0} \left( C_{2.129,1}^{\frac{1}{2}} (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\lfloor \frac{p-1}{2} \rfloor + 1} n_{\max}^2 k h^p \|\theta_b\|_{p-1,b,n} \right. \\
&\quad \quad \left. + C_{2.129,2}^{\frac{1}{2}} b \|\rho\|_{H_k^1(D)} \right) \left. \right] \\
&\quad P_{p-2}(n) (C_{\text{FEM},1} b + C_{\text{FEM},2} C_{\text{stab}} (hk)^p) \|\theta_b\|_{L^2(D)} \\
&\quad + \|\rho\|_{L^2(D)} \|\theta_b\|_{L^2(D)} \\
&\leq \left[ \left( C_C (C_{\text{proj},0} + 2C_{\text{proj},1}) n_{\max} + C_{\text{MT}} C_{\text{inv},p} C_{\text{proj},0} C_{2.129,2}^{\frac{1}{2}} + \frac{C_{\text{proj},0}}{C_{\text{FEM},1}} n_{\text{var}} \right) \|\rho\|_{H_k^1(D)} \right. \\
&\quad + \left( C_{2.156} (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\lfloor \frac{p-1}{2} \rfloor + 1} n_{\max}^2 \right. \\
&\quad \left. + C_{\text{MT}} C_{\text{inv},p} C_{\text{proj},0} C_{2.129,1}^{\frac{1}{2}} (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\lfloor \frac{p-1}{2} \rfloor + 1} n_{\max}^2 \right) k^2 h^p \|\theta_b\|_{p-1,b,n} \left. \right] \\
&\quad P_{p-2}(n) (C_{\text{FEM},1} b + C_{\text{FEM},2} C_{\text{stab}} (hk)^p) \|\theta_b\|_{L^2(D)}. \tag{2.158}
\end{aligned}$$

Rearranging (2.158) and using Lemma 2.60 and the fact that  $hk \leq 1$  we have

$$\begin{aligned}
\|\theta_b\|_{L^2(D)}^2 &\leq \left[ \left( C_C (C_{\text{proj},0} + 2C_{\text{proj},1}) + C_{\text{MT}} C_{\text{inv},p} C_{\text{proj},0} C_{2.129,2}^{\frac{1}{2}} + \frac{C_{\text{proj},0}}{C_{\text{FEM},1}} \right) \|\rho\|_{H_k^1(D)} \right. \\
&\quad + \left( C_{2.156} + C_{\text{MT}} C_{\text{inv},p} C_{\text{proj},0} C_{2.129,1}^{\frac{1}{2}} \right) k^2 h^p \|\theta_b\|_{p-1,b,n} \left. \right] \\
&\quad (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\lfloor \frac{p-1}{2} \rfloor + 1} n_{\max}^2 P_{p-2}(n) (C_{\text{FEM},1} b + C_{\text{FEM},2} C_{\text{stab}} (hk)^p) \|\theta_b\|_{L^2(D)}. \tag{2.159}
\end{aligned}$$

Using Young's inequality (2.84) to separate out the  $\|\theta_b\|_{L^2(D)}$  term on the right-hand side of (2.159) and then moving the resulting  $\|\theta_b\|_{L^2(D)}^2$  term to the left hand side, followed by (2.83), we obtain

$$\begin{aligned}
\|\theta_b\|_{L^2(D)} &\leq \left[ \left( C_C (C_{\text{proj},0} + 2C_{\text{proj},1}) + C_{\text{MT}} C_{\text{inv},p} C_{\text{proj},0} C_{2.129,2}^{\frac{1}{2}} + \frac{C_{\text{proj},0}}{C_{\text{FEM},1}} \right) \|\rho\|_{H_k^1(D)} \right. \\
&\quad + \left( C_{2.156} + C_{\text{MT}} C_{\text{inv},p} C_{\text{proj},0} C_{2.129,1}^{\frac{1}{2}} \right) k^2 h^p \|\theta_b\|_{p-1,b,n} \left. \right] \\
&\quad (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\lfloor \frac{p-1}{2} \rfloor + 1} n_{\max}^2 P_{p-2}(n) (C_{\text{FEM},1} b + C_{\text{FEM},2} C_{\text{stab}} (hk)^p). \tag{2.160}
\end{aligned}$$

Upon rearranging (2.160) we obtain (2.150).  $\square$

With all our lemmas proved, we can now prove our main theorem.

*Proof of Theorem 2.39.* By inserting (2.137) (with  $m = p - 1$ ) into (2.150) and using the fact that  $n_{\max}, \mathcal{C}_{\text{err}}(n), n_{\text{var}} \geq 1$ , we have

$$\begin{aligned} \|\theta_b\|_{L^2(D)} &\leq \left( C_{2.150,1} + C_{2.150,2} k^2 h^p C_{2.137,p-1,2} \left( (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\frac{1}{2}(\lfloor \frac{p-1}{2} \rfloor + 1)} n_{\max} n_{\min}^{-\frac{p}{2}} \right)^{p-1} \right. \\ &\quad \left. n_{\text{var}}^2 n_{\max} n_{\min}^{1-(p-1)} h^{1-(p-1)} \right) \\ &\quad (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\lfloor \frac{p-1}{2} \rfloor + 1} n_{\max}^2 P_{p-2}(n) (C_{\text{FEM},1} h + C_{\text{FEM},2} C_{\text{stab}} (hk)^p) \|\rho\|_{H_k^1(D)} \\ &+ C_{2.150,2} k^2 h^p C_{2.137,p-1,1} \left( (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\frac{1}{2}(\lfloor \frac{p-1}{2} \rfloor + 1)} n_{\max} n_{\min}^{-\frac{p}{2}} \right)^{p-1} k^{p-1} \\ &\quad (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\lfloor \frac{p-1}{2} \rfloor + 1} n_{\max}^2 P_{p-2}(n) (C_{\text{FEM},1} h + C_{\text{FEM},2} C_{\text{stab}} (hk)^p) \|\theta_b\|_{L^2(D)}, \end{aligned} \quad (2.161)$$

and observe that the second summand in (2.161) is equal to

$$\begin{aligned} &C_{2.150,2} C_{2.137,p-1,1} \left( (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\frac{1}{2}(\lfloor \frac{p-1}{2} \rfloor + 1)} n_{\max} n_{\min}^{-\frac{p}{2}} \right)^{p-1} \\ &(\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\lfloor \frac{p-1}{2} \rfloor + 1} n_{\max}^2 P_{p-2}(n) (C_{\text{FEM},1} (hk)^{p+1} + C_{\text{FEM},2} C_{\text{stab}} h^{2p} k^{2p+1}) \|\theta_b\|_{L^2(D)} \end{aligned}$$

Choosing  $h$  according to (2.62), (2.161) simplifies to

$$\begin{aligned} \|\theta_b\|_{L^2(D)} &\leq \left( C_{2.150,1} + C_{2.150,2} k^2 h^p C_{2.137,p-1,2} \left( (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\frac{1}{2}(\lfloor \frac{p-1}{2} \rfloor + 1)} n_{\max} n_{\min}^{-\frac{p}{2}} \right)^{p-1} \right. \\ &\quad \left. n_{\text{var}}^2 n_{\max} n_{\min}^{1-(p-1)} h^{1-(p-1)} \right) \\ &\quad (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\lfloor \frac{p-1}{2} \rfloor + 1} n_{\max}^2 P_{p-2}(n) (C_{\text{FEM},1} h + C_{\text{FEM},2} C_{\text{stab}} (hk)^p) \|\rho\|_{H_k^1(D)} \\ &+ \frac{1}{2} \|\theta_b\|_{L^2(D)}, \end{aligned}$$

and therefore (since  $k^2 h^p h^{1-(p-1)} = k^2 h^2 \leq 1$ ) it follows that

$$\begin{aligned} \|\theta_b\|_{L^2(D)} &\leq \left( (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\frac{1}{2}(\lfloor \frac{p-1}{2} \rfloor + 1)} n_{\max} n_{\min}^{-\frac{p}{2}} \right)^{p-1} n_{\text{var}}^5 n_{\min}^{-(p+1)} (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\frac{1}{2}(\lfloor \frac{p-1}{2} \rfloor + 1)} P_{p-2}(n) \\ &\quad (C_{2.162,1} h + C_{2.162,2} C_{\text{stab}} (hk)^p) \|\rho\|_{H_k^1(D)} \end{aligned} \quad (2.162)$$

$$= \left( \frac{\mathcal{C}_{L^2}(n)}{P_{p-2}(n)} - 1 \right) (C_{2.162,1} h + C_{2.162,2} C_{\text{stab}} (hk)^p) \|\rho\|_{H_k^1(D)} \quad (2.163)$$

(recall the definition of  $\mathcal{C}_{L^2}(n)$  from (2.65)).

By Lemmas 2.56 and 2.60 (using (2.83) and the fact that  $k \geq 1$ )

$$\|\rho\|_{H_k^1(D)} \leq C_{2.164} P_{p-2}(n) (h + C_{\text{stab}} (hk)^p) C_{f,g_r}, \quad (2.164)$$

and therefore

$$\|\theta_b\|_{L^2(D)} \leq (\mathcal{C}_{L^2}(n) - P_{p-2}(n)) C_{2.165} (h^2 + C_{\text{stab}} h(hk)^p + C_{\text{stab}}^2 (hk)^{2p}) C_{f, g_I}. \quad (2.165)$$

We can now bound  $\|u - u_b\|_{L^2(D)}$ :

$$\begin{aligned} \|u - u_b\|_{L^2(D)} &\leq \|\rho\|_{L^2(D)} + \|\theta_b\|_{L^2(D)} \\ &\leq C_{\text{proj},0} P_{p-2}(n) (C_{\text{FEM},1} h^2 + C_{\text{FEM},2} C_{\text{stab}} h(hk)^p) C_{f, g_I} + \|\theta_b\|_{L^2(D)}, \\ &\quad \text{by Lemmas 2.56 and 2.60,} \\ &\leq C_{\text{proj},0} \max\{C_{\text{FEM},1}, C_{\text{FEM},2}\} P_{p-2}(n) (h^2 + C_{\text{stab}} h(hk)^p + C_{\text{stab}}^2 (hk)^{2p}) C_{f, g_I} \\ &\quad + (\mathcal{C}_{L^2}(n) - P_{p-2}(n)) C_{2.165} (h^2 + C_{\text{stab}} h(hk)^p + C_{\text{stab}}^2 (hk)^{2p}) C_{f, g_I}, \end{aligned}$$

which gives (2.63), as required.

We can now proceed similarly as above to bound  $|\theta_b|_{H^1(D)}$ , and hence to bound the error  $\|u - u_b\|_{H_k^1(D)}$ . By Lemmas 2.73 and 2.79 (with  $m = 1$ ) we have

$$\begin{aligned} |\theta_b|_{H^1(D)} &\leq A_{\min}^{\frac{1}{2}} \left[ C_{2.137,1,1} \left( (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\frac{1}{2}} (\lfloor \frac{p-1}{2} \rfloor + 1) n_{\max} n_{\min}^{-\frac{p}{2}} \right) k \|\theta_b\|_{L^2(D)} \right. \\ &\quad \left. + C_{2.137,1,2} \left( (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\frac{1}{2}} (\lfloor \frac{p-1}{2} \rfloor + 1) n_{\max} n_{\min}^{-\frac{p}{2}} \right) n_{\text{var}}^2 n_{\max} \|\rho\|_{H_k^1(D)} \right], \quad (2.166) \end{aligned}$$

and by combining (2.163) and (2.166) we obtain (since  $hk \leq 1$ )

$$\begin{aligned} |\theta_b|_{H^1(D)} &\leq C_{2.167} \left[ \left( (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\frac{1}{2}} (\lfloor \frac{p-1}{2} \rfloor + 1) n_{\max} n_{\min}^{-\frac{p}{2}} \right) \right. \\ &\quad \left. \left( \frac{\mathcal{C}_{L^2}(n)}{P_{p-2}(n)} - 1 \right) (1 + C_{\text{stab}} k (hk)^p) \right. \\ &\quad \left. + \left( (\mathcal{C}_{\text{err}}(n) n_{\text{var}})^{\frac{1}{2}} (\lfloor \frac{p-1}{2} \rfloor + 1) n_{\max} n_{\min}^{-\frac{p}{2}} \right) n_{\text{var}}^2 n_{\max} \right] \|\rho\|_{H_k^1(D)}. \quad (2.167) \end{aligned}$$

Substituting (2.164) into (2.167) we have

$$\begin{aligned}
|\theta_b|_{H^1(D)} &\leq C_{2.167}C_{2.164} \left[ \left( (\mathcal{C}_{\text{err}}(n)n_{\text{var}})^{\frac{1}{2}(\lfloor \frac{p-1}{2} \rfloor + 1)} n_{\text{max}} n_{\text{min}}^{-\frac{p}{2}} \right) \left( \frac{\mathcal{C}_{L^2}(n)}{P_{p-2}(n)} - 1 \right) \right. \\
&\quad \left. + \left( (\mathcal{C}_{\text{err}}(n)n_{\text{var}})^{\frac{1}{2}(\lfloor \frac{p-1}{2} \rfloor + 1)} n_{\text{max}} n_{\text{min}}^{-\frac{p}{2}} \right) n_{\text{var}}^2 n_{\text{max}} \right] P_{p-2}(n) \\
&\quad (1 + C_{\text{stab}}k(hk)^p)(h + C_{\text{stab}}(hk)^p)C_{f,g_I} \\
&\leq 2C_{2.167}C_{2.164} \left[ \left( (\mathcal{C}_{\text{err}}(n)n_{\text{var}})^{\frac{1}{2}(\lfloor \frac{p-1}{2} \rfloor + 1)} n_{\text{max}} n_{\text{min}}^{-\frac{p}{2}} \right) \left( \frac{\mathcal{C}_{L^2}(n)}{P_{p-2}(n)} - 1 \right) \right. \\
&\quad \left. + \left( (\mathcal{C}_{\text{err}}(n)n_{\text{var}})^{\frac{1}{2}(\lfloor \frac{p-1}{2} \rfloor + 1)} n_{\text{max}} n_{\text{min}}^{-\frac{p}{2}} \right) n_{\text{var}}^2 n_{\text{max}} \right] P_{p-2}(n) \\
&\quad (h + C_{\text{stab}}(hk)^p + C_{\text{stab}}^2 k(hk)^{2p})C_{f,g_I}. \tag{2.168}
\end{aligned}$$

We can now proceed to bound  $\|u - u_b\|_{H_k^1(D)}$ :

$$\begin{aligned}
\|u - u_b\|_{H_k^1(D)} &\leq \|\rho\|_{H_k^1(D)} + \|\theta_b\|_{H_k^1(D)} \\
&\leq \|\rho\|_{H^1(D)} + k\|\rho\|_{L^2(D)} + |\theta_b|_{H^1(D)} + k\|\theta_b\|_{L^2(D)} \\
&\leq 2(C_{\text{proj},-1} + C_{\text{proj},0})P_{p-2}(n)(C_{\text{FEM},1}h + C_{\text{FEM},2}C_{\text{stab}}(hk)^p)C_{f,g_I} \\
&\quad + |\theta_b|_{H^1(D)} + k\|\theta_b\|_{L^2(D)}, \text{ by Lemmas 2.56 and 2.60,} \\
&\leq 2(C_{\text{proj},-1} + C_{\text{proj},0})P_{p-2}(n)(C_{\text{FEM},1}h + C_{\text{FEM},2}C_{\text{stab}}(hk)^p)C_{f,g_I} \\
&\quad + 2C_{2.167}C_{2.164} \left[ \left( (\mathcal{C}_{\text{err}}(n)n_{\text{var}})^{\frac{1}{2}(\lfloor \frac{p-1}{2} \rfloor + 1)} n_{\text{max}} n_{\text{min}}^{-\frac{p}{2}} \right) \left( \frac{\mathcal{C}_{L^2}(n)}{P_{p-2}(n)} - 1 \right) \right. \\
&\quad \left. + \left( (\mathcal{C}_{\text{err}}(n)n_{\text{var}})^{\frac{1}{2}(\lfloor \frac{p-1}{2} \rfloor + 1)} n_{\text{max}} n_{\text{min}}^{-\frac{p}{2}} \right) n_{\text{var}}^2 n_{\text{max}} \right] P_{p-2}(n) \\
&\quad (h + C_{\text{stab}}(hk)^p + C_{\text{stab}}^2 k(hk)^{2p})C_{f,g_I} \\
&\quad + (\mathcal{C}_{L^2}(n) - P_{p-2}(n))C_{2.165}k(h^2 + C_{\text{stab}}h(hk)^p + C_{\text{stab}}^2(hk)^{2p})C_{f,g_I},
\end{aligned}$$

by (2.165) and (2.168). Using the fact that  $hk \leq 1$ , and the definitions of  $\mathcal{C}_{H^1}(n)$  and  $C_{\text{FEM},H^1}$ , we then obtain (2.64), as required.  $\square$

### 2.4.7 Constants from Section 2.4

To summarise the constants used in Section 2.4, we use the following table, where the constants are given in order of appearance in the text. As well as giving the definitions of the constants, we also state the place (Theorem, Lemma, etc.) where they are defined. If a constant is not defined in terms of other constants, but rather is given in the statement of a Theorem or Lemma (as for, e.g., the definition of  $C_{\text{MT}}$  in Theorem 2.57), then the ‘Definition’ column is left blank.

We recall that where a constant is only used inside a proof it will usually be numbered using

Table 2.4: The constants from Section 2.4

| Constant                      | Definition                                                        | Defined/Introduced    |
|-------------------------------|-------------------------------------------------------------------|-----------------------|
| $C_{BA,m}$                    | —                                                                 | Lemma 2.22            |
| $C_{\text{stab}}$             | —                                                                 | Assumption 2.36       |
| $C_{f,g_l}$                   | —                                                                 | Assumption 2.36       |
| $C_{\text{int},A,l}$          | —                                                                 | Proof of Theorem 2.51 |
| $C_{\text{scat},A,l}$         | —                                                                 | Proof of Theorem 2.51 |
| $C_{\text{trunc},A,l}$        | —                                                                 | Proof of Theorem 2.51 |
| $C_{A,l}$                     | $C_{\text{int},A,l} + C_{\text{scat},A,l} + C_{\text{trunc},A,l}$ | Theorem 2.51          |
| $C_{\text{Tr},m}$             | —                                                                 | Theorem 2.54          |
| $C_{H^m,\text{prod}}$         | —                                                                 | Proof of Theorem 2.55 |
| $C_{\text{mult},m,\tilde{m}}$ | —                                                                 | Theorem 2.55          |

Table 2.4: (continued)

| Constant                 | Definition                                                                                                                                                                                                                                              | Defined/Introduced    |
|--------------------------|---------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| $C_{\text{mult}}$        | $\max_{m=2,\dots,p-2} C_{\text{mult},m,\max\{p-1,[d/2]+1\}}$                                                                                                                                                                                            | Proof of Theorem 2.49 |
| $C_{\text{expansion},j}$ | $\begin{cases} \max C_{A,0}, & j = 0 \\ \max\{1, C_{A,1}\}(1 + C_{\text{Tr},2})C_{\text{expansion},0}, & j = 1 \\ C_{A,j}(1 + C_{\text{Tr},j+1})\max\{C_{\text{mult}}C_{\text{expansion},j-2}, C_{\text{expansion},j-1}\} & j \in [2, p-2] \end{cases}$ | Theorem 2.49          |
| $C_{\text{rem},j}$       | $\begin{cases} C_{A,1}(1 + C_{\text{Tr},2}), & j = 1 \\ C_{A,2}(1 + C_{\text{Tr},3}(1 + C_{\text{rem},1})), & j = 2 \\ C_{A,j}(1 + C_{\text{Tr},j+1}(C_{\text{rem},j-1} + C_{\text{rem},j-1})), & j = 3, \dots, p-1 \end{cases}$                        | Proof of Theorem 2.49 |
| $C_{\text{osc}}$         | $C_{\text{rem},p-1}$                                                                                                                                                                                                                                    | Theorem 2.49          |
| $C_{\text{FEM},1}$       | $\sum_{j=0}^{p-2} C_{\text{BA},j+2}C_{\text{expansion},j}$                                                                                                                                                                                              | Lemma 2.56            |
| $C_{\text{FEM},2}$       | $C_{\text{FEM},2} = C_{\text{BA},p+1}C_{\text{osc}}$                                                                                                                                                                                                    | Lemma 2.56            |

Table 2.4: (continued)

| Constant                | Definition                                                                                                                                                                                                                                                                                    | Defined/Introduced  |
|-------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|---------------------|
| $C_{\text{proj},-m}$    | $\begin{cases} \frac{2\ A\ _{L^\infty(D;\text{op})}}{\min\{1/\frac{1}{C_p^2}\}A_{\min}}, & m = -1 \\ \ A\ _{L^\infty(D;\text{op})}C_{\text{BA},2}C_{A,0}C_{\text{proj},-1}, & m = 0 \\ C_{\text{BA},m+2}C_{A,m}\ A\ _{L^\infty(D;\text{op})}C_{\text{proj},-1}, & m \in [1, p-1] \end{cases}$ | Lemma 2.60          |
| $C_{\text{weight},-m}$  | $\begin{cases} C_{\text{proj},m}, & m = -1, 0 \\ C_{\text{mult},m,\max\{p-1,[d/2]+1\}}C_{\text{proj},m}, & m \in [1, p-1] \end{cases}$                                                                                                                                                        | Lemma 2.60          |
| $C_{\text{inv},p}$      | —                                                                                                                                                                                                                                                                                             | Lemma 2.71          |
| $C_{\text{disc,inv},p}$ | $C_{\text{inv},p}\ A\ _{L^\infty(D;\text{op})}^{1/2}$                                                                                                                                                                                                                                         | Lemma 2.72          |
| $C_{\text{shift},-m}$   | $\begin{cases} 1/A_{\min}, & m = -1 \\ C_{A,0}, & m = 0 \\ 2C_{A,m}C_{\text{mult},m,\max\{p-1,[d/2]+1\}}, & m \in [1, p-1] \end{cases}$                                                                                                                                                       | Lemma 2.74          |
| $C_{2.111,m}$           | $C_{\text{weight},-m,1}C_{A,0}C_{\text{BA},2}$                                                                                                                                                                                                                                                | Proof of Lemma 2.75 |
| $C_{2.113,m,t}$         | See (2.114)–(2.116)                                                                                                                                                                                                                                                                           | Proof of Lemma 2.75 |

Table 2.4: (continued)

| Constant        | Definition                                                                                                                                          | Defined/Introduced  |
|-----------------|-----------------------------------------------------------------------------------------------------------------------------------------------------|---------------------|
| $C_{2.120,m,t}$ | See (2.121)–(2.123)                                                                                                                                 | Proof of Lemma 2.75 |
| $C_{2.110,j}$   | $\max_{m \in \{0, \dots, l\}} \left\{ C_{2.113,m,l}, \frac{(1 + C_{\text{proj},1}) \ A\ _{L^\infty(D;\text{op})}}{A_{\min}} C_{2.120,m,l} \right\}$ | Proof of Lemma 2.75 |
| $C_{2.133}$     | $C_{2.110,p-1} \left( \sum_{m=0}^{p-1} C_{\text{BA},(p-1-m)} C_{\text{weight},0} + C_{\text{weight},-(p-1-m)} \right)$                              | Proof of Lemma 2.78 |
| $C_{\text{MT}}$ | —                                                                                                                                                   | Theorem 2.57        |
| $C_{2.129,1}$   | $C_{2.133}^2$                                                                                                                                       | Lemma 2.78          |
| $C_{2.129,2}$   | $(1 + C_{\text{MT}}^2 C_{\text{proj},0})/2$                                                                                                         | Lemma 2.78          |
| $C_{2.140,1}$   | $C_{\text{MT}} C_{\text{inv},p}^{1/2} C_{\text{disc,inv},p}^{p-1} C_{2.129,1}^{1/2}$                                                                | Proof of Lemma 2.79 |
| $C_{2.140,2}$   | $C_{\text{MT}} C_{\text{inv},p}^{1/2} C_{\text{disc,inv},p}^{m-1} C_{2.129,2}^{1/2}$                                                                | Proof of Lemma 2.79 |
| $C_{2.143}$     | $C_{\text{MT}}^2 C_{\text{inv},p}^{1/2} C_{\text{disc,inv},p}^{m-1} C_{\text{proj},0}^{1/2}$                                                        | Proof of Lemma 2.79 |

Table 2.4: (continued)

| Constant        | Definition                                                                                                                                                               | Defined/Introduced    |
|-----------------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| $C_{2.149,1}$   | $\sqrt{\frac{3}{2}} + C_{2.140,1}^{\frac{1}{2}} + \frac{C_{2.140,2}}{\sqrt{2}} + \frac{C_{2.143}}{\sqrt{2}}$                                                             | Proof of Lemma 2.79   |
| $C_{2.149,2,m}$ | $\frac{1}{\sqrt{2}} C_{\text{disc,inv},p}^{m-1} C_{\text{weight},0} (1 + C_{\text{BA},0}) + 1$                                                                           | Proof of Lemma 2.79   |
| $C_{2.137,m,1}$ | $C_{2.149,1}^m$                                                                                                                                                          | Proof of Lemma 2.79   |
| $C_{2.137,m,2}$ | $\sum_{j=1}^m C_{2.149,1}^{m-j} C_{2.149,2,j}$                                                                                                                           | Proof of Lemma 2.79   |
| $C_{2.156}$     | $2C_{2.110,p-1} \sum_{m=0}^{p-1} (C_{\text{BA},p-1-m} C_{\text{proj},0} + C_{\text{weight},p-1-m} C_{\text{proj},-1})$                                                   | Proof of Lemma 2.81   |
| $C_C$           | $2 \max \left\{ \ A\ _{L^\infty(D;\text{op})}, 1, \frac{C_{\text{MT}}^2}{2} \right\}$                                                                                    | Lemma 2.80            |
| $C_{2.150,1}$   | $C_C (C_{\text{proj},0} + 2C_{\text{proj},1}) + C_{\text{MT}} C_{\text{inv},p} C_{\text{proj},0} C_{2.129,2}^{\frac{1}{2}} + \frac{C_{\text{proj},0}}{C_{\text{FEM},1}}$ | Proof of Lemma 2.81   |
| $C_{2.150,2}$   | $C_{2.156} + C_{\text{MT}} C_{\text{inv},p} C_{\text{proj},0} C_{2.129,1}^{\frac{1}{2}}$                                                                                 | Proof of Lemma 2.81   |
| $C_{2.162,1}$   | $2(C_{2.150,1} + C_{2.150,2} C_{2.137,p-1,1}) C_{\text{FEM},1}$                                                                                                          | Proof of Theorem 2.39 |

Table 2.4: (continued)

| Constant             | Definition                                                                                                                                           | Defined/Introduced    |
|----------------------|------------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| $C_{2.162,2}$        | $2(C_{2.150,1} + C_{2.150,2}C_{2.137,p-1,1})C_{\text{FEM},2}$                                                                                        | Proof of Theorem 2.39 |
| $C_{2.164}$          | $2(C_{\text{proj},-1} + C_{\text{proj},0})\max\{C_{\text{FEM},1}, C_{\text{FEM},2}\}$                                                                | (2.164)               |
| $C_{2.167}$          | $A_{\min}^{\frac{1}{2}}\max\{C_{2.137,1,1}C_{2.162,1}, C_{2.137,1,1}C_{2.162,2}, C_{2.137,1,2}\}$                                                    | Proof of Theorem 2.39 |
| $C_{2.165}$          | $2C_{2.164}\max\{C_{2.162,1}, C_{2.162,2}\}$                                                                                                         | (2.165)               |
| $C_{\text{FEM},L^2}$ | $\max\{C_{\text{proj},0}\max\{C_{\text{FEM},1}, C_{\text{FEM},2}\}, C_{2.165}\}$                                                                     | Theorem 2.39          |
| $C_{\text{FEM},H^1}$ | $\max\{2(C_{\text{proj},-1} + C_{\text{proj},0})\max\{C_{\text{FEM},1}, C_{\text{FEM},2}\}, 2C_{2.167}C_{2.164}, C_{2.165}\}$                        | Theorem 2.39          |
| $C_{\text{cond}}$    | $\frac{1}{4^{2p}}(C_{2.150,2}C_{2.137,p-1,1})^{-\frac{1}{2p}}\min\left\{C_{\text{FEM},1}^{-\frac{1}{2p}}, C_{\text{FEM},2}^{-\frac{1}{p+1}}\right\}$ | Theorem 2.39          |

the equation number of its first appearance.

## 2.5 SUMMARY AND FUTURE WORK

### 2.5.1 Summary

In this chapter we gave the requisite background theory and setup for Helmholtz problems in heterogeneous media and their finite-element discretisation, before proving new finite-element-error bounds for the Helmholtz equation in heterogeneous media. In particular:

- In Section 2.2 we gave the setup for deterministic heterogeneous Helmholtz problems, reviewed the literature around the  $k$ -dependence of a priori bounds on the solution of the Helmholtz equation, and discussed these results in the context of trapping phenomena.
- In Sections 2.3.1 and 2.3.2 we gave the setup for the finite-element discretisation of heterogeneous Helmholtz problems.
- In Section 2.3.3 we introduced concepts such as  $(hk^a, hk^b)$ -accuracy as a means of classifying results on the  $h$ - and  $k$ -dependence of finite-element discretisations of the Helmholtz equation, and gave a complete survey of the literature on rigorous quasi-optimality and error bounds.
- In Section 2.3.4 we discussed proof techniques for quasi-optimality and error bounds for finite-element discretisations of the Helmholtz equation, giving a detailed discussion and taxonomy of the proof techniques in the literature.
- Finally, in Section 2.4 we proved new finite-element error bounds for higher-order finite-element methods for the Helmholtz equation in heterogeneous media. These results are the first for higher-order methods and heterogeneous media, and are explicit in their dependence on the squared slowness  $n$ , and are sharp in their  $h$ - and  $k$ -dependence.

### 2.5.2 Future work

There are several possibilities for future work building on the new finite-element error bounds in Section 2.4.1. E.g.,

- Using simpler proof techniques, such as Modified Schatz arguments for data-accuracy to prove error bounds that (may) have a simpler  $n$ -dependence than those in Section 2.4. However, such proof techniques may not give sharp  $h$ - and  $k$ -dependence for elements of degree  $p > 1$ , see the discussions in Sections 2.3.3 and 2.3.4.
- Numerical experiments confirming the  $h$ - and  $k$ -dependence of the results in Section 2.4.
- Numerical experiments investigating whether the dependence on  $C_{\text{stab}}$  in our main result, Theorem 2.39, is as predicted, or whether this dependence is pessimistic, given one expects trapping behaviour to only be manifested for very few values of  $k$ .

# Well-posedness of formulations of the stochastic Helmholtz equation

## 3.1 INTRODUCTION

The goals of this chapter are to prove results on the well-posedness of variational formulations of the stochastic Helmholtz equation

$$\nabla \cdot (A(\omega)\nabla u(\omega)) + k^2 n(\omega)u(\omega) = -f(\omega), \quad (3.1)$$

as well as a priori bounds on its solution that are explicit in the wavenumber  $k$  and the material coefficients  $A$  and  $n$ .

We consider (3.1) with physical domain either  $\mathbb{R}^d$ ,  $d = 2, 3$ , or  $\mathbb{R}^d \setminus \overline{D_-}$ , where  $D_-$  (referred to as the *obstacle*) is as in Problem 2.1, and

- $\omega$  is an element of the underlying probability space,
- $A$  is a symmetric-positive-definite matrix-valued random field such that  $\text{supp}(I - A)$  is compact,
- $n$  is a positive real-valued random field such that  $\text{supp}(1 - n)$  is compact,
- $f$  is a real-valued random field such that  $\text{supp} f$  is compact, and
- $k > 0$  is the wavenumber,

and as in the rest of this thesis we are particularly interested in the case where the wavenumber  $k$  is large. See Section 3.1.1 below for a rigorous definition of the problems we consider.

*Motivation* The motivation for establishing well-posedness and proving a priori bounds on the solution of (3.1) is the growing interest in Uncertainty Quantification (UQ) for the Helmholtz equation; see, e.g., [220, 209, 32, 87, 80, 81, 138, 117, 14]. (In this PDE context, by ‘UQ’ we mean theory and algorithms for computing statistics of quantities of interest involving PDEs *either* posed on a random domain *or* having random coefficients.) There is a large literature on UQ for the stationary diffusion equation

$$-\nabla \cdot (\chi(\omega)\nabla u(\omega)) = f(\omega), \quad (3.2)$$

due in part to its large number of applications (e.g. in modelling groundwater flow), and a priori bounds on the solution are vital for the rigorous analysis of UQ algorithms; see e.g. [8, 7, 96, 157,

42]. In contrast, whilst (3.1) has many applications (see, e.g., Section 1.1.1 above), there is much less rigorous theory of UQ for the Helmholtz equation. The main reason for this is that the (deterministic) PDE theory of (3.1) when  $k$  is large is much more complicated than the analogous theory for (3.2).

*Related previous work* To our knowledge, the only work that considers (3.1) with large  $k$  and attempts to establish either (i) well-posedness of variational formulations or (ii) a priori bounds is [80], which considers both (i) and (ii) for (3.1) posed in a bounded domain with an impedance boundary condition. We discuss the results of [80] further in Section 3.1.4, but we highlight here that (a) [80] considers  $A = I$  and  $n = 1 + \eta$ , with  $\eta$  random and the magnitude of  $\eta$  decreasing with  $k$ , whereas we consider classes of  $A$  and  $n$  that allow  $k$ -independent random perturbations, and (b) in its well-posedness result, [80] invokes Fredholm theory to conclude existence of a solution, but this relies on an incorrect assumption about compact inclusion of Bochner spaces—see Appendix A below. In Section 3.1.4 we also discuss the papers [32, 122, 123, 117, 191, 71] on the theory of UQ for either (3.1) or the related time-harmonic Maxwell's equations; in these papers either the  $k$ -explicit well-posedness is not a primary concern or  $k$  is assumed to be small. Our hope is that the results in this chapter can be used in the rigorous theory of UQ for Helmholtz problems with large  $k$ .

*The contributions of this chapter* The main results in this chapter, Theorems 3.7 and 3.10 below, concern well-posedness and a priori bounds for the solutions of various formulations of the stochastic Helmholtz equation; these formulations include those used in sampling-based UQ algorithms (Problems 3.1 and 3.2 below) and in the stochastic Galerkin method (Problem 3.3 below). These are the first such results for arbitrarily large  $k$  and for  $A$  and  $n$  varying independently of  $k$ . These results are proved by combining:

1. bounds for the Helmholtz equation in [105] (detailed in Section 2.2.2 above) with  $A$  and  $n$  deterministic but spatially-varying, with
2. general arguments (i.e. not specific to Helmholtz) presented here for proving a priori bounds and well-posedness of variational formulations of linear time-independent SPDEs.

Regarding 1: the  $k$ -dependence of the bounds on  $u$  in terms of  $f$  depends crucially on whether or not  $A$ ,  $n$ , and  $D_-$  are such that there exist trapped rays. In the trapping case, the solution operator can grow exponentially in  $k$ ; in contrast, in the nontrapping case, the solution operator is bounded uniformly in  $k$  (see the review in Section 2.2.3 above). The bounds in [105] are under conditions on  $A$ ,  $n$ , and  $D_-$  that ensure nontrapping of rays; the significance of these bounds is that they are the first (deterministic) bounds for the Helmholtz scattering problem in which both  $A$  and  $n$  vary and the bounds are explicit in  $A$  and  $n$  (as well as in  $k$ ). This feature of being explicit in  $A$  and  $n$  is crucial in allowing us to prove the results in this chapter when  $A$  and  $n$  are random fields.

Regarding 2: the main reason these general arguments are needed is the fact that the variational formulations of both the deterministic and the stochastic Helmholtz equation are not coercive,

and so one cannot use the Lax–Milgram theorem to conclude well-posedness and an a priori bound. In the deterministic case, the remedy for the lack of coercivity of the Helmholtz equation is to use Fredholm theory, but this is *not* applicable to the stochastic variational formulation of the Helmholtz equation because the necessary compactness results do not hold in Bochner spaces (see Appendix A below). Our solution to this lack of coercivity and failure of Fredholm theory is to use well-posedness results and bounds from the deterministic case to prove results for the stochastic case. We work ‘pathwise’ by integrating the deterministic results over probability space, identifying conditions under which the necessary quantities are indeed integrable. Our approach is given in a general framework that, given (i) deterministic well-posedness results and a priori bounds that are explicit in all the coefficients, and (ii) measurability and integrability conditions on the stochastic quantities, returns corresponding well-posedness results, a priori bounds, and equivalence results for different formulations of the stochastic problem. One reason we state our well-posedness results in general (i.e. not only in the specific case of the Helmholtz equation) is that we expect that they can be used in the future to prove well-posedness results for the time-harmonic Maxwell’s equations in random media. A nontechnical summary of the ideas behind our general well-posedness results is given in Remark 3.29 below. Some of these results are similar in spirit to the results about the PDE (3.2) in [96, 157] (which deal with the failure of Lax–Milgram for the stochastic variational problem for (3.2) in the case when the coefficient  $\chi$  is not uniformly bounded above and below), and our general arguments use some of the ideas and technical tools from these two papers.

### 3.1.1 Statement of main results

*Notation and basic definitions* We now give the setting for our stochastic Helmholtz problems. Whilst this setting is similar to the deterministic setting in Section 2.2.1 above, one key difference is that we specify the ball  $B_R$  which contains the inhomogeneities in  $A$ ,  $n$ , and  $f$ . In contrast, in Section 2.2.1 we simply assume  $I - A$ ,  $1 - n$ , and  $f$  have compact support. Let either (i)  $D_- \subset \mathbb{R}^d$ ,  $d = 2, 3$ , be a bounded Lipschitz open set such that  $\mathbf{0} \in D_-$  and the open complement  $D_+ := \mathbb{R}^d \setminus \overline{D_-}$  is connected, or (ii)  $D_- = \emptyset$ . Let  $\Gamma_D = \partial D_-$ . Fix  $R > 0$  and let  $B_R$  be the ball of radius  $R$  centred at the origin. Define  $\Gamma_R := \partial B_R$  and  $D_R := D_+ \cap B_R$  (see Figure 2.1). Let  $\gamma$  denote the trace operator from  $D_R$  to  $\partial D_R = \Gamma_D \cup \Gamma_R$  and define  $H_{0,D}^1(D_R) := \{v \in H^1(D_R) : \gamma v = 0 \text{ on } \Gamma_D\}$ .

Let  $T_R : H^{1/2}(\Gamma_R) \rightarrow H^{-1/2}(\Gamma_R)$  be the Dirichlet-to-Neumann map for the deterministic equation  $\Delta u + k^2 u = 0$  posed in the exterior of  $B_R$  with the Sommerfeld radiation condition

$$\frac{\partial u}{\partial r}(\mathbf{x}) - ik u(\mathbf{x}) = o\left(\frac{1}{r^{(d-1)/2}}\right) \text{ as } r := |\mathbf{x}| \rightarrow \infty, \text{ uniformly in } \frac{\mathbf{x}}{|\mathbf{x}|}; \quad (3.3)$$

see [158, Section 2.6.3] and [38, Equations 3.5 and 3.6] for an explicit expression for  $T_R$  in terms of Hankel functions and Fourier series ( $d = 2$ )/spherical harmonics ( $d = 3$ ). Let  $(\cdot, \cdot)_{\Gamma_R}$  be the duality pairing on  $\Gamma_R$  between  $H^{-1/2}(\Gamma_R)$  and  $H^{1/2}(\Gamma_R)$  and write  $d\lambda$  for Lebesgue measure. Throughout this chapter we explicitly include the measure when writing integrals, to help distinguish between integrals over the spatial domain (using  $d\lambda$ ) and integrals over the probability space (using  $d\mathbb{P}$ ).

For  $A_0 \in \mathbb{R}^{d \times d}$ , we write  $\|A_0\|_{\text{op}}$  for the operator norm induced by the Euclidean vector norm on  $\mathbb{C}^d$ , and for  $A : D_R \rightarrow \mathbb{R}^{d \times d}$ , we write  $\|A\|_{L^\infty(D_R; \text{op})}$  for the norm  $\left\| \|A(\mathbf{x})\|_{\text{op}} \right\|_{L^\infty(D_R; \mathbb{R})}$ . Observe that, by the equivalence of all norms on the finite-dimensional space  $\mathbb{R}^{d \times d}$ ,  $\|\cdot\|_{L^\infty(D_R; \text{op})}$  is equivalent to the more standard norm

$$\|A\|_{L^\infty(D_R; \mathbb{R}^{d \times d})} := \sup_{i,j=1,\dots,d} \|A_{i,j}\|_{L^\infty(D_R; \mathbb{R})}. \quad (3.4)$$

We define  $\|\cdot\|_{W^{1,\infty}(D_R; \mathbb{R}^{d \times d})}$  by

$$\|A\|_{W^{1,\infty}(D_R; \mathbb{R}^{d \times d})} := \sup_{i,j=1,\dots,d} \|A_{i,j}\|_{W^{1,\infty}(D_R; \mathbb{R})}. \quad (3.5)$$

We write  $D_1 \subset\subset D_2$  if  $D_1$  is a compact subset of the open set  $D_2$ . Throughout this chapter, unless stated otherwise we equip a topological space with its Borel  $\sigma$ -algebra. See Appendix B for a summary of the measure-theoretic concepts used in this chapter. Let

- $(\Omega, \mathcal{F}, \mathbb{P})$  be a complete probability space,
- $f : \Omega \rightarrow L^2(D_+)$  be such that  $\text{supp } f(\omega) \subset\subset B_R$  almost surely,
- $n : \Omega \rightarrow L^\infty(D_+; \mathbb{R})$  be such that  $\text{supp}(1-n(\omega)) \subset\subset B_R$  almost surely and there exist  $n_{\min}, n_{\max} : \Omega \rightarrow \mathbb{R}$  such that  $0 < n_{\min}(\omega) \leq n(\omega)(\mathbf{x}) \leq n_{\max}(\omega)$  for almost every  $\mathbf{x} \in D_+$  almost surely, and
- $A : \Omega \rightarrow L^\infty(D_+; \text{SPD})$  be such that  $\text{supp}(I-A(\omega)) \subset\subset B_R$  and there exist  $A_{\min}, A_{\max} : \Omega \rightarrow \mathbb{R}$  such that  $0 < A_{\min}(\omega) < A_{\max}(\omega)$  almost surely and  $A_{\min}(\omega)|\xi|^2 \leq (A(\omega)(\mathbf{x})\xi) \cdot \xi \leq A_{\max}(\omega)|\xi|^2$  for almost every  $\mathbf{x} \in D_+$  and for all  $\xi \in \mathbb{C}^d$  almost surely.

If  $v : \Omega \rightarrow Z$  for some function space  $Z$  of functions on  $\mathbb{R}^d$ , we abuse notation slightly and write  $v(\omega, \mathbf{x})$  instead of  $v(\omega)(\mathbf{x})$ .

*Variational Formulations* We consider three different formulations of the *Helmholtz stochastic exterior Dirichlet problem* (stochastic EDP); Problems 3.1–3.3 below.

Define the sesquilinear form  $a(\omega)$  on  $H_{0,D}^1(D_R) \times H_{0,D}^1(D_R)$  by

$$[a(\omega)](v_1, v_2) := \int_{D_R} \left( (A(\omega)\nabla v_1) \cdot \nabla \bar{v}_2 - k^2 n(\omega) v_1 \bar{v}_2 \right) d\lambda - \langle T_R \gamma v_1, \gamma v_2 \rangle_{\Gamma_R}, \quad (3.6)$$

and the antilinear functional  $L(\omega)$  on  $H_{0,D}^1(D_R)$  by

$$[L(\omega)](v_2) := \int_{D_R} f(\omega) \bar{v}_2 d\lambda. \quad (3.7)$$

Define the sesquilinear form  $\mathfrak{a}$  on  $L^2(\Omega; H_{0,D}^1(D_R)) \times L^2(\Omega; H_{0,D}^1(D_R))$  and the antilinear func-

tional  $\mathfrak{L}$  on  $L^2(\Omega; H_{0,D}^1(D_R))$  by

$$\mathfrak{a}(v_1, v_2) := \int_{\Omega} [a(\omega)](v_1(\omega), v_2(\omega)) d\mathbb{P}(\omega) \quad \text{and} \quad \mathfrak{L}(v_2) := \int_{\Omega} [L(\omega)](v_2(\omega)) d\mathbb{P}(\omega). \quad (3.8)$$

We consider the following three problems:

**Problem 3.1** (Measurable EDP almost surely). *Find a measurable  $u : \Omega \rightarrow H_{0,D}^1(D_R)$  such that*

$$[a(\omega)](u(\omega), v) = [L(\omega)](v) \text{ for all } v \in H_{0,D}^1(D_R) \text{ almost surely.}$$

**Problem 3.2** (Second-order EDP almost surely). *Find  $u \in L^2(\Omega; H_{0,D}^1(D_R))$  such that*

$$[a(\omega)](u(\omega), v) = [L(\omega)](v) \text{ for all } v \in H_{0,D}^1(D_R) \text{ almost surely.}$$

**Problem 3.3** (Stochastic variational EDP). *Find  $u \in L^2(\Omega; H_{0,D}^1(D_R))$  such that*

$$\mathfrak{a}(u, v) = \mathfrak{L}(v) \text{ for all } v \in L^2(\Omega; H_{0,D}^1(D_R)).$$

Problem 3.2 is the foundation of sampling-based UQ methods, such as Monte-Carlo and Stochastic-Collocation methods; its analogue for the stationary diffusion equation is well-studied in, e.g., [218, 7, 160, 41, 42, 204, 131, 114]. Similarly Problem 3.3 is the foundation of the Stochastic Galerkin method (a finite-element method in  $\Omega \times D$ , where  $D$  is the spatial domain), and is studied for the Helmholtz Interior Impedance Problem in [80], and its analogue for the stationary diffusion equation is considered in, e.g., [8, 129, 15, 108].

**Remark 3.4** (Why consider Problem 3.1?). *The difference between Problems 3.1 and 3.2 is that Problem 3.1 requires no integrability of  $u$  over  $\Omega$ , but Problem 3.2 requires  $u \in L^2(\Omega; H_{0,D}^1(D_R))$ . Since all the theory for sampling-based UQ methods assume some integrability of the solution, the natural question is: why consider Problem 3.1 at all?*

*The main reason we consider Problem 3.1 is that, given the existing PDE theory for the Helmholtz equation, we can prove existence of a solution to Problem 3.1 under general conditions on  $A$  and  $n$ , but there is no current prospect of proving existence of a solution to Problem 3.2 under general conditions on  $A$  and  $n$ . The explanation for this consists of the following three points:*

1. *The only two known ways to obtain a solution to Problem 3.2 are: (i) obtain a deterministic a priori bound, explicit in all parameters, and integrate (followed, e.g., in [42] for (3.2) with lognormal coefficients) and (ii) obtain a solution to Problem 3.3 and show this is a solution to Problem 3.2. In the Helmholtz case, doing (ii) is difficult as neither the Lax–Milgram theorem nor Fredholm theory is applicable (as explained in the introduction), and so we follow the approach in (i).*
2. *The only known bounds on the solution of the Helmholtz equation explicit in all parameters are those recently obtained for nontrapping scenarios in [105, 83].*

3. Obtaining a bound explicit in all the parameters for any general class of  $A$  and  $n$ , e.g.,  $A \in W^{1,\infty}(D_R; \text{SPD})$  and  $n \in L^\infty(D_R; \mathbb{R})$  is well beyond current techniques. Indeed, a general class of  $A$  and  $n$  will include both trapping and nontrapping scenarios, and such a bound would need to capture the exponential blow-up in  $k$  for trapping  $A$  and  $n$ , the uniform boundedness in  $k$  for nontrapping  $A$  and  $n$ , and be explicit in  $A$  and  $n$ .

Given this fact that there is no current prospect of proving existence of a solution to Problem 3.2 under general conditions on  $A$  and  $n$  we keep Problem 3.1 so that we prove an (albeit weaker) existence result for the Helmholtz equation with general coefficients.

**Remark 3.5** (Measurability of  $u$  in Problem 3.1). *It is natural to construct the solution of Problem 3.1 pathwise; that is, one defines  $u(\omega)$  to be the solution of the deterministic problem with coefficients  $A(\omega)$  and  $n(\omega)$ . However, it is then not obvious that  $u$  is measurable. In the proof of Theorem 3.7 below, we show that the measurability of  $u$  follows from*

1. a natural condition on the measurability of the coefficients and data (Condition C1 below), and
2. the continuity of the map taking the coefficients of the deterministic PDE to the solution of the deterministic PDE (see Lemma 3.54 below).

In Theorems 3.7 and 3.10 we prove results on the well-posedness of Problems 3.1–3.3 under conditions on  $A$ ,  $n$ ,  $f$ , and  $D_-$ . Although  $A$ ,  $n$ , and  $f$  are defined on  $D_+$ , since  $\text{supp}(I - A)$ ,  $\text{supp}(1 - n)$ , and  $\text{supp} f$  are compactly contained in  $D_R$  we can consider  $A$ ,  $n$ , and  $f$  as functions on  $D_R$ .

**Condition 3.6** (Regularity and stochastic regularity of  $f$ ,  $A$ , and  $n$ ). *The random fields  $f$ ,  $A$ , and  $n$  satisfy  $f \in L^2(\Omega; L^2(D_R))$ ,  $A : \Omega \rightarrow W^{1,\infty}(D_R; \text{SPD})$  with  $A \in L^\infty(\Omega; L^\infty(D_R; \mathbb{R}^{d \times d}))$ , and  $n \in L^\infty(\Omega; L^\infty(D_R; \mathbb{R}))$ .*

**Theorem 3.7** (Equivalence of variational problems). *Under Condition 3.6:*

- The maps  $\mathfrak{a}$  and  $\mathfrak{L}$  (defined by (3.8)) are well-defined.
- $u \in L^2(\Omega; H_{0,D}^1(D_R))$  solves Problem 3.2 if and only if  $u$  solves Problem 3.3.
- If  $u \in L^2(\Omega; H_{0,D}^1(D_R))$  solves Problem 3.2, then any member of the equivalence class of  $u$  solves Problem 3.1.
- The solution of Problem 3.1 exists and is unique up to modification on a set of measure zero in  $\Omega$ .
- The solution of Problems 3.2 and 3.3 is unique in  $L^2(\Omega; H_{0,D}^1(D_R))$ .

The proof of Theorem 3.7 is on page 135 below.

Observe that the only relationship between formulations not proved in Theorem 3.7 is: if  $u : \Omega \rightarrow H_{0,D}^1(D_R)$  solves Problem 3.1 then  $u \in L^2(\Omega; H_{0,D}^1(D_R))$  and  $u$  solves Problem 3.2. Theorem 3.10 below includes this relationship, but we need additional assumptions on  $A$ ,  $n$ , and  $D_-$ . We use the notation established in Definition 2.3.

**Condition 3.8** ( $k$ -independent nontrapping conditions on (random)  $A$  and  $n$ ).

The random fields  $A$  and  $n$  satisfy  $A : \Omega \rightarrow W^{1,\infty}(D_R; \text{SPD})$  and  $n : \Omega \rightarrow W^{1,\infty}(D_R; \mathbb{R})$  with  $\text{supp}(I - A(\omega)) \subset\subset B_R$  and  $\text{supp}(1 - n(\omega)) \subset\subset B_R$  almost surely. Furthermore, there exist  $\mu_1, \mu_2 : \Omega \rightarrow \mathbb{R}$ , independent of  $f$ , with  $\mu_1(\omega), \mu_2(\omega) > 0$  almost surely and  $1/\mu_1, 1/\mu_2 \in L^2(\Omega; \mathbb{R})$  such that  $A(\omega) \in \text{NT}_{\text{mat}, D_R}(\mu_1(\omega))$  almost surely and  $n(\omega) \in \text{NT}_{\text{scal}, D_R}(\mu_2(\omega))$  almost surely.

**Definition 3.9** (Star-shaped). The set  $D \subseteq \mathbb{R}^d$  is star-shaped with respect to the point  $\mathbf{x}_0$  if for any  $\mathbf{x} \in D$  the line segment  $[\mathbf{x}_0, \mathbf{x}] \subseteq D$ .

**Theorem 3.10** (Equivalence of variational problems in a nontrapping case).

Let  $D_-$  be star-shaped with respect to the origin. Under Conditions 3.6 and 3.8:

- The maps  $\mathfrak{a}$  and  $\mathfrak{L}$  (defined by (3.8)) are well-defined.
- Problems 3.1–3.3 are all equivalent.
- The solution  $u \in L^2(\Omega; H_{0,D}^1(D_R))$  of these problems exists, is unique, and, given  $k_0 > 0$ , satisfies the bound

$$\|\nabla u\|_{L^2(\Omega; L^2(D_R))}^2 + k^2 \|u\|_{L^2(\Omega; L^2(D_R))}^2 \leq \|C_1\|_{L^1(\Omega)} \|f\|_{L^2(\Omega; L^2(D_R))}^2 \quad (3.9)$$

for all  $k \geq k_0$ , where  $C_1 : \Omega \rightarrow \mathbb{R}$  is given by

$$C_1 = \max\left\{\frac{1}{\mu_1}, \frac{1}{\mu_2}\right\} \left(\frac{R^2}{\mu_1} + \frac{2}{\mu_2} \left(R + \frac{d-1}{2k_0}\right)^2\right). \quad (3.10)$$

The proof of Theorem 3.10 is on page 136 below.

As highlighted above, Theorem 3.10 is obtained by combining deterministic a priori bounds from Theorem 2.6 with the general arguments in Section 3.2 about well-posedness of variational formulations of stochastic PDEs. Theorem 3.10 uses the most basic a priori bound proved in [105] (from [105, Theorem 2.5]), but [105] contains several extensions of this bound. Remarks 3.11–3.15 outline the implications of these (deterministic) extensions for the stochastic Helmholtz equation.

**Remark 3.11** (Dirichlet boundary conditions on  $\Gamma_D$  and plane-wave incidence).

The formulations of the stochastic EDP above assume that  $u = 0$  on the boundary  $\Gamma_D$ . An important scattering problem for which  $u \neq 0$  on  $\Gamma_D$  is when  $u$  is the field scattered by an incident plane wave; in this case  $\gamma u = -\gamma u_I$ , where  $u_I$  is the incident plane wave [39, p. 107].

The results in this chapter can be easily extended to the case when  $u \neq 0$  on  $\Gamma_D$  using [105, Theorem 2.19(ii)] which proves a priori (deterministic) bounds in this case. One subtlety, however, is that  $f$  is then not necessarily independent of  $\mu_1$  and  $\mu_2$ . Indeed in this case  $f = -\nabla \cdot (A \nabla u_I) - k^2 n u_I$ . If  $\mu_1$  depends on  $A$  and  $\mu_2$  depends on  $n$  then  $f$  may be not be independent of  $\mu_1$  and  $\mu_2$ . One can produce an analogue of Theorem 3.10 in the case where  $f, \mu_1$ , and  $\mu_2$  are dependent, but one requires  $1/\mu_1, 1/\mu_2 \in L^4(\Omega)$  and  $f \in L^4(\Omega; L^2(D))$ ; see Remark 3.58 below.

**Remark 3.12** (The case when either  $n = 1$  or  $A = I$ ). *When either  $n = 1$  or  $A = I$ , [105, Theorem 2.19] gives deterministic bounds under weaker conditions on  $A$  and  $n$  respectively; the corresponding results for the stochastic case are that:*

- *When  $n = 1$  almost surely, the condition  $A(\omega) \in \text{NT}_{\text{mat}, D_R}(\mu_1(\omega))$  in Condition 3.8 can be improved to  $2A(\omega) - (\mathbf{x} \cdot \nabla)A(\omega) \geq \mu_1(\omega)$  for almost every  $\mathbf{x} \in D_+$ , almost surely.*
- *When  $A = I$  almost surely, the condition  $n(\omega) \in \text{NT}_{\text{scal}, D_R}(\mu_2(\omega))$  in Condition 3.8 can be improved to:*

$$2n(\omega) + \mathbf{x} \cdot \nabla n(\omega) \geq \mu_2(\omega) \text{ for almost every } \mathbf{x} \in D_+, \text{ almost surely.} \quad (3.11)$$

**Remark 3.13** (The Helmholtz stochastic truncated exterior Dirichlet problem).

*When applying the Galerkin method to Problems 3.1–3.3, the Dirichlet-to-Neumann map  $T_R$  is expensive to compute. Therefore, it is common to approximate the DtN map on  $\Gamma_R$  by an ‘absorbing boundary condition’ (see, e.g., [118, Section 3.3] and the references therein), the simplest of which is the impedance boundary condition  $\partial u / \partial \nu - ik u = 0$ . We call the Helmholtz stochastic EDP posed in  $D_R$  with an impedance boundary condition on  $\Gamma_R$  the stochastic truncated exterior Dirichlet problem (stochastic TEDP). In fact, since we no longer need to know the DtN map explicitly on the truncation boundary, the truncation boundary can be arbitrary (i.e. it does not have to be just a circle/sphere). Note that in the case when the obstacle is the empty set, the TEDP is just the Interior Impedance Problem.*

*The results in this chapter also hold for the stochastic TEDP (with arbitrary Lipschitz truncation boundary) under an analogue of Condition 3.8 based on the deterministic bounds in [105, Theorem A.6(i)] instead of [105, Theorem 2.5].*

**Remark 3.14** (Discontinuous  $A$  and  $n$ ). *The requirements on  $A$  and  $n$  in Condition 3.8 require them to be continuous (since  $W^{1,\infty}(D_R) = C^{0,1}(D_R)$  as  $D_R$  is Lipschitz; see, e.g., [74, Section 4.2.3, Theorem 5]). In addition to proving deterministic a priori bounds for the class of  $A$  and  $n$  in Condition 3.8, the paper [105] proves deterministic bounds for discontinuous  $A$  and  $n$  satisfying (2.9) and (2.10) in a distributional sense; see [105, Theorem 2.7]. In this case, when moving outward from the obstacle to infinity,  $A$  can jump downwards and  $n$  can jump upwards on interfaces that are star-shaped. (When the jumps are in the opposite direction, the problem is trapping; see [178] and [152, Section 6]). The well-posedness results and a priori bounds in this chapter can therefore be adapted to prove results about the stochastic Helmholtz equation for a class of random  $A$  and  $n$  that allows nontrapping jumps on randomly-placed star-shaped interfaces.*

**Remark 3.15** ( $k$ -dependent  $A$  and  $n$ ). *In this chapter we focus on random fields  $A$  and  $n$  varying independently of  $k$ ; this corresponds to a fixed physical medium, characterised by  $A$  and  $n$ , with waves of frequency  $k$  passing through. In Section 3.1.2 below we construct  $A$  and  $n$  as ( $k$ -independent)  $W^{1,\infty}$  perturbations of random fields  $A_0$  and  $n_0$  satisfying Condition 3.8. We note, however, that results for  $A$  and  $n$  being  $k$ -dependent  $L^\infty$  perturbations (i.e. rougher, but  $k$ -dependent perturbations) of  $A_0$  and  $n_0$  satisfying Condition 3.8 can easily be obtained.*

The basis for these bounds is observing that deterministic a priori bounds hold when (a)  $A \in \text{NT}_{\text{mat}, D_R}(\mu_1)$ ,  $n = n_0 + \eta$ , where  $n_0 \in \text{NT}_{\text{scal}, D_R}(\mu_2)$  and  $k\|\eta\|_{L^\infty(D_R; \mathbb{R})}$  is sufficiently small, and (b)  $A = A_0 + B$ ,  $n = n_0 + \eta$ , where  $A_0 \in \text{NT}_{\text{mat}, D_R}(\mu_1)$ ,  $n_0 \in \text{NT}_{\text{scal}, D_R}(\mu_2)$ ,  $k\|\eta\|_{L^\infty(D_R; \mathbb{R})}$  and  $k\|B\|_{W^{1,\infty}(D_R; \mathbb{R}^{d \times d})}$  are both sufficiently small, and  $A, n$ , and  $D_-$  are such that  $u \in H^2(D_R)$  (see, e.g., [146, Theorem 4.18(i)] or [107, Theorems 2.3.3.2 and 2.4.2.5] for these latter requirements). Given these deterministic bounds, the general arguments in this chapter can then be used to prove well-posedness of the analogous stochastic problems.

To understand why bounds hold in the case (a), observe that one can write the PDE as

$$\nabla \cdot (A \nabla u) + k^2 n_0 u = -f - k^2 \eta u; \quad (3.12)$$

if  $k\|\eta\|_{L^\infty(D_R; \mathbb{R})}$  is sufficiently small then the contribution from the  $k^2 \eta u$  term on the right-hand side of (3.12) can be absorbed into the  $k^2 \|u\|_{L^2(D_R)}^2$  term appearing on the left-hand side of the bound (the deterministic analogue of (3.9)). In the case  $n_0 = 1$ , this is essentially the argument used to prove the a priori bound in [80, Theorem 2.4] (see [105, Remark 2.15]). The reason bounds hold in the case (b) is similar, except now we need the  $H^2$  norm of  $u$  on the left-hand side of the bound (as well as the  $H^1$  norm) to absorb the contribution from the  $\nabla \cdot (B \nabla u)$  term on the right-hand side.

### 3.1.2 Random fields satisfying Condition 3.8

The main focus of this chapter is proving well-posedness of the variational formulations of the stochastic Helmholtz equation, and a priori bounds on the solution, for the most-general class of  $A$  and  $n$  allowed by the deterministic bounds in [105]. However, in this section, motivated by the Karhunen-Loève expansion (see e.g. [143, p. 201ff.]) and similar expansions of material coefficients for the stationary diffusion equation [131, Section 2.1], we consider  $A$  and  $n$  as series expansions around known non-random fields  $A_0$  and  $n_0$  satisfying Condition 3.8 (i.e., Condition 3.8 is satisfied for  $n_0, A_0$  independent of  $\omega \in \Omega$ , and therefore  $\mu_1, \mu_2$  independent of  $\omega$ ). Define

$$A(\omega, \mathbf{x}) = A_0(\mathbf{x}) + \sum_{j=1}^{\infty} Y_j(\omega) \sqrt{\Lambda_j} \Psi_j(\mathbf{x}) \quad \text{and} \quad n(\omega, \mathbf{x}) = n_0(\mathbf{x}) + \sum_{j=1}^{\infty} Z_j(\omega) \sqrt{\lambda_j} \psi_j(\mathbf{x}), \quad (3.13)$$

where:

- $\text{supp}(1 - A_0), \text{supp}(I - n_0) \subset\subset B_R$ ,
- $A_0$  and  $n_0$  satisfy Condition 3.8 with  $\mu_1$  and  $\mu_2$  independent of  $\omega \in \Omega$
- $Y_j, Z_j \sim \text{Unif}(-1/2, 1/2)$  i.i.d.,
- $\Lambda_j, \lambda_j > 0$  for all  $j = 1, \dots, \infty$ ,
- $\Psi_j \in W^{1,\infty}(D_R; \text{SPD})$  with  $\text{supp} \Psi_j \subset\subset B_R$  for all  $j = 1, \dots, \infty$ ,

$$\sum_{j=1}^{\infty} \sqrt{\Lambda_j} \|\Psi_j\|_{W^{1,\infty}(D_R; \mathbb{R}^{d \times d})} < \infty, \quad \text{and} \quad (3.14)$$

$$\sum_{j=1}^{\infty} \sqrt{\Lambda_j} \|\Psi_j\|_{L^\infty(D_R; \text{op})} < 2A_{0,\min}, \quad (3.15)$$

where  $A_{0,\min} > 0$  is such that  $A_{0,\min} |\xi|^2 \leq (A_0(\mathbf{x})\xi) \cdot \xi$  for almost every  $\mathbf{x} \in D_+$  and for all  $\xi \in \mathbb{C}^d$ .

- $\psi_j \in W^{1,\infty}(D_R; \mathbb{R})$  with  $\text{supp } \psi_j \subset\subset B_R$  for all  $j = 1, \dots, m$ ,

$$\sum_{j=1}^{\infty} \sqrt{\lambda_j} \|\psi_j\|_{W^{1,\infty}(D_R; \mathbb{R})} < \infty, \text{ and} \quad (3.16)$$

$$\sum_{j=1}^{\infty} \sqrt{\lambda_j} \|\psi_j\|_{L^\infty(D_R; \mathbb{R})} < 2n_{0,\min}, \quad (3.17)$$

where  $n_{0,\min} := \text{ess inf}_{\mathbf{x} \in D_R} n_0(\mathbf{x})$ .

The assumptions (3.15) and (3.17) ensure that  $A > 0$  (in the sense of quadratic forms) and  $n > 0$  almost surely, and the assumptions (3.14) and (3.16) are used to prove  $A$  and  $n$  are measurable.

Regarding the measurability of  $A$  and  $n$  defined by (3.13): the proof that  $A$  and  $n$  given by (3.13) are measurable is given in Lemma C.12, and relies on the proof that the sum of measurable functions is measurable. This latter result is standard, but we have not been able to find this result for this particular setting of mappings into a separable subspace of a general normed vector space, and so we briefly give it in Lemma C.7.

The following lemmas give sufficient conditions for the series in (3.13) to satisfy Condition 3.8.

**Lemma 3.16** (Series expansion of  $A$  satisfies Condition 3.8). *Let  $\mu > 0$ ,  $\delta \in (0, 1)$  be fixed. If  $A_0 \in \text{NT}_{\text{mat}, D_R}(\mu)$ , and*

$$\sum_{j=1}^{\infty} \sqrt{\Lambda_j} \|\Psi_j(\mathbf{x}) - (\mathbf{x} \cdot \nabla) \Psi_j(\mathbf{x})\|_{L^\infty(D_R; \text{op})} \leq 2\delta \mu, \quad (3.18)$$

*then  $A \in \text{NT}_{\text{mat}, D_R}((1 - \delta)\mu)$  almost surely.*

*Proof of Lemma 3.16.* Since  $A_0 \in \text{NT}_{\text{mat}, D_R}(\mu)$ , we have

$$\left( (A(\omega, \mathbf{x}) - (\mathbf{x} \cdot \nabla) A(\omega, \mathbf{x})) \xi \right) \cdot \bar{\xi} \geq \mu |\xi|^2 + \sum_{j=1}^{\infty} \left( Y_j(\omega) \sqrt{\Lambda_j} (\Psi_j(\mathbf{x}) - (\mathbf{x} \cdot \nabla) \Psi_j(\mathbf{x})) \xi \right) \cdot \bar{\xi} \quad (3.19)$$

for all  $\xi \in \mathbb{C}^d$ , for almost every  $\mathbf{x} \in D_R$ , almost surely. As  $Y_j \sim \text{Unif}(-1/2, 1/2)$  for all  $j$  and the bound (3.18) holds, the right-hand side of (3.19) is bounded below by

$$\mu |\xi|^2 - \frac{1}{2} 2\delta \mu |\xi|^2 = (1 - \delta) \mu |\xi|^2 \text{ almost surely.}$$

Since  $\xi \in \mathbb{C}^d$  was arbitrary, it follows that  $A(\omega) \in \text{NT}_{\text{mat}, D_R}((1 - \delta)\mu)$  almost surely, as required.  $\square$

**Lemma 3.17** (Series expansion of  $n$  satisfies Condition 3.8). *Let  $\mu > 0$  and  $\delta \in (0, 1)$ . If  $n_0 \in \text{NT}_{\text{scal}, D_R}(\mu)$  and*

$$\sum_{j=1}^{\infty} \sqrt{\lambda_j} \left\| \psi_j(\mathbf{x}) + \mathbf{x} \cdot \nabla \psi_j(\mathbf{x}) \right\|_{L^\infty(D_R; \mathbb{R})} \leq 2\delta \mu, \quad (3.20)$$

*then  $n \in \text{NT}_{\text{scal}, D_R}((1 - \delta)\mu)$ .*

The proof of Lemma 3.17 is omitted, since it is similar to the proof of Lemma 3.16; in fact it is simpler, because it involves scalars rather than matrices.

### 3.1.3 Outline of the chapter

In Section 3.1.4 we discuss our results in the context of related literature. In Section 3.2 we state general results on a priori bounds and well-posedness for stochastic variational formulations. In Section 3.3 we prove the results in Section 3.2. In Section 3.4 we prove Theorems 3.7 and 3.10.

### 3.1.4 Discussion of the main results in the context of other work on UQ for time-harmonic wave equations

In this section we discuss existing results on well-posedness of (3.1), as well as analogous results for the elastic wave equation and the time-harmonic Maxwell's equations. The most closely-related work to this chapter is [80] (and its analogue for elastic waves [75]), in that a large component of [80] consists of attempting to prove well-posedness and a priori bounds for the stochastic variational formulation (i.e. Problem 3.3) of the Helmholtz Interior Impedance Problem; i.e., (3.1) with  $A = I$  and stochastic  $n$  posed in a bounded domain with an impedance boundary condition  $\partial u / \partial \nu - iku = g$ . (Recall that this boundary condition is a simple approximation to the Dirichlet-to-Neumann map  $T_R$  defined above (3.3).) Under the assumption of existence, [80] shows that for any  $k > 0$  the solution is unique and satisfies an a priori bound of the form (3.9) (with different constant  $C_1$ ), provided  $n = 1 + \eta$  where the random field  $\eta$  satisfies (almost surely)  $\|\eta\|_{L^\infty} \leq C/k$  for some  $C > 0$  independent of  $k$ . [80] then invokes Fredholm theory to conclude existence, but this relies on an incorrect assumption about compact inclusion of Bochner spaces—see Appendix A below. However, combining Theorem 3.7 and Remarks 3.13 and 3.15 with  $A = I$  and  $n_0 = 1 + \eta$  (with  $\eta$  as above) produces an analogous result to Theorem 3.10, and gives a correct proof of [80, Theorem 2.5]. Therefore the analysis of the Monte Carlo interior penalty discontinuous Galerkin method in [80] can proceed under the assumptions of Theorem 3.7 and Remarks 3.13 and 3.15.

The papers [117] and [191] consider the Helmholtz transmission problem with a stochastic interface, i.e. (3.1) posed in  $\mathbb{R}^d$  with both  $A$  and  $n$  piecewise constant and jumping on a common, randomly-located interface. A component of this work is establishing well-posedness of Problem 3.1 for this setup. To do this, the authors make the assumption that  $k$  is small (to avoid problems with trapping mentioned above—see the comments after [117, Theorem 4.3]); the sesquilinear form  $a$  is then coercive and a priori bounds (in principle explicit in  $A$  and  $n$ ) follow in Sobolev norms [117, Lemma 4.5] and Hölder norms [191, Theorem 5.1 and Corollary 5.2].

By Remark 3.14, the results of this chapter can be used to obtain the analogous well-posedness result for large  $k$  in the case of nontrapping jumps.

The paper [32] studies the *Bayesian inverse problem* associated to (3.1) with  $A = I$  and  $n = 1$  posed in the exterior of a Dirichlet obstacle. That is, [32] analyses computing the posterior distribution of the shape of the obstacle given noisy observations of the acoustic field in the exterior of the obstacle. A component of the analysis in [32] is the well-posedness of the forward problem for an obstacle with a variable boundary [32, Proposition 3.5]. Instead of mapping the problem to one with a fixed domain and variable  $A$  and  $n$ , [32] instead works with the variability of the obstacle directly, using boundary-integral equations. The  $k$ -dependence of the solution operator is not considered, but would enter in [32, Lemma 3.1].

The papers [123] and [122] consider the time-harmonic Maxwell's equations with (i) the material coefficients  $\varepsilon, \mu$  constant in the exterior of a perfectly-conducting random obstacle and (ii)  $\varepsilon, \mu$  piecewise constant and jumping on a common randomly located interface; in both cases these problems are mapped to problems where the domain/interface is fixed and  $\varepsilon$  and  $\mu$  are random and heterogeneous. The papers [123] and [122] essentially consider the analogue of Problem 3.1 for the time-harmonic Maxwell's equations, obtaining well-posedness from the corresponding results for the related deterministic problems. Similarly the paper [71] considers analogues of Problem 3.1 for the Helmholtz equation with constant coefficients and a random (penetrable or impenetrable) obstacle. The paper [71] obtains *deterministic* tensor-product boundary-integral equations for the statistical moments of  $u$ .

## 3.2 GENERAL RESULTS PROVING A PRIORI BOUNDS AND WELL-POSEDNESS OF STOCHASTIC VARIATIONAL FORMULATIONS

In this section we state general results for proving a priori bounds and well-posedness results for variational formulations of linear time-independent SPDEs.

### 3.2.1 Notation and definitions of the variational formulations

Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be a complete probability space. Let  $X$  and  $Y$  be separable Banach spaces over a field  $\mathbb{F}$ , (where  $\mathbb{F} = \mathbb{R}$  or  $\mathbb{C}$ ). Let  $B(X, Y^*)$  denote the space of bounded linear maps  $X \rightarrow Y^*$ . Let  $\mathcal{C}$  be a topological space with topology  $\mathcal{T}_{\mathcal{C}}$ . Given maps

$$c : \Omega \rightarrow \mathcal{C}, \quad \mathcal{A} : \mathcal{C} \rightarrow B(X, Y^*), \quad \text{and} \quad \mathcal{L} : \mathcal{C} \rightarrow Y^*,$$

let  $\mathfrak{A} : L^2(\Omega; X) \rightarrow L^2(\Omega; Y)^*$  and  $\mathfrak{L} \in L^2(\Omega; Y)^*$  be defined by

$$[\mathfrak{A}(v_1)](v_2) := \int_{\Omega} [\mathcal{A}_{c(\omega)} v_1(\omega)](v_2(\omega)) d\mathbb{P}(\omega) \quad \text{and} \quad \mathfrak{L}(v_2) := \int_{\Omega} \mathcal{L}_{c(\omega)}(v_2(\omega)) d\mathbb{P}(\omega) \quad (3.21)$$

for  $v_1 \in L^2(\Omega; X)$ ,  $v_2 \in L^2(\Omega; Y)$ . Recall that a bounded linear map  $X \rightarrow Y^*$  is equivalent to a sesquilinear (or bilinear) form on  $X \times Y$ ; see e.g. [189, Lemma 2.1.38]. To keep notation compact,

we write  $\mathcal{A}_{c(\omega)} = (\mathcal{A} \circ c)(\omega)$  and  $\mathcal{L}_{c(\omega)} = (\mathcal{L} \circ c)(\omega)$ .

**Remark 3.18** (Interpretation of the space  $\mathcal{C}$ ). *The space  $\mathcal{C}$  is the ‘space of inputs’. For the stochastic Helmholtz EDP in Section 3.1.1 the space  $\mathcal{C}$  is defined in Definition 3.47 below, but the upshot of this definition is that for any  $\omega \in \Omega$  the triple  $(A(\omega), n(\omega), f(\omega))$  is an element of  $\mathcal{C}$ . The maps  $c$ ,  $\mathcal{A}$ , and  $\mathcal{L}$  are given by  $c = (A, n, f)$ ,  $\mathcal{A} = a$ , and  $\mathcal{L} = L$ , where  $a$  and  $L$  are given by (3.6) and (3.7) respectively and the equality  $\mathcal{A} = a$  is meant in the sense of the one-to-one correspondence between  $\mathcal{B}(X, Y^*)$  and sesquilinear forms on  $X \times Y$ .*

The following three problems are the analogues in this general setting of Problems 3.1–3.3 in Section 3.1.

**Problem MAS** (Measurable variational formulation almost surely). *Find a measurable function  $u : \Omega \rightarrow X$  such that*

$$\mathcal{A}_{c(\omega)} u(\omega) = \mathcal{L}_{c(\omega)} \text{ in } Y^* \quad (3.22)$$

*almost surely.*

**Problem SOAS** (Second-order moment variational formulation almost surely).

*Find  $u \in L^2(\Omega; X)$  such that (3.22) holds almost surely.*

**Problem SV** (Stochastic variational formulation). *Find  $u \in L^2(\Omega; X)$  such that*

$$\mathfrak{A}u = \mathfrak{L} \text{ in } L^2(\Omega; Y^*). \quad (3.23)$$

**Remark 3.19** (Immediate relationships between formulations). *Since  $L^2(\Omega; X) \subseteq \mathcal{B}(\Omega, X)$  (the space of all measurable functions  $\Omega \rightarrow X$ ) it is immediate that if  $u$  solves Problem SOAS then every member of the equivalence class of  $u$  solves Problem MAS.*

### 3.2.2 Conditions on $\mathcal{A}$ , $\mathcal{L}$ , and $c$

We now state all the conditions under which we prove results about the equivalence of Problems MAS–SV.

**Condition A1** ( $\mathcal{A}$  is continuous). *The function  $\mathcal{A} : \mathcal{C} \rightarrow \mathcal{B}(X, Y^*)$  is continuous, where we place the norm topology on  $X$ , the dual norm topology on  $Y^*$ , and the operator norm topology on  $\mathcal{B}(X, Y^*)$ .*

**Condition A2** (Regularity of  $\mathcal{A} \circ c$ ). *The map  $\mathcal{A} \circ c \in L^\infty(\Omega; \mathcal{B}(X, Y^*))$ .*

We note that Condition A2 is violated in the well-studied case of a log-normal coefficient  $\chi$  for the stationary diffusion equation (3.2); in order to ensure the stochastic variational formulation is well-defined in this case, one must change the space of test functions as in [96, 157].

**Condition L1** ( $\mathcal{L}$  is continuous). *The function  $\mathcal{L} : \mathcal{C} \rightarrow Y^*$  is continuous, where we place the dual norm topology on  $Y^*$ .*

**Condition L2** (Regularity of  $\mathcal{L} \circ c$ ). *The map  $\mathcal{L} \circ c \in L^2(\Omega; Y^*)$ .*

**Condition C1** ( $c$  is measurable). *The function  $c : \Omega \rightarrow \mathcal{C}$  is measurable.*

To state the next condition, we need to recall the following definition.

**Definition 3.20** ( $\mathbb{P}$ -essentially separably valued [186, p26]). *Let  $(S, \mathcal{T}_S)$  be a topological space. A function  $h : \Omega \rightarrow S$  is  $\mathbb{P}$ -essentially separably valued if there exists  $E \in \mathcal{F}$  such that  $\mathbb{P}(E) = 1$  and  $h(E)$  is contained in a separable subset of  $S$ .*

**Condition C2** ( $c$  is  $\mathbb{P}$ -essentially separably valued). *The map  $c : \Omega \rightarrow \mathcal{C}$  is  $\mathbb{P}$ -essentially separably valued.*

**Remark 3.21** (Why do we need Condition C2?). *The theory of Bochner spaces requires strong measurability of functions (see Definitions B.9 and B.14 below). However, the proof techniques used in this chapter rely heavily on the measurability of functions (see Definition B.1 below). In separable spaces these two notions are equivalent (see Corollary B.19). However, some of the spaces we encounter (such as  $L^\infty(D_R; \mathbb{R})$ ) are not separable. Therefore, in our arguments we use Condition C2 along with the Pettis Measurability Theorem (Theorem B.18 below) to conclude that measurable functions are strongly measurable.*

**Condition B** (A priori bound almost surely). *There exist  $C_j, f_j : \Omega \rightarrow \mathbb{R}$ ,  $j = 1, \dots, m$  such that  $C_j f_j \in L^1(\Omega)$  for all  $j = 1, \dots, m$  and the bound*

$$\|u(\omega)\|_X^2 \leq \sum_{j=1}^m C_j(\omega) f_j(\omega) \quad (3.24)$$

*holds almost surely.*

**Remark 3.22** (Notation in the a priori bound). *We use the notation  $f_j$  in the right-hand side of (3.24) to emphasise the fact that typically these terms relate to the right-hand sides of the PDE in question. For the stochastic Helmholtz EDP,  $m = 1$ ,  $f_1 = \|f\|_{L^2(D)}^2$ , and  $C_1$  is given by (3.10).*

**Condition U** (Uniqueness almost surely).  *$\ker(\mathcal{A}_{c(\omega)}) = \{0\}$   $\mathbb{P}$ -almost surely.*

The condition  $\ker(\mathcal{A}_{c(\omega)}) = \{0\}$   $\mathbb{P}$ -almost surely can be stated as: given  $L \in Y^*$ , for  $\mathbb{P}$ -almost every  $\omega \in \Omega$  the deterministic problem  $\mathcal{A}_{c(\omega)} u_0 = L$  has a unique solution.

### 3.2.3 Results on the equivalence of Problems MAS, SOAS, and SV

**Theorem 3.23** (Measurable solution implies second-order solution).

*Under Condition B, if  $u$  solves Problem MAS then  $u$  solves Problem SOAS and satisfies the stochastic a priori bound*

$$\|u\|_{L^2(\Omega; X)}^2 \leq \sum_{j=1}^m \|C_j f_j\|_{L^1(\Omega)}. \quad (3.25)$$

The proof of Theorem 3.23 is on page 126 below.

Note that the right-hand side of the stochastic a priori bound (3.25) is the expectation of the right-hand side of the bound (3.24).

**Lemma 3.24** (Stochastic variational formulation well-defined).

Under Conditions A1, A2, L1, L2, C1, and C2, the maps  $\mathfrak{A}$  and  $\mathfrak{L}$  defined by (3.21) are well-defined in the sense that

$$[\mathfrak{A}(v_1)](v_2), \mathfrak{L}(v_2) < \infty \quad \text{for all } v_1 \in L^2(\Omega; X), \text{ for all } v_2 \in L^2(\Omega; Y). \quad (3.26)$$

The proof of Lemma 3.24 is on page 126 below.

**Theorem 3.25** (Second-order solution implies stochastic variational solution).

Under Conditions L1, L2, C1, and C2, if  $u$  solves Problem SOAS then  $u$  solves Problem SV.

The proof of Theorem 3.25 is on page 127 below.

**Theorem 3.26** (Stochastic variational solution implies second-order solution). *If Problem SV is well-defined and  $u$  solves Problem SV, then  $u$  solves Problem SOAS.*

The proof of Theorem 3.26 is on page 128 below.

Theorems 3.23, 3.25, and 3.26 and Lemma 3.24 are summarised in Figure 3.1.

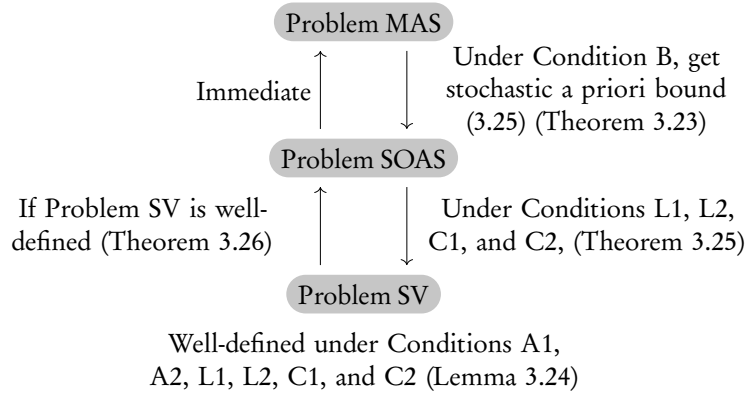


Figure 3.1: The relationship between the variational formulations. An arrow from Problem P to Problem Q with Conditions R indicates ‘under Conditions R, the solution of Problem P is a solution of Problem Q’

**Remark 3.27** (Condition L2 in Theorem 3.25). *In Theorem 3.25 we could replace Condition L2 with Condition A2, and the result would still hold—see the proof for further details. However, Condition L2 is less restrictive than Condition A2, as it only requires  $L^2$  integrability of  $\mathcal{L} \circ c$  as opposed to essential boundedness of  $\mathcal{A} \circ c$ .*

**Lemma 3.28** (Showing uniqueness of the solution to Problems MAS–SV).

If Condition U holds, then

1. the solution to Problem MAS (if it exists) is unique up to modification on a set of  $\mathbb{P}$ -measure 0 in  $\Omega$ ,
2. the solution to Problem SOAS (if it exists) is unique in  $L^2(\Omega; X)$ , and

3. if Problem SV is well-defined, the solution to Problem SV (if it exists) is unique in  $L^2(\Omega; X)$ .

The proof of Lemma 3.28 is on page 129 below.

**Remark 3.29** (Informal discussion on the ideas behind the equivalence results). *The diagram in Figure 3.1 summarises the relationships between the variational formulations, and the conditions under which they hold. Moving ‘up’ the left-hand side of the diagram, we prove a solution of Problem SV is a solution of Problem SOAS in Theorem 3.26; the key idea in this theorem is to use a particular set of test functions and the general measure-theory result of Lemma B.22 below; this approach was used for the stationary diffusion equation (3.2) with log-normal coefficients in [96], and for a wider class of coefficients in [157].*

*Moving ‘down’ the right-hand side, we prove a solution of Problem MAS is a solution of Problem SOAS in Theorem 3.23; the key part of this proof is that the bound in Condition B gives information on the integrability of the solution  $u$ . (In the case of (3.2) with uniformly coercive and bounded coefficient  $\kappa$ , the analogous integrability result follows from the Lax–Milgram theorem; [41, Proposition 2.4] proves an equivalent result for (3.2) with lognormal coefficient  $\kappa$  with an isotropic Lipschitz covariance function.) Proving a solution of Problem SOAS is a solution of Problem SV in Theorem 3.25 essentially amounts to posing conditions such that the quantities  $[\mathcal{A}_{c(\omega)}(u(\omega))](v(\omega))$  and  $\mathcal{L}_{c(\omega)}(v(\omega))$  are Bochner integrable for any  $v \in L^2(\Omega; Y)$ , so that (3.23) makes sense. Lemma 3.24 shows that the stronger property (3.26) holds, and requires stronger assumptions than Theorem 3.25, since the proof of Theorem 3.25 uses the additional information that  $u$  solves Problem SOAS.*

**Remark 3.30** (Changing the condition  $u \in L^2(\Omega; X)$ ). *Here we seek the solution  $u \in L^2(\Omega; X)$  but we could instead require  $u \in L^p(\Omega; X)$ , for some  $p > 0$  and require  $\mathfrak{A}u = \mathfrak{L}$  in  $L^q(\Omega; Y)^*$ , for some  $q > 0$  (i.e. use test functions in  $L^q(\Omega; Y)$ ). In this case, the proof of Theorem 3.26 would be nearly identical, as the space  $\mathcal{D}$  of test functions used there is a subset of  $L^q(\Omega; Y)$  for all  $q > 0$ . One could also develop analogues of Theorems 3.23 and 3.25 and Lemma 3.24 in this setting—see, e.g., [96, Theorem 3.20] for an example of this approach for the stationary diffusion equation with lognormal diffusion coefficient.*

**Remark 3.31** (Non-reliance on the Lax–Milgram theorem). *The above results hold for an arbitrary sesquilinear form and hence are applicable to a wide variety of PDEs; their main advantage is that they apply to PDEs whose stochastic variational formulations are not coercive. For example, as noted in Section 3.1, for the stationary diffusion equation (3.2) with coefficient  $\kappa$  bounded uniformly below in  $\omega$ , the bilinear form of Problem SV is coercive; existence and uniqueness follow from the Lax–Milgram theorem, and hence the chain of results above leading to the well-posedness of Problem SV is not necessary.*

**Remark 3.32** (Overview of how these results are applied to the Helmholtz equation in Section 3.4). *We obtain the results for the Helmholtz equation via the following steps (which could also be applied to other SPDEs fitting into this framework):*

1. Define the map  $c$  (via  $A, n$ , and  $f$ ) such that for almost every  $\omega \in \Omega$  there exists a solution of the deterministic Helmholtz EDP corresponding to  $c(\omega)$ .

2. Define  $u : \Omega \rightarrow X$  to map  $\omega$  to the solution of the deterministic problem corresponding to  $c(\omega)$ .
3. Prove that Conditions A1, A2, L1, L2, C1, C2, B, and U hold, so that one can apply Theorems 3.23, 3.25, and 3.26 along with Lemmas 3.24 and 3.28 to show Problem 3.3 is well-defined and  $u$  is unique and satisfies Problems 3.1–3.3.

Steps 1 and 2 can be thought of as constructing a solution pathwise.

### 3.3 PROOF OF THE RESULTS IN SECTION 3.2

A key ingredient in proving that the stochastic variational formulation is well-defined (Lemma 3.24) is showing that the maps  $\omega \mapsto [(\mathcal{L} \circ c)(\omega)](v_2(\omega))$  and  $\omega \mapsto [[(\mathcal{A} \circ c(\omega)](v_1(\omega))](v_2(\omega))$  are measurable, for appropriate functions  $v_1$  and  $v_2$ . Showing that these functions are measurable is not straightforward, because they both depend on  $\omega$  in multiple places. However, the structure of the  $\omega$ -dependence in each case is similar, and so we first prove some general results that will be applicable to both of these cases.

#### 3.3.1 Preliminary lemmas

Throughout this section, we assume we have two separable Banach spaces  $Z_1$  and  $Z_2$ , and maps  $\mathcal{P} : \Omega \rightarrow \mathbb{B}(Z_1, Z_2)$  and  $v : \Omega \rightarrow Z_1$ . To simplify notation, we introduce the following definition.

**Definition 3.33** (Pairing map). *We define the map  $\pi_{\mathcal{P},v} : \Omega \rightarrow Z_2$  by*

$$\pi_{\mathcal{P},v}(\omega) := [\mathcal{P}(\omega)](v(\omega)). \quad (3.27)$$

**Definition 3.34** (Product map). *Let  $P_{\mathcal{P},v} : \Omega \rightarrow \mathbb{B}(Z_1, Z_2) \times Z_1$  be defined by*

$$P_{\mathcal{P},v}(\omega) = (\mathcal{P}(\omega), v(\omega)).$$

**Lemma 3.35** (Product map is measurable). *When  $\mathbb{B}(Z_1, Z_2) \times Z_1$  is equipped with the product topology, if  $\mathcal{P}$  and  $v$  are measurable, then  $P_{\mathcal{P},v} : \Omega \rightarrow \mathbb{B}(Z_1, Z_2) \times Z_1$  is measurable.*

*Proof of Lemma 3.35.* By the result on the measurability of the Cartesian product of measurable functions (Lemma B.6),  $P_{\mathcal{P},v}$  is measurable with respect to  $(\mathcal{F}, \mathcal{B}(\mathbb{B}(Z_1, Z_2)) \otimes \mathcal{B}(Z_1))$  (where  $\mathcal{B}$  denotes the Borel  $\sigma$ -algebra—see Definition B.2), as both of the coordinate functions  $\mathcal{P}$  and  $v$  are measurable. Since  $\mathbb{B}(Z_1, Z_2)$  and  $Z_1$  are both metric spaces, they are both Hausdorff. As  $Z_1$  is separable, Lemma B.7 on the product of Borel  $\sigma$ -algebras implies  $\mathcal{B}(\mathbb{B}(Z_1, Z_2)) \otimes \mathcal{B}(Z_1) = \mathcal{B}(\mathbb{B}(Z_1, Z_2) \times Z_1)$ . Hence  $P_{\mathcal{P},v}$  is measurable with respect to  $(\mathcal{F}, \mathcal{B}(\mathbb{B}(Z_1, Z_2) \times Z_1))$ .  $\square$

**Definition 3.36** (Evaluation map). *The function  $\eta_{Z_1, Z_2} : \mathbb{B}(Z_1, Z_2) \times Z_1 \rightarrow Z_2$  is defined by*

$$\eta_{Z_1, Z_2}((\mathcal{H}, v)) := \mathcal{H}(v) \quad \text{for } \mathcal{H} \in \mathbb{B}(Z_1, Z_2) \text{ and } v \in Z_1. \quad (3.28)$$

Observe that the pairing, product, and evaluation maps ( $\pi_{\mathcal{P},v}$ ,  $P_{\mathcal{P},v}$ , and  $\eta_{Z_1,Z_2}$  respectively) are related by  $\pi_{\mathcal{P},v} = \eta_{Z_1,Z_2} \circ P_{\mathcal{P},v}$ .

**Lemma 3.37** (Evaluation map is continuous). *The map  $\eta_{Z_1,Z_2}$  is continuous with respect to the product topology on  $B(Z_1, Z_2) \times Z_1$  and the norm topology on  $Z_2$ .*

The proof of Lemma 3.37 is straightforward and omitted.

**Lemma 3.38** (Pairing map is measurable). *If  $\mathcal{P}$  and  $v$  are measurable, then  $\pi_{\mathcal{P},v}$  is measurable.*

*Proof of Lemma 3.38.* By Lemma 3.35  $P_{\mathcal{P},v}$  is measurable and by Lemma 3.37  $\eta_{Z_1,Z_2}$  is continuous. Therefore Lemma B.4 implies that  $\pi_{\mathcal{P},v} = \eta_{Z_1,Z_2} \circ P_{\mathcal{P},v}$  is measurable.  $\square$

**Lemma 3.39** ( $(\mathcal{L} \circ c)(v)$  is measurable). *Under Conditions L1 and C1, for any measurable  $v_2 : \Omega \rightarrow Y$ , the function  $\omega \mapsto [(\mathcal{L} \circ c(\omega))(v(\omega))]$  is measurable.*

*Proof of Lemma 3.39.* The map  $c$  is measurable (by Condition C1) and  $\mathcal{L}$  is continuous (by Condition L1), therefore Lemma B.4 implies that  $\mathcal{L} \circ c$  is measurable. Applying Lemma 3.38 with  $Z_1 = Y$ ,  $Z_2 = \mathbb{F}$  (because  $Y^* = B(Y, \mathbb{F})$ ),  $\mathcal{P} = \mathcal{L} \circ c$ , and  $v = v_2$ , the result follows.  $\square$

**Lemma 3.40** ( $((\mathcal{A} \circ c)(v_1))(v_2)$  is measurable). *If Conditions A1 and C1 hold and  $v_1 : \Omega \rightarrow X$  and  $v_2 : \Omega \rightarrow Y$  are measurable, then the function  $\omega \mapsto [(\mathcal{A} \circ c(\omega))(v_1(\omega))](v_2(\omega))$  is measurable.*

*Proof of Lemma 3.40.* Since Conditions A1 and C1 hold,  $\mathcal{A} \circ c$  is measurable by Lemma B.4. Therefore by Lemma 3.38 with  $Z_1 = X$ ,  $Z_2 = Y^*$ ,  $\mathcal{P} = \mathcal{A} \circ c$  and  $v = v_1$ , the map  $\omega \rightarrow [(\mathcal{A} \circ c(\omega))(v_1(\omega))]$  is measurable. Therefore applying Lemma 3.38 again with  $Z_1 = Y$ ,  $Z_2 = \mathbb{F}$ ,  $\mathcal{P}(\omega) = [(\mathcal{A} \circ c(\omega))(v_1(\omega))]$ , and  $v = v_2$ , the result follows.  $\square$

### 3.3.2 Proofs of Theorems 3.23, 3.25, and 3.26 and Lemmas 3.24 and 3.28

*Proof of Theorem 3.23.* We need to show  $u : \Omega \rightarrow X$  is strongly measurable, satisfies the bound (3.25), and therefore is Bochner integrable and is in the space  $L^2(\Omega; X)$ . Our plan is to use Corollary B.12 to show  $u$  is Bochner integrable, and establish (3.25) as a by-product. Since  $u$  solves Problem MAS,  $u$  is measurable. As  $X$  is separable, it follows from Corollary B.19 that  $u$  is strongly measurable. Define  $N : X \rightarrow \mathbb{R}$  by  $N(v) := \|v\|_X^2$ . Since  $N$  is continuous, Lemma B.4 implies  $N \circ u : \Omega \rightarrow \mathbb{R}$  is measurable. Therefore, since both the left- and right-hand sides of (3.24) are measurable and (3.24) holds for almost every  $\omega \in \Omega$  we can integrate (3.24) over  $\Omega$  with respect to  $\mathbb{P}$  and obtain

$$\int_{\Omega} \|u(\omega)\|_X^2 d\mathbb{P}(\omega) \leq \sum_{j=1}^m \|C_j f_j\|_{L^1(\Omega)}, \quad (3.29)$$

the right-hand side of which is finite since Condition B includes that  $C_j f_j \in L^1(\Omega)$  for all  $j = 1, \dots, m$ . Since  $u$  is strongly measurable, the bound (3.29) and Corollary B.12 with  $p = 2$  imply that  $u$  is Bochner integrable. The norm  $\|u\|_{L^2(\Omega; X)}$  is thus well-defined by Definition B.13 and (3.29) shows that (3.25) holds, and so in particular  $\|u\|_{L^2(\Omega; X)} < \infty$ .  $\square$

*Proof of Lemma 3.24.* We must show that for any  $v_1 \in L^2(\Omega; X)$  and any  $v_2 \in L^2(\Omega; Y)$ :

- The quantities  $[\mathcal{A}_{c(\omega)}v_1(\omega)](v_2(\omega))$  and  $\mathcal{L}_{c(\omega)}(v_2(\omega))$  are Bochner integrable, so that the definitions of  $\mathfrak{A}$  and  $\mathfrak{L}$  as integrals over  $\Omega$  make sense.
- The maps  $\mathfrak{A}(v_1)$  and  $\mathfrak{L}$  are linear and bounded on  $L^2(\Omega; Y)$ , that is,  $\mathfrak{A} : L^2(\Omega; X) \rightarrow L^2(\Omega; Y)^*$  and  $\mathfrak{L} \in L^2(\Omega; Y)^*$ .

It follows from these two points that  $\mathfrak{A}$  and  $\mathfrak{L}$  are well-defined. Thanks to the groundwork laid in Section 3.3.1,  $[\mathcal{A}_{c(\omega)}v_1(\omega)](v_2(\omega))$  and  $\mathcal{L}_{c(\omega)}(v_2(\omega))$  are measurable by Lemmas 3.39 and 3.40 (which need Conditions A1, L1, and C2). Their  $\mathbb{P}$ -essential separability follows from Conditions A1, L1, and C2 and Lemma B.20 and thus their strong measurability follows from Corollary B.19 on the equivalence of measurability and strong measurability when the image is separable. Their Bochner integrability then follows from the Bochner integrability condition in Theorem B.11 (with  $V = \mathbb{F}$ ) and the Cauchy–Schwarz inequality since

$$\begin{aligned} \int_{\Omega} \left| \mathcal{L}_{c(\omega)}(v_2(\omega)) \right| d\mathbb{P}(\omega) &\leq \int_{\Omega} \|(\mathcal{L} \circ c)(\omega)\|_{Y^*} \|v_2(\omega)\|_Y d\mathbb{P}(\omega) \\ &\leq \|\mathcal{L} \circ c\|_{L^2(\Omega; Y^*)} \|v_2\|_{L^2(\Omega; Y)}, \end{aligned} \quad (3.30)$$

which is finite by Condition L2, and

$$\begin{aligned} \int_{\Omega} \left| [\mathcal{A}_{c(\omega)}v_1(\omega)](v_2(\omega)) \right| d\mathbb{P}(\omega) &\leq \operatorname{ess\,sup}_{\omega \in \Omega} \left\| \mathcal{A}_{c(\omega)} \right\|_{\mathbb{B}(X, Y^*)} \int_{\Omega} \|v_1(\omega)\|_X \|v_2(\omega)\|_Y d\mathbb{P}(\omega) \\ &\leq \|\mathcal{A} \circ c\|_{L^\infty(\Omega; \mathbb{B}(X, Y^*))} \|v_1\|_{L^2(\Omega; X)} \|v_2\|_{L^2(\Omega; Y)}, \end{aligned} \quad (3.31)$$

which is finite by Condition A2.

We now show  $\mathfrak{L} \in L^2(\Omega; Y)^*$  and  $\mathfrak{A} : L^2(\Omega; X) \rightarrow L^2(\Omega; Y)^*$ . Observe that  $|\mathfrak{L}(v_2)| \leq \int_{\Omega} \left| \mathcal{L}_{c(\omega)}(v_2(\omega)) \right| d\mathbb{P}(\omega)$  and  $|\mathfrak{A}(v_1)| \leq \int_{\Omega} \left| [\mathcal{A}_{c(\omega)}v_1(\omega)](v_2(\omega)) \right| d\mathbb{P}(\omega)$  and thus by (3.30) and (3.31)  $\mathfrak{L}$  and  $\mathfrak{A}(v_1)$  are bounded. They are clearly linear, and so it follows that  $\mathfrak{L} \in L^2(\Omega; Y)^*$  and  $\mathfrak{A}(v_1) \in L^2(\Omega; Y)^*$ , i.e.,  $\mathfrak{A} : L^2(\Omega; X) \rightarrow L^2(\Omega; Y)^*$ .  $\square$

*Proof of Theorem 3.25.* In order to show that  $u$  solves Problem SV, we must show:

1. either the functional  $\mathfrak{L} \in L^2(\Omega; Y)^*$  or the functional  $\mathfrak{A}(u) \in L^2(\Omega; Y)^*$ , and
2. the equality (3.23) holds.

For Point 1 we show that  $\mathfrak{L} \in L^2(\Omega; Y)^*$ , (since this is easier than showing  $\mathfrak{A}(u) \in L^2(\Omega; Y)^*$ ); in fact the proof of this is contained in the proof of Lemma 3.24.

For Point 2, since  $u$  solves Problem SOAS, for  $\mathbb{P}$ -almost every  $\omega \in \Omega$  we have  $\mathcal{A}_{c(\omega)}u(\omega) = \mathcal{L}_{c(\omega)}$  in  $Y^*$ . Hence, for any  $v \in L^2(\Omega; Y)$  we have

$$[\mathcal{A}_{c(\omega)}u(\omega)](v(\omega)) = \mathcal{L}_{c(\omega)}(v(\omega)) \quad (3.32)$$

for  $\mathbb{P}$ -almost every  $\omega \in \Omega$ . Since  $\mathfrak{L} \in L^2(\Omega; Y)^*$ , the right-hand side of (3.32) is a strongly measurable function with finite integral. Hence the left-hand side of (3.32) is as well, and we can integrate over  $\Omega$  to conclude  $[\mathfrak{A}u](v) = \mathfrak{L}(v)$  for all  $v \in L^2(\Omega; Y)$ , that is,  $\mathfrak{A}u = \mathfrak{L}$  in  $L^2(\Omega; Y)^*$ .  $\square$

The following lemma is needed for the proof of Theorem 3.26.

**Lemma 3.41.** *Let  $\delta : \Omega \times Y \rightarrow \mathbb{F}$ . For  $y \in Y$ , define  $\Omega_y := \{\omega \in \Omega : \delta(\omega, y) = 0\}$  and define  $\tilde{\Omega} := \{\omega \in \Omega : \delta(\omega, y) = 0 \text{ for all } y \in Y\}$ . If*

- *for all  $\omega \in \Omega$ ,  $\delta(\omega, \cdot)$  is a continuous functional on  $Y$  and*
- *for all  $y \in Y$ , the map  $\delta(\cdot, y) : \Omega \rightarrow \mathbb{F}$  is measurable and  $\mathbb{P}(\Omega_y) = 1$ ,*

*then  $\mathbb{P}(\tilde{\Omega}) = 1$ .*

*Proof of Lemma 3.41.* We must show that the set  $\tilde{\Omega} \in \mathcal{F}$ , and  $\mathbb{P}(\tilde{\Omega}) = 1$ . Observe that, for any  $y \in Y$ , the set  $\Omega_y \in \mathcal{F}$ , since  $\Omega_y = \delta(\cdot, y)^{-1}(\{0\})$ , which is the preimage under a measurable map of a measurable set.

Since  $Y$  is a Hilbert space, it is separable, and therefore it has a countable dense subset  $(y_n)_{n \in \mathbb{N}}$ . We will show that  $\mathbb{P}(\cap_{n \in \mathbb{N}} \Omega_{y_n}) = 1$  and  $\tilde{\Omega} = \cap_{n \in \mathbb{N}} \Omega_{y_n}$ . The set  $\cap_{n \in \mathbb{N}} \Omega_{y_n} \in \mathcal{F}$ , as  $\mathcal{F}$  is a  $\sigma$ -algebra and  $\mathbb{P}(\cup_{n \in \mathbb{N}} \Omega_{y_n}^c) \leq \sum_{n \in \mathbb{N}} \mathbb{P}(\Omega_{y_n}^c) = 0$ , and hence  $\mathbb{P}(\cap_{n \in \mathbb{N}} \Omega_{y_n}) = 1$ . To next show  $\tilde{\Omega} = \cap_{n \in \mathbb{N}} \Omega_{y_n}$  we observe that  $\tilde{\Omega} = \cap_{y \in Y} \Omega_y$  and  $\cap_{y \in Y} \Omega_y \subseteq \cap_{n \in \mathbb{N}} \Omega_{y_n}$ . It therefore suffices to show  $\cap_{n \in \mathbb{N}} \Omega_{y_n} \subseteq \cap_{y \in Y} \Omega_y$  to conclude  $\tilde{\Omega} = \cap_{n \in \mathbb{N}} \Omega_{y_n}$ .

Fix  $y \in Y$ . By density of  $(y_n)_{n \in \mathbb{N}}$ , there exists a subsequence  $(y_{n_m})_{m \in \mathbb{N}}$  such that  $y_{n_m} \rightarrow y$  as  $m \rightarrow \infty$ . Fix  $\omega \in \cap_{n \in \mathbb{N}} \Omega_{y_n}$ . Note that  $\omega \in \cap_{m \in \mathbb{N}} \Omega_{y_{n_m}}$ ; that is, for all  $m \in \mathbb{N}$ ,  $\delta(\omega, y_{n_m}) = 0$ . As  $\delta(\omega, \cdot)$  is a continuous function on  $Y$ ,  $\delta(\omega, y_{n_m}) \rightarrow \delta(\omega, y)$  as  $m \rightarrow \infty$ . But as previously noted,  $\delta(\omega, y_{n_m}) = 0$  for all  $m \in \mathbb{N}$ . Hence we must have  $\delta(\omega, y) = 0$ , and thus  $\omega \in \Omega_y$ . Since  $\omega \in \cap_{n \in \mathbb{N}} \Omega_{y_n}$  was arbitrary, it follows that  $\cap_{n \in \mathbb{N}} \Omega_{y_n} \subseteq \Omega_y$ , and since  $y \in Y$  was arbitrary, it follows that  $\cap_{n \in \mathbb{N}} \Omega_{y_n} \subseteq \cap_{y \in Y} \Omega_y$  as required.  $\square$

*Proof of Theorem 3.26.* Let  $u \in L^2(\Omega; X)$  solve Problem SV. We need to show that  $u$  solves Problem SOAS. Observe that  $u$  solving Problem SOAS means  $\mathcal{A}_{c(\omega)}(u(\omega)) = (\mathcal{L}_{c(\omega)})(\omega)$  in  $Y^*$  for almost every  $\omega \in \Omega$ . We now use an idea from [96, Theorem 3.3]. Our plan is to use test functions of the form  $y \mathbb{1}_E$ , where  $y \in Y$  and  $E \in \mathcal{F}$  to reduce Problem SV to the statement

$$\int_E [\mathcal{A}_{c(\omega)}(u(\omega))](y(\omega)) d\mathbb{P}(\omega) = \int_E [(\mathcal{L}_{c(\omega)})(\omega)](y(\omega)) d\mathbb{P}(\omega) \quad \text{for all } E \in \mathcal{F}$$

and then show this implies  $u$  satisfies Problem SOAS via Lemma B.22.

First define the space  $\mathcal{D} := \{y \mathbb{1}_E : y \in Y, E \in \mathcal{F}\}$ . It is straightforward to see that the elements of  $\mathcal{D}$  are maps from  $\Omega$  to  $Y$ . The fact that  $\mathcal{D} \subseteq L^2(\Omega; Y)$  follows via the following three steps:

1. The elements of  $\mathcal{D}$  are measurable, indeed the indicator function of a measurable set is a measurable function  $\Omega \rightarrow \mathbb{R}$ , and multiplication by  $y \in Y$  is a continuous function  $\mathbb{R} \rightarrow Y$ . Hence elements of  $\mathcal{D}$  are measurable by Lemma B.4.
2. As  $Y$  is a separable Hilbert space, it follows from Corollary B.19 that the elements of  $\mathcal{D}$  are strongly measurable.
3.  $\|y \mathbb{1}_E\|_{L^2(\Omega; Y)} = \sqrt{\mathbb{P}(E)} \|y\|_Y < \infty$  for all  $y \in Y, E \in \mathcal{F}$ .

Since Problem SV is well-defined, and  $u$  solves Problem SV, and  $\mathcal{D} \subseteq L^2(\Omega; Y)$ , we have that  $[\mathfrak{A}u](v) = \mathfrak{L}(v)$  for all  $v \in \mathcal{D}$ . Therefore, we have

$$\int_{\Omega} [\mathcal{A}_{c(\omega)}(u(\omega))](y \mathbb{1}_E(\omega)) d\mathbb{P}(\omega) = \int_{\Omega} [\mathcal{L}_{c(\omega)}](y \mathbb{1}_E(\omega)) d\mathbb{P}(\omega) \quad (3.33)$$

for all  $y \in Y$  and  $E \in \mathcal{F}$ . If we define  $\delta : \Omega \times Y \rightarrow \mathbb{F}$  by  $\delta(\omega, y) := [\mathcal{A}_{c(\omega)}(u(\omega)) - \mathcal{L}_{c(\omega)}](y)$  then, by the definition of  $\mathbb{1}_E$ , (3.33) becomes

$$\int_E \delta(\omega, y) d\mathbb{P}(\omega) = 0 \quad \text{for all } E \in \mathcal{F}. \quad (3.34)$$

To conclude  $u$  solves Problem SOAS we must show  $\delta(\omega, y) = 0$  for all  $y \in Y$ , almost surely. We will use Lemma B.22, so the first step is to show that for all  $y \in Y$   $\delta(\cdot, y)$  is Bochner integrable. This follows from the fact that Problem SV is well-defined, and thus the quantities  $[\mathcal{A}_{c(\omega)}v_1(\omega)](v_2(\omega))$  and  $\mathcal{L}_{c(\omega)}(v_2(\omega))$  are Bochner integrable for any  $v_1 \in L^2(\Omega; X)$ ,  $v_2 \in L^2(\Omega; Y)$ . In particular, they are Bochner integrable when  $v_1 = u$ , and  $v_2 = y \mathbb{1}_E$  and thus their difference  $\delta$  is Bochner integrable. Secondly,  $\delta(\omega, \cdot)$  is a continuous function on  $Y$  since  $\mathcal{A}_{c(\omega)}(u(\omega)), (\mathcal{L}_{c(\omega)})(\omega) \in Y^*$ , for all  $\omega \in \Omega$ .

We now show  $\delta(\omega, y) = 0$  for all  $y \in Y$ , almost surely. For  $y \in Y$  define the set  $\Omega_y := \{\omega \in \Omega : \delta(\omega, y) = 0\}$ ; by (3.34) and Lemma B.22 we have that  $\mathbb{P}(\Omega_y) = 1$  for all  $y \in Y$ . By Lemma 3.41,  $\delta(\omega, y) = 0$  for all  $y \in Y$ , almost surely, that is,  $\mathcal{A}_{c(\omega)}u(\omega) = \mathcal{L}_{c(\omega)}$  almost surely; it follows that  $u$  solves Problem SOAS.  $\square$

**Remark 3.42** (Connection with the argument in [157, Remark 2.2]). *The argument in Lemma 3.41 and the final part of Theorem 3.26 closely mirrors the result in [157, Remark 2.2]. Indeed, we prove in general that*

$$\mathbb{P}(\delta(\omega, y) = 0) = 1 \text{ for all } y \in Y \quad \text{implies} \quad \mathbb{P}(\delta(\omega, y) = 0 \text{ for all } y \in Y) = 1,$$

and [157, Remark 2.2] shows an analogous result for the stationary diffusion equation (3.2) with non-uniformly coercive and unbounded coefficient  $\kappa$ .

*Proof of Lemma 3.28. Proof of Part 1.* Suppose  $u_1, u_2 : \Omega \rightarrow X$  solve Problem MAS. Let  $E = \{\omega \in \Omega : u_1(\omega) \neq u_2(\omega)\}$ . Denote by  $E_1$  and  $E_2$  the sets (of measure zero) where the variational problems for  $u_1$  and  $u_2$  fail to hold, i.e.  $E_1, E_2 \in \mathcal{F}$  with  $\mathbb{P}(E_1) = \mathbb{P}(E_2) = 0$  and

$$\mathcal{A}_{c(\omega)}(u_1(\omega)) \neq \mathcal{L}_{c(\omega)} \text{ iff } \omega \in E_1, \quad \text{and} \quad \mathcal{A}_{c(\omega)}(u_2(\omega)) \neq \mathcal{L}_{c(\omega)} \text{ iff } \omega \in E_2.$$

As  $\ker(\mathcal{A}_{c(\omega)}) = \{0\}$   $\mathbb{P}$ -almost surely, there exists  $E_3 \in \mathcal{F}$  such that  $\mathbb{P}(E_3) = 0$  and  $\ker(\mathcal{A}_{c(\omega)}) \neq \{0\}$  iff  $\omega \in E_3$ . We claim  $E \subseteq E_1 \cup E_2 \cup E_3$ . Indeed, if  $u_1(\omega) \neq u_2(\omega)$  then either: (i) at least one of  $u_1$  and  $u_2$  does not solve Problem MAS at  $\omega$  or (ii)  $u_1$  and  $u_2$  both solve Problem MAS at  $\omega$ , but  $\ker(\mathcal{A}_{c(\omega)}) \neq \{0\}$ . Since  $\mathbb{P}(E_j) = 0$ ,  $j = 1, 2, 3$ , we have  $\mathbb{P}(E_1 \cup E_2 \cup E_3) = 0$ . Therefore  $E \in \mathcal{F}$  and  $\mathbb{P}(E) = 0$  since  $(\Omega, \mathcal{F}, \mathbb{P})$  is a complete probability space; hence  $u_1 = u_2$

almost surely, as required.

*Proof of Part 2.* By Remark 3.19, if  $u_1, u_2 \in L^2(\Omega; X)$  solve Problem SOAS, then all the representatives of the equivalence classes of  $u_1$  and  $u_2$  solve Problem MAS. Hence, by Part 1, any representative of  $u_1$  and any representative of  $u_2$  differ only on some set (depending on the representatives) of  $\mathbb{P}$ -measure zero in  $\Omega$ . Therefore  $u_1 = u_2$  in  $L^2(\Omega; X)$ , by definition of  $L^2(\Omega; X)$ .

*Proof of Part 3.* As Problem SV is well-defined, by Remark 3.19 and Theorem 3.26, if  $u_1$  and  $u_2$  solve Problem SV, then  $u_1$  and  $u_2$  also solve Problem MAS. We then repeat the reasoning in the proof of Part 2 to show  $u_1 = u_2$  in  $L^2(\Omega; X)$ .  $\square$

### 3.4 PROOFS OF THEOREMS 3.7 AND 3.10

In Section 3.4.1 we place the Helmholtz stochastic EDP into the framework developed in Section 3.2. In Section 3.4.2 we give sufficient conditions for the Helmholtz stochastic EDP to satisfy Conditions A1, L1, C1, etc.. In Section 3.4.3 we apply the general theory developed in Section 3.2 to prove Theorems 3.7 and 3.10.

#### 3.4.1 Placing the Helmholtz stochastic EDP into the framework of Section 3.2

Recall  $R > 0$  is fixed. We let  $X = Y = H_{0,D}^1(D_R)$  and define the norm  $\|v\|_{H_k^1(D_R)}^2 := \|\nabla v\|_{L^2(D_R)}^2 + k^2 \|v\|_{L^2(D_R)}^2$  on  $H_{0,D}^1(D_R)$ . Throughout this section,  $A_0, n_0$ , and  $f_0$  will be deterministic functions. Recall that since the supports of  $1 - n$ ,  $I - A$ , and  $f$  are compactly contained in  $B_R$ , we can consider  $A, n$ , and  $f$  as functions on  $D_R$  rather than on  $D_+$ . In order to define the space  $\mathcal{C}$  and the maps  $c, \mathcal{A}$ , and  $\mathcal{L}$  we define the following function spaces on  $D_R$ .

**Definition 3.43** (Compact-support spaces). *Let*

$$\begin{aligned} L_R^2(D_R) &:= \{f_0 \in L^2(D_R) : \text{supp}(f_0) \subset\subset B_R\}, \\ L_{R,\min}^\infty(D_R; \mathbb{R}) &:= \{n_0 \in L^\infty(D_R; \mathbb{R}) : \text{supp}(1 - n_0) \subset\subset B_R, \\ &\quad \text{there exists } \alpha_{n_0} > 0 \text{ such that } n_0(\mathbf{x}) \geq \alpha_{n_0} \text{ almost everywhere}\}, \\ L_R^\infty(D_R; \text{SPD}) &:= \{A_0 \in L^\infty(D_R; \text{SPD}) : \text{supp}(I - A_0) \subset\subset B_R\}, \text{ and} \\ W_R^{1,\infty}(D_R; \text{SPD}) &:= \{A_0 \in L_R^\infty(D_R; \text{SPD}) : A_0 \in W^{1,\infty}(D_R; \text{SPD})\}. \end{aligned}$$

Observe that the norms on  $L^\infty(D_R; \mathbb{R})$ ,  $L^\infty(D_R; \mathbb{R}^{d \times d})$ ,  $W^{1,\infty}(D_R; \mathbb{R}^{d \times d})$ , and  $L^2(D_R)$  induce metrics on  $L_{R,\min}^\infty(D_R; \mathbb{R})$ ,  $L_R^\infty(D_R; \text{SPD})$ ,  $W_R^{1,\infty}(D_R; \text{SPD})$ , and  $L_R^2(D_R)$  respectively. These spaces are not vector spaces, and are not complete, but completeness and being a vector space is not required in what follows—we only need them to be metric spaces.

**Definition 3.44** (Deterministic form and functional). *Given some  $(A_0, n_0, f_0) \in L_R^\infty(D_R; \text{SPD}) \times L_{R,\min}^\infty(D_R; \mathbb{R}) \times L_R^2(D_R)$ , let the sesquilinear form  $a_{A_0, n_0}$  on  $H_{0,D}^1(D_R) \times H_{0,D}^1(D_R)$  and the antilinear*

functional  $L_{f_0}$  on  $H_{0,D}^1(D_R)$  be given by

$$a_{A_0, n_0}(v_1, v_2) := \int_{D_R} \left( (A_0 \nabla v_1) \cdot \nabla \overline{v_2} - k^2 n_0 v_1 \overline{v_2} \right) d\lambda - \langle T_R \gamma v_1, \gamma v_2 \rangle_{\Gamma_R}, \quad \text{and}$$

$$L_{f_0}(v_2) := \int_{D_R} f_0 \overline{v_2} d\lambda, \quad \text{for } v_1, v_2 \in H_{0,D}^1(D_R).$$

We now restate Problem 2.10 in the notation of this chapter.

**Problem 3.45** (Helmholtz EDP). For  $(A_0, n_0, f_0) \in L_R^\infty(D_R; \text{SPD}) \times L_R^\infty(D_R; \mathbb{R}) \times L_R^2(D_R)$  find  $u_0 \in H_{0,D}^1(D_R)$  such that  $a_{A_0, n_0}(u_0, v) = L_{f_0}(v)$  for all  $v \in H_{0,D}^1(D_R)$ .

**Definition 3.46** ( $d_\infty$  metric). Let  $(X_1, d_1), \dots, (X_m, d_m)$  be metric spaces. The  $d_\infty$  metric on the Cartesian product  $X_1 \times \dots \times X_m$  is defined by

$$d_\infty((x_1, \dots, x_m), (y_1, \dots, y_m)) := \max_{j=1, \dots, m} d_j(x_j, y_j).$$

**Definition 3.47** (The input space  $\mathcal{C}$ ). We let  $\mathcal{C} := W_R^{1,\infty}(D_R; \text{SPD}) \times L_{R,\min}^\infty(D_R; \mathbb{R}) \times L_R^2(D_R)$  with topology given by the  $d_\infty$  metric.

**Definition 3.48** (The input map  $c$ ). Define  $c : \Omega \rightarrow \mathcal{C}$  by  $c(\omega) = (A(\omega), n(\omega), f(\omega))$ .

**Definition 3.49** (The maps  $\mathcal{A}$  and  $\mathcal{L}$  for the Helmholtz stochastic EDP). Let

$$\mathcal{A}((A_0, n_0, f_0)) := a_{A_0, n_0} \quad \text{and} \quad \mathcal{L}((A_0, n_0, f_0)) := L_{f_0}, \quad (3.35)$$

where the definition of  $\mathcal{A}$  is understood in terms of the equivalence between  $B(X, Y^*)$  and sesquilinear forms on  $X \times Y$ .

### 3.4.2 Verifying the Helmholtz stochastic EDP satisfies the general conditions in Section 3.2

**Lemma 3.50** (Conditions C1 and C2 for Helmholtz stochastic EDP). If  $A, n$ , and  $f$  are strongly measurable, then  $c$  defined by Definition 3.48 satisfies Conditions C1 and C2.

*Proof.* Since  $A, n$ , and  $f$  are strongly measurable, by Theorem B.18 they are measurable and  $\mathbb{P}$ -essentially separably valued. By Lemma B.6, it follows that  $c$  is measurable, so  $c$  satisfies Condition C1. By Lemma B.23, it follows that  $c$  is  $\mathbb{P}$ -essentially separably valued, so  $c$  satisfies Condition C2.  $\square$

**Lemma 3.51** (Conditions A1 and L1 for Helmholtz stochastic EDP). The two maps  $\mathcal{A}$  and  $\mathcal{L}$  given by (3.35) satisfy Conditions A1 and L1.

*Proof of Lemma 3.51.* We need to show that if we have  $(A_m, n_m, f_m) \rightarrow (A_0, n_0, f_0)$  in  $\mathcal{C}$ , then  $\mathcal{A}((A_m, n_m, f_m)) \rightarrow \mathcal{A}((A_0, n_0, f_0))$  in  $B(X, Y^*)$ , and similarly for  $\mathcal{L}$ . We have, for  $v_1 \in X, v_2 \in$

$Y$ ,

$$\begin{aligned}
& \left| \left[ \left[ \mathcal{A}(A_m, n_m, f_m) - \mathcal{A}(A_0, n_0, f_0) \right](v_1) \right](v_2) \right| \\
&= \left| \int_{D_R} \left( (A_m - A_0) \nabla v_1 \cdot \nabla \bar{v}_2 - k^2 (n_m - n_0) v_1 \bar{v}_2 \right) d\lambda \right| \\
&\leq \|A_m - A_0\|_{L^\infty(D_R; \text{op})} \|\nabla v_1\|_{L^2(D_R)} \|\nabla v_2\|_{L^2(D_R)} \\
&\quad + k^2 \|n_m - n_0\|_{L^\infty(D_R; \mathbb{R})} \|v_1\|_{L^2(D_R)} \|v_2\|_{L^2(D_R)} \\
&\leq 2d_\infty((A_m, n_m, f_m), (A_0, n_0, f_0)) \|v_1\|_{H_k^1(D_R)} \|v_2\|_{H_k^1(D_R)},
\end{aligned}$$

Hence if  $(A_m, n_m, f_m) \rightarrow (A_0, n_0, f_0)$  in  $\mathcal{C}$ , then  $\mathcal{A}((A_m, n_m, f_m)) \rightarrow \mathcal{A}((A_0, n_0, f_0))$  in  $B(X, Y^*)$ . We also have

$$\left| \left[ \mathcal{L}((A_m, n_m, f_m)) - \mathcal{L}((A_0, n_0, f_0)) \right](v_2) \right| = \left| \int_{D_R} (f_m - f_0) \bar{v}_2 d\lambda \right| \leq \|f_m - f_0\|_{L^2(D_R)} \frac{\|v_2\|_{H_k^1(D_R)}}{k}.$$

Hence if  $(A_m, n_m, f_m) \rightarrow (A_0, n_0, f_0)$  in  $\mathcal{C}$ , then  $\mathcal{L}((A_m, n_m, f_m)) \rightarrow \mathcal{L}((A_0, n_0, f_0))$  in  $Y^*$ .  $\square$

**Definition 3.52** (The solution operator  $\mathcal{S}$ ). *Define the operator  $\mathcal{S} : \mathcal{C} \rightarrow H_{0,D}^1(D_R)$  by letting  $\mathcal{S}(A_0, n_0, f_0) \in H_{0,D}^1(D_R)$  be the solution of the Helmholtz EDP (Problem 3.45).*

**Theorem 3.53** ( $\mathcal{S}$  is well defined). *For  $(A_0, n_0, f_0) \in \mathcal{C}$  the solution  $\mathcal{S}((A_0, n_0, f_0))$  of the Helmholtz EDP (Problem 3.45) exists, is unique, and depends continuously on  $f_0$ .*

*Proof of Theorem 3.53.* Since  $\Re(-\langle T_R \gamma v, \gamma v \rangle_{\Gamma_R}) \geq 0$  for all  $v \in H_{0,D}^1(D_R)$  (see, e.g. [158, Theorem 2.6.4]),  $a_{A_0, n_0}$  satisfies a Gårding inequality. Since the inclusion  $H_{0,D}^1(D_R) \hookrightarrow L^2(D_R)$  is compact, Fredholm theory shows that uniqueness implies well-posedness (see, e.g. [146, Theorem 2.34]). Since  $A$  is Lipschitz and  $n$  is  $L^\infty$ , uniqueness follows from the unique continuation results in [124, 89]; see [99, Section 2] for these results specifically applied to Helmholtz problems.  $\square$

**Lemma 3.54** (Continuity of solution operator for Helmholtz stochastic EDP). *For the Helmholtz stochastic EDP, the solution operator  $\mathcal{S} : \mathcal{C} \rightarrow H_{0,D}^1(D_R)$  is continuous.*

*Sketch Proof of Lemma 3.54.* Let  $(A_0, n_0, f_0), (A_1, n_1, f_1) \in \mathcal{C}$ , such that  $\mathcal{S}((A_0, n_0, f_0)) = u_0$  and  $\mathcal{S}((A_1, n_1, f_1)) = u_1$ . Then for any  $v \in H_{0,D}^1(D_R)$  we have, for  $j = 0, 1$ ,

$$\left[ \left[ \mathcal{A}((A_j, n_j, f_j)) \right](u_j) \right](v) = \left[ \mathcal{L}((A_j, n_j, f_j)) \right](v).$$

Continuity of  $\mathcal{S}$  then follows from:

1. Deriving the Helmholtz equation with coefficients  $A_0$  and  $n_0$  satisfied by  $u_d := u_0 - u_1$ .
2. Recalling that the well-posedness result of Theorem 3.53 holds when  $f_0 \in L_R^2(D_R)$  is replaced by a right-hand side in  $(H_{0,D}^1(D_R))^*$ ; see, e.g., [146, Theorem 2.34].

3. Applying the result in Point 2 to obtain a bound  $\|u_d\|_{H_k^1(D_R)} \leq C(A_0, n_0) \|F\|_{(H_{0,D}^1(D_R))^*}$ .
4. Showing  $\|F\|_{(H_{0,D}^1(D_R))^*}$  depends on each of  $\|\nabla u_1\|_{L^2(D_R)}$ ,  $\|u_1\|_{L^2(D_R)}$ ,  $\|A_1 - A_0\|_{L^\infty(D_R; \text{op})}$ ,  $\|n_1 - n_0\|_{L^\infty(D_R; \mathbb{R})}$ , and  $\|f_0 - f_1\|_{L^2(D)}$ .
5. Eliminating the dependence on  $u_1$  by writing  $u_1 = u_0 - u_d$  and moving terms in  $u_d$  to the left-hand side, to obtain a bound on  $u_d$  of the form

$$\begin{aligned} & \|\nabla u_d\|_{L^2(D_R)} + k\|u_d\|_{L^2(D_R)} \\ & \leq \tilde{C}\left(u_0, A_0, n_0, \|A_1 - A_0\|_{L^\infty(D_R; \text{op})}, \|n_1 - n_0\|_{L^\infty(D_R; \mathbb{R})}, \|f_0 - f_1\|_{L^2(D_R)}\right). \end{aligned}$$

6. Concluding that  $u_d \rightarrow 0$  in  $H_{0,D}^1(D_R)$  as  $(A_1, n_1, f_1) \rightarrow (A_0, n_0, f_0)$  in  $\mathcal{C}$ .

□

**Lemma 3.55** (Condition U for the Helmholtz stochastic EDP).

*The Helmholtz stochastic EDP satisfies Condition U.*

*Proof of Lemma 3.55.* This condition holds immediately from Theorem 3.53. □

To prove that Condition B holds for the Helmholtz stochastic EDP, we first state the deterministic analogue of Theorem 3.10. This is essentially a restatement of Theorem 2.6, but on the set  $D_R$ , rather than on  $D_+$ . Theorem 3.56 uses the notation of Definition 2.3.

**Theorem 3.56** (Well-posedness of the Helmholtz EDP [105, Theorem 2.5]).

*Let  $(A_0, n_0, f_0) \in \mathcal{C}$  and suppose  $A_0 \in \text{NT}_{\text{mat}, D_R}(\tau_1)$  and  $n_0 \in \text{NT}_{\text{scal}, D_R}(\tau_2)$ . Then the solution of the Helmholtz EDP (Problem 3.45) exists and is unique. Furthermore, given  $k_0 > 0$ , for all  $k \geq k_0$ , the solution  $u_0$  of the Helmholtz EDP satisfies the bound*

$$\tau_1 \|\nabla u_0\|_{L^2(D_R)}^2 + \tau_2 k^2 \|u_0\|_{L^2(D_R)}^2 \leq C_1 \|f_0\|_{L^2(D_R)}^2, \text{ where } C_1 := 4 \left[ \frac{R^2}{\tau_1} + \frac{1}{\tau_2} \left( R + \frac{d-1}{2k_0} \right)^2 \right]. \quad (3.36)$$

We can now prove Condition B holds for the Helmholtz stochastic EDP.

**Lemma 3.57** (Condition B for Helmholtz stochastic EDP). *If Conditions 3.6 and 3.8 hold, then Condition B holds for the Helmholtz stochastic EDP.*

*Proof of Lemma 3.57.* As Condition 3.8 holds, the conditions of Theorem 3.56 hold for  $\mathbb{P}$ -almost every  $\omega \in \Omega$  (with  $A_0 = A(\omega)$ ,  $n_0 = n(\omega)$ ,  $\tau_1 = \mu_1(\omega)$ , and  $\tau_2 = \mu_2(\omega)$ ). Hence, by Theorem 3.56 the bound (3.24) holds for all  $k \geq k_0$ , with  $X = H_{0,D}^1(D_R)$ ,  $m = 1$ ,

$$C_1(\omega) = \frac{4}{\min\{\mu_1(\omega), \mu_2(\omega)\}} \left[ \frac{R^2}{\mu_1(\omega)} + \frac{1}{\mu_2(\omega)} \left( R + \frac{d-1}{2k_0} \right)^2 \right],$$

and  $f_1 = \|f(\omega)\|_{L^2(D_R)}^2$ . It now remains to show that  $C_1 \|f\|_{L^2(D_R)}^2 \in L^1(\Omega)$ . We first show that  $C_1 \|f\|_{L^2(D_R)}^2$  is measurable and then show that it lies in  $L^1(\Omega)$ . To show measurability, we rewrite  $C_1(\omega)$  as

$$C_1(\omega) = \max \left\{ \frac{2R^2}{\mu_1^2(\omega)} + \frac{2}{\mu_1(\omega)\mu_2(\omega)} \left( R + \frac{d-1}{2k_0} \right)^2, \frac{2R^2}{\mu_1(\omega)\mu_2(\omega)} + \frac{2}{\mu_2^2(\omega)} \left( R + \frac{d-1}{2k_0} \right)^2 \right\}.$$

The functions  $\mu_1^{-1}$  and  $\mu_2^{-1}$  are measurable by assumption; to conclude  $C_1$  is measurable we use the facts (see e.g. [110, Theorems 19.C, 20.A]): (i) the square of a measurable function is measurable, and (ii) the product, sum, and maximum of two measurable functions are measurable. Under Condition 3.6, the function  $f$  lies in the Bochner space  $L^2(\Omega; L^2(D_R))$ . Therefore,  $f$  is strongly measurable and hence  $f$  is measurable by Theorem B.18. The map  $f \mapsto \|f\|_{L^2(D_R)}^2$  is clearly continuous, and therefore  $f_1$  is measurable by Lemma B.4. As the product of two measurable functions is measurable, it follows that  $C_1 \|f\|_{L^2(D_R)}^2$  is measurable.

We now show that  $C_1 \|f\|_{L^2(D_R)}^2 \in L^1(\Omega)$ . The assumptions  $1/\mu_1, 1/\mu_2 \in L^2(\Omega)$  and the Cauchy-Schwarz inequality imply  $1/(\mu_1\mu_2) \in L^1(\Omega)$ . Therefore the maps,

$$\omega \mapsto \frac{2R^2}{\mu_1^2(\omega)} + \frac{2}{\mu_1(\omega)\mu_2(\omega)} \left( R + \frac{d-1}{2k_0} \right)^2 \quad \text{and} \quad \omega \mapsto \frac{2R^2}{\mu_1(\omega)\mu_2(\omega)} + \frac{2}{\mu_2^2(\omega)} \left( R + \frac{d-1}{2k_0} \right)^2$$

are in  $L^1(\Omega)$ . Since the maximum of two functions in  $L^1(\Omega)$  is also in  $L^1(\Omega)$ , it follows that  $C_1 \in L^1(\Omega)$ . Condition 3.6 implies that  $\|f\|_{L^2(D_R)}^2 \in L^1(\Omega)$ .

To conclude  $C_1 \|f\|_{L^2(D_R)}^2 \in L^1(\Omega)$ , observe that the only dependence of  $C_1$  on  $\omega$  is through  $\mu_1$  and  $\mu_2$ . As  $\mu_1$  and  $\mu_2$  are assumed independent of  $f$ , and measurable functions of independent random variables are independent [142, p.236] it follows that  $C_1$  and  $\|f\|_{L^2(D_R)}^2$  are independent, and therefore

$$\begin{aligned} \|C_1 \|f\|_{L^2(D_R)}^2 \|_{L^1(\Omega)} &= \int_{\Omega} C_1(\omega) \|f(\omega)\|_{L^2(D_R)}^2 d\mathbb{P}(\omega) \\ &= \left( \int_{\Omega} C_1(\omega) d\mathbb{P}(\omega) \right) \left( \int_{\Omega} \|f(\omega)\|_{L^2(D_R)}^2 d\mathbb{P}(\omega) \right) = \|C_1\|_{L^1(\Omega)} \| \|f\|_{L^2(D_R)}^2 \|_{L^1(\Omega)} < \infty. \end{aligned} \quad (3.37)$$

Therefore  $C_1 \|f\|_{L^2(D)}^2 \in L^1(\Omega)$  as required. We take the expectation (equivalently, the  $L^1$  norm) of (3.36) (with  $A_0 = A(\omega)$  etc.) and use (3.37) to obtain (3.9).  $\square$

**Remark 3.58** (The case when  $f$ ,  $\mu_1$ , and  $\mu_2$  are not independent). *Remark 3.11 shows that in the physically relevant case of scattering by a plane wave,  $f$ ,  $\mu_1$ , and  $\mu_2$  may not be independent. In this case, if we replace the requirements in Condition 3.8 that  $f \in L^2(\Omega; L^2(D))$  and  $1/\mu_1, 1/\mu_2 \in L^2(\Omega)$  with the stronger requirements  $f \in L^4(\Omega; L^2(D))$  and  $1/\mu_1, 1/\mu_2 \in L^4(\Omega)$ , then one can obtain the bound*

$$\|\nabla u\|_{L^2(\Omega; H_{0,D}^1(D_R))}^2 + k^2 \|u\|_{L^2(\Omega; H_{0,D}^1(D_R))}^2 \leq \|C_1\|_{L^2(\Omega)} \|f\|_{L^4(\Omega; L^2(D_R))}^2.$$

Indeed, instead of independence, we use the Cauchy–Schwarz inequality in (3.37) to conclude

$$\left\| C_1 \|f\|_{L^2(D_R)}^2 \right\|_{L^1(\Omega)} \leq \|C_1\|_{L^2(\Omega)} \left\| \|f\|_{L^2(D_R)}^2 \right\|_{L^2(\Omega)} = \|C_1\|_{L^2(\Omega)} \|f\|_{L^4(\Omega; L^2(D_R))}^2.$$

**Lemma 3.59** (Condition L2 for Helmholtz stochastic EDP). *If  $f \in L^2(\Omega; L^2(D_R))$  and  $A$  and  $n$  are strongly measurable, then Condition L2 holds for the Helmholtz stochastic EDP.*

*Proof of Lemma 3.59.* Since  $A, n$ , and  $f$  are strongly measurable, Conditions C1 and C2 hold by Lemma 3.50; i.e.,  $c$  is both measurable and  $\mathbb{P}$ -essentially separably valued. Furthermore, by Theorem B.18  $c$  is strongly measurable. By Lemma 3.51, Condition L1 holds, so the map  $\mathcal{L}$  is continuous. Hence, by Lemma B.21,  $\mathcal{L} \circ c$  is strongly measurable. We also have that  $\|(\mathcal{L} \circ c)(\omega)\|_{Y^*} = \|f(\omega)\|_{L^2(D_R)}/k$ , and thus  $\mathcal{L} \circ c \in L^2(\Omega; Y^*)$  since  $f \in L^2(\Omega; L^2(D_R))$ , i.e. Condition L2 holds.  $\square$

**Lemma 3.60** (Condition A2 for the Helmholtz stochastic EDP).

*If  $A \in L^\infty(\Omega; L^\infty(D_R; \text{SPD}))$ ,  $n \in L^\infty(\Omega; L^\infty(D_R; \mathbb{R}))$ , and  $f$  is strongly measurable, then Condition A2 holds for the Helmholtz stochastic EDP.*

*Proof of Lemma 3.60.* A near-identical argument to the argument at the beginning of the proof of Lemma 3.59 shows  $\mathcal{A} \circ c$  is strongly measurable. Recall that the Dirichlet-to-Neumann operator  $T_R$  is continuous from  $H^{1/2}(\Gamma_R)$  to  $H^{-1/2}(\Gamma_R)$ , see e.g. [158, Theorem 2.6.4]. Let  $v_1 \in X, v_2 \in Y$ , and observe that the Cauchy–Schwarz inequality and these properties of  $T_R$  imply that there exists  $C(k) > 0$  such that

$$\begin{aligned} \left| \left[ \left[ \mathcal{A}_{c(\omega)} \right] (v_1) \right] (v_2) \right| &= \left| \int_{D_R} \left( (A(\omega) \nabla v_1) \cdot \nabla \bar{v}_2 - k^2 n(\omega) v_1 \bar{v}_2 \right) d\lambda - \langle T_R v_1, v_2 \rangle_{\Gamma_R} \right| \\ &\leq \|A(\omega)\|_{L^\infty(D_R; \text{op})} \|\nabla v_1\|_{L^2(D_R)} \|\nabla v_2\|_{L^2(D_R)} \\ &\quad + k^2 \|n(\omega)\|_{L^\infty(D_R; \mathbb{R})} \|v_1\|_{L^2(D_R)} \|v_2\|_{L^2(D_R)} + C(k) \|\gamma v_1\|_{H^{1/2}(\Gamma_R)} \|\gamma v_2\|_{H^{1/2}(\Gamma_R)}. \end{aligned}$$

Since the trace operator  $\gamma$  is continuous from  $H^1(D_R)$  to  $H^{1/2}(\Gamma_R)$  (see, e.g. [146, Theorem 3.38]), there exists  $\tilde{C} > 0$  such that

$$\|(\mathcal{A} \circ c)(\omega)\|_{\mathbb{B}(X, Y^*)} \leq \tilde{C} \max \left\{ \|A(\omega)\|_{L^\infty(D_R; \text{op})}, \|n(\omega)\|_{L^\infty(D_R; \mathbb{R})}, C(k) \right\} \|v_1\|_{H_k^1(D_R)} \|v_2\|_{H_k^1(D_R)}.$$

and hence  $\mathcal{A} \circ c \in L^\infty(\Omega; \mathbb{B}(X, Y^*))$ .  $\square$

### 3.4.3 Proofs of Theorems 3.7 and 3.10

*Proof of Theorem 3.7.* We construct a solution of Problem 3.1 by letting  $u = \mathcal{S} \circ c$  (which is well-defined by Theorem 3.53), and observe that, by construction,  $[a(\omega)](u(\omega), v) = [L(\omega)](v)$  for all  $v \in H_{0,D}^1(D_R)$  almost surely. It follows that  $u$  is measurable by Condition 3.6 and Lemmas 3.54 and B.4, and so  $u$  solves Problem 3.1. We therefore proceed to apply the general theory.

Conditions A1 and L1 hold by Lemma 3.51; Condition A2 holds by Lemma 3.60; Condition L2 holds by Lemma 3.59; Conditions C1 and C2 hold by Lemma 3.50 and Condition 3.6; and

Condition U holds by Lemma 3.55. Hence we can apply Theorems 3.25 and 3.26 and Lemmas 3.24 and 3.28 to conclude the results.  $\square$

*Proof of Theorem 3.10.* All the conclusions of Theorem 3.7 hold, and we only need to show that if  $u$  solves Problem 3.1 then it also solves Problem 3.2. Condition B holds by Conditions 3.6 and 3.8 and Lemma 3.57. The result then follows from Theorem 3.23.  $\square$

## 3.5 SUMMARY AND FUTURE WORK

### 3.5.1 Summary

In this chapter we proved existence, uniqueness, and a  $k$ -independent a priori bound for the Helmholtz equation in random media under a  $k$ -independent nontrapping condition. In particular:

- In Section 3.1.1 we gave sufficient conditions for (i) the equivalence of three different formulations of the Helmholtz equation in random media, (ii) existence and uniqueness of a solution, and (iii) a  $k$ -independent a priori bound on the solution.
- In Section 3.2 in an abstract setting we gave three different formulations of linear spatial stochastic PDEs, along with abstract sufficient conditions for (i) the equivalence of these formulations, (ii) existence and uniqueness of a solution, and (iii) an a priori bound on the solution. These abstract conditions were then applied to the Helmholtz equation.

### 3.5.2 Future work

There are several possibilities for applying and extending the results in this chapter:

- Applying the abstract results in Section 3.2 to other linear spatial stochastic PDEs, especially those whose standard variational formulations are not coercive, similar to the Helmholtz equation. One such example would be the time-harmonic Maxwell's equations, which are closely related to the Helmholtz equation, see the discussion in Section 1.1.1.
- Extending the abstract results in Section 3.2 to cover coefficients that are not bounded almost surely; this extension should be straightforward, as the proofs in Section 3.3 are based on those in [96, 157], which treat coefficients that are not bounded almost surely.
- Attempting to extend the abstract results in Section 3.2 to nonlinear spatial stochastic PDEs. We expect this possibility would require substantially adapting the ideas and proofs in Sections 3.2 and 3.3 to the nonlinear case. However, if possible it would potentially allow one to prove well-posedness results for certain classes of nonlinear spatial stochastic PDEs.

# Nearby preconditioning for the Helmholtz equation

## 4.1 INTRODUCTION AND MOTIVATION FROM UQ

### 4.1.1 Motivation from uncertainty quantification for the Helmholtz equation

Consider the stochastic Helmholtz equation

$$\nabla \cdot (A(\omega, \mathbf{x}) \nabla u(\omega, \mathbf{x})) + k^2 n(\omega, \mathbf{x}) u(\omega, \mathbf{x}) = -f(\mathbf{x}), \quad \mathbf{x} \in D_+, \quad (4.1)$$

as defined in Chapter 3. If  $Q(u)$  is some quantity of interest of the solution, then the simplest way to approximate  $\mathbb{E}[Q(u)]$  is via a sampling-based method, i.e. using the approximation

$$\mathbb{E}[Q(u)] \approx \frac{1}{N} \sum_{l=1}^N Q(u(\omega^l)), \quad (4.2)$$

where the  $\omega^l$  are elements of the sample space  $\Omega$ . To calculate the right-hand side of (4.2), one must solve many deterministic Helmholtz problems, corresponding to different samples  $\omega^l$ , i.e. corresponding to different realisations of the coefficients  $A(\omega, \cdot)$  and  $n(\omega, \cdot)$ . Solving all these deterministic problems is a very computationally-intensive task because linear systems arising from discretisations of the Helmholtz equation are notoriously difficult to solve; see the discussion in Section 1.1.2 above. In particular, direct solvers involving a sparse LU decomposition of the linear system have a computational cost of the order  $\mathcal{O}(N^{3/2})$  in 2-d and  $\mathcal{O}(N^2)$  in 3-d (see [61, Section 1] and [61, Equation 3], respectively, for a particular regular grid).

However, if one already has access to the LU decomposition, then the cost of applying a direct solver using the LU decomposition is much cheaper;  $\mathcal{O}(N \log N)$  in 2-d [61, Section 1] and  $\mathcal{O}(N^{4/3})$  in 3-d [61, Equation 4]. In the context of Uncertainty Quantification for the Helmholtz equation this reduction in cost when one has access to an LU decomposition suggests the following question: When can the LU decomposition corresponding to a particular realisation of (4.1) be used as a preconditioner for other realisations of (4.1)?

This question of reusing preconditioners is more widely applicable than just for LU decompositions. For *any* preconditioner for the Helmholtz equation, one could ask when the preconditioner corresponding to one realisation of (4.1) can be re-used for other realisations. In this chapter, for simplicity, we restrict our attention to the case where the preconditioner is an exact LU decomposition.

One expects this reuse of the preconditioner to work well if the two realisations are ‘nearby’ in some sense. This idea of reusing preconditioners is the ‘nearby preconditioning’ strategy

proposed in this chapter. To analyse this ‘nearby preconditioning’ strategy rigorously, we first consider the following problem and question.

Let  $A^{(j)}, n^{(j)}$ ,  $j = 1, 2$  satisfy the properties of  $A$  and  $n$  in Problem 2.10 or Problem 2.12 (we will prove results for both problems), with  $u^{(j)}$  the corresponding solution and  $D_-, f$ , etc. as in Problem 2.10 or Problem 2.12. Let  $A^{(j)}$ ,  $j = 1, 2$ , be the Galerkin matrices for the corresponding  $h$ -finite-element discretisations (see (4.9) below for a precise definition of  $A^{(j)}$ ). We seek to answer:

Q1. How small must  $\|A^{(1)} - A^{(2)}\|$  and  $\|n^{(1)} - n^{(2)}\|$  be (in some norm to be defined, in terms of  $k$ -dependence) for GMRES applied to  $(A^{(1)})^{-1}A^{(2)}$  to converge in a  $k$ -independent number of iterations for arbitrarily large  $k$ ?

The rigorous answer of Q1 is contained in Theorems 4.11 and 4.12 below. However, an informal statement of the answer to Q1 is that if

$$k \|A^{(1)} - A^{(2)}\|_{L^\infty} \quad \text{and} \quad k \|n^{(1)} - n^{(2)}\|_{L^\infty} \quad \text{are both sufficiently small,} \quad (4.3)$$

then GMRES applied to  $(A^{(1)})^{-1}A^{(2)}$  in some weighted norm converges in a  $k$ -independent number of iterations (and a similar result for standard GMRES with (4.3) replaced by a slightly stronger condition).

### 4.1.2 Outline of the chapter

In Section 4.2 we state and discuss the main results of this chapter on the effectiveness of nearby preconditioning, and give their analogues on the PDE level. In Section 4.3 we describe numerical experiments investigating the sharpness of the nearby-preconditioning results in Section 4.2. In Section 4.4 we prove the results in Section 4.2. In Section 4.5 we extend the results in Section 4.2 to hold in weaker spatial norms, and we describe numerical experiments investigating the sharpness of these new results. In Section 4.6 we then apply the idea of nearby preconditioning to a Quasi-Monte-Carlo (QMC) method for the stochastic Helmholtz equation; in Section 4.6.2 we describe two algorithms for applying nearby preconditioning to QMC methods and in Section 4.6.3 we describe numerical experiments on the effectiveness of nearby preconditioning applied to QMC methods. In Section 4.7 we briefly review the related literature. Finally, in Section 4.8, we show how one can prove probabilistic results about the behaviour of nearby preconditioning, and we describe numerical experiments that investigate these probabilistic results.

## 4.2 STATEMENT OF THE MAIN RESULTS

### 4.2.1 Definition of variational problems and conditions used to prove main results

As this chapter concerns finite-element discretisations of the Helmholtz equation, we will work with the variational formulation of the Helmholtz equation. However, because the arguments we use do not directly rely on the boundary condition used to truncate the computational domain, we

will state our Helmholtz problems in sufficient generality to include both the EDP (Problem 2.10) and TEDP (Problem 2.12) above.

**Problem 4.1** (General variational Helmholtz problem). *Let  $D, A$ , and  $n$  be as in Problem 2.2. We say  $u \in H_{0,D}^1(D)$  satisfies the variational formulation of a general exterior Dirichlet problem with  $g_D = 0$  if*

$$a_G(u, v) = L_G(v) \text{ for all } v \in H_{0,D}^1(D), \quad (4.4)$$

where

$$a_G(w, v) := \int_D ((A \nabla w) \cdot \nabla \bar{v} - k^2 n w \bar{v}) - (T \gamma_I w, \gamma_I v)_{\Gamma_I}, \quad (4.5)$$

$T : H^{1/2}(\Gamma_I) \rightarrow H^{-1/2}(\Gamma_I)$  is a bounded linear map,  $(\cdot, \cdot)_{\Gamma_I}$  is the duality pairing on  $\Gamma_I$ , and  $L_G \in (H_{0,D}^1(D))^*$ .

**Remark 4.2** (Problem 4.1 is a generalisation of Problems 2.10 and 2.12). *With the exception of some overlap in notation, it is straightforward to see that appropriate choices of  $D, \Gamma_I, T$ , and  $L_G$  allow Problem 4.1 to be either Problem 2.10 or Problem 2.12. Taking  $D = D, \Gamma_I = \Gamma_R, T = T_R$  and  $L_G(v) = \int_D f \bar{v}$  (for  $f$  as in Problem 2.10) in Problem 4.1, we see Problem 4.1 becomes Problem 2.10. Additionally, taking  $D$  and  $\Gamma_I$  in Problem 4.1 to be the same as the  $D$  and  $\Gamma_I$  in Problem 2.12, taking  $T = ik$ , and  $L_G(v) = \int_D f \bar{v} + \int_{\Gamma_I} g_I \gamma_I \bar{v}$  (for  $f$  and  $g_I$  as in Problem 2.12), Problem 4.1 becomes Problem 2.12.*

**Remark 4.3** (Problem 4.1 allows for other boundary conditions). *The strength of the general formulation in Problem 4.1 is that it allows us to treat a wide variety of Helmholtz problems at once. Indeed, any Helmholtz problem that can be written in the form (4.4) and (4.5) and satisfies Conditions 4.8 and 4.9 below can be treated using the analysis in this chapter.*

For the remainder of this chapter, we let  $(V_{h,p})_{h>0}$  be the family of finite-element spaces.

**Assumption 4.4** (Properties of finite-element spaces). *We assume  $(V_{h,p})_{h>0}$  is a family of finite-dimensional subspaces of  $H_{0,D}^1(D)$ , whose union is dense in  $H_{0,D}^1(D)$ . Moreover, we assume  $V_{h,p}$  consists of nodal finite-element functions given by piecewise-polynomials on a quasi-uniform simplicial mesh  $\mathcal{T}_h$  with mesh-size  $h$  and fixed polynomial degree  $p$ .*

Note that the dimension  $N$  of  $V_{h,p}$  satisfies  $N \sim h^{-d}$ , with hidden constant dependent on  $p$ . (The assumption of quasi-uniformity can, in principle, be relaxed, see Remark 4.7 below.) As in Remark 2.17 above, we ignore any variational crimes resulting from this discretisation. We now define the finite-element approximation of Problem 4.1.

**Problem 4.5** (Finite-element approximation of Problem 4.1). *Find  $u_h \in V_{h,p}$  such that*

$$a_G(u_h, v_h) = L_G(v_h) \text{ for all } v_h \in V_{h,p}. \quad (4.6)$$

We say that  $u_h \in V_{h,p}$  is the finite-element approximation of  $u$  (the solution to Problem 4.1).

## 4.2.2 Definition of finite-element matrices, weighted norms, and weighted GMRES

### Finite-element matrices and weighted norms

We now define the matrices associated with our finite-element discretisation. First, let  $\{\phi_i, i = 1, \dots, N\}$  be a basis for  $V_{b,p}$  with each  $\phi_i$  *real-valued*. Let

$$(S_A)_{ij} := \int_D (A \nabla \phi_j) \cdot \nabla \phi_i, \quad (M_n)_{ij} := \int_D n \phi_i \phi_j, \quad \text{and} \quad (\mathbf{N})_{ij} := \int_{\Gamma_R} T(\gamma \phi_j) \gamma \phi_i \quad (4.7)$$

be the stiffness, domain-mass, and boundary-mass matrices, respectively. Note that both  $S_A$  and  $M_n$  are *real-valued*, but in general  $\mathbf{N}$  is *complex-valued* (because both the DtN operator  $T_R$  and the impedance operator  $ik$  are complex-valued). Let

$$A := S_A - k^2 M_n - \mathbf{N}, \quad (4.8)$$

and let  $u_b := \sum_j u_j \phi_j$ . Then (4.6) implies that the coefficient vector  $u = (u_i)_{i=1}^N \in \mathbb{C}^N$  satisfies

$$Au = f,$$

where  $(f)_i := L(\phi_i)$ . Similarly to above we let

$$A^{(j)} := S_{A^{(j)}} - k^2 M_{n^{(j)}} - \mathbf{N}. \quad (4.9)$$

Our main results about Q1 are Theorems 4.11 and 4.12 below. Theorem 4.12 gives results in the *Euclidean* norm on matrices, denoted by  $\|\cdot\|_2$  (induced by the Euclidean norm on vectors), whereas Theorem 4.11 gives results in the *weighted* norms  $\|\cdot\|_{D_k}$  and  $\|\cdot\|_{D_k^{-1}}$ . These weighted norms are induced by the corresponding vector norms

$$\|v\|_{D_k}^2 := (D_k v, v)_2 = \|v_b\|_{H_k^1(D)}^2 \quad \text{and} \quad \|v\|_{D_k^{-1}}^2 := (D_k^{-1} v, v)_2 \quad (4.10)$$

where  $D_k$  is given in terms of familiar finite-element stiffness- and mass-matrices by

$$D_k := S_I + k^2 M_1,$$

and  $v_b = \sum_i v_i \phi_i$ , where the  $\phi_i$  are the finite-element basis functions.

As described in Section 2.2.3, the PDE analysis of the Helmholtz equation naturally takes place in the norm  $\|\cdot\|_{H_k^1(D)}$ , and (4.10) shows that the norm  $\|\cdot\|_{D_k}$  is simply the norm on the finite-element space induced by  $\|\cdot\|_{H_k^1(D)}$ . The norms  $\|\cdot\|_{D_k}$  and  $\|\cdot\|_{D_k^{-1}}$  recently appeared in results about the convergence of domain-decomposition methods for the Helmholtz equation [102, 106], and a related norm appeared in similar results for the time-harmonic Maxwell equations [27].

The statement and proof our main results, Theorems 4.11 and 4.12 will require the following lemma.

**Lemma 4.6** (Norm equivalences of FE functions). *There exist  $m_{\pm} > 0$  and  $s_+ > 0$ , independent of  $h$  (but dependent on  $p$ ), such that*

$$m_- h^{d/2} \|v\|_2 \leq \|v_b\|_{L^2(D)} \leq m_+ h^{d/2} \|v\|_2, \quad (4.11)$$

and

$$\|\nabla v_b\|_{L^2(D)} \leq s_+ h^{d/2-1} \|v\|_2, \quad (4.12)$$

for all finite-element functions  $v_b = \sum_i v_i \phi_i \in V_{b,p}$ , with  $v = (v_i)_{i=1}^N \in \mathbb{C}^N$ .

The proof of Lemma 4.6 is on page 147 below.

Written in terms of the matrices  $M_1$  and  $S_I$  defined in (4.7), the bounds (4.11) and (4.12) are, respectively, the bounds

$$(M_1 v, v)_2 \sim h^d \|v\|_2^2 \quad \text{and} \quad (S_I v, v)_2 \lesssim h^{d-2} \|v\|_2^2.$$

**Remark 4.7** (Relaxing the assumption of quasi-uniformity). *We assume that  $\{\mathcal{T}_b\}_{b>0}$  is a quasi-uniform family of meshes so that the proof of Lemma 4.6 is straightforward. However, this assumption can almost certainly be relaxed. In [86] (on which the bulk of the arguments in this chapter are based) Gander, Graham, and Spence prove results both for quasi-uniform meshes and also for shape-regular meshes (see [86, Sections 3.4 and 4.1.2]). Given the results in [86] for shape-regular meshes are analagous to those they obtain for quasi-uniform meshes, we expect the results in this chapter can also be extended to shape-regular meshes. However, we note that [86] only contains bounds on preconditioned mass matrices (analagous to Lemma 4.19 below) but not preconditioned stiffness matrices (analagous to Lemma 4.20 below). Therefore it remains open to prove that our results in this chapter can be extended in their entirety to shape-regular meshes.*

## Weighted GMRES

We now give the set-up for weighted GMRES, first introduced in by Essai in [72]; we largely follow [102, Section 5]. Consider the abstract linear system  $Cx = d$  in  $\mathbb{C}^N$ , where  $C \in \mathbb{C}^{N \times N}$  is invertible. Let  $x^0$  be the initial guess, and define the initial residual  $r^0 := d - Cx^0$  and the standard Krylov spaces:

$$\mathcal{K}^m(C, r^0) := \text{span}\{C^j r^0 : j = 0, \dots, m-1\}.$$

Analogously to the definition of  $\|\cdot\|_{D_k}$  above, let  $(\cdot, \cdot)_D$  denote the inner product on  $\mathbb{C}^n$  induced by some Hermitian positive-definite matrix  $D$ , i.e.  $(v, w)_D := (Dv, w)_2$ , and let  $\|\cdot\|_D$  be the induced norm. For  $m \geq 1$ , define the  $m$ th GMRES iterate  $x^m$  to be the unique element of  $\mathcal{K}^m$  satisfying the minimal residual property:

$$\|r_m\|_D := \|d - Cx^m\|_D = \min_{x \in \mathcal{K}^m(C, r^0)} \|d - Cx\|_D.$$

Observe that when  $D = I$  this is the standard GMRES algorithm. We also note that in general, weighted GMRES requires the use of weighted Arnoldi process, also introduced by Essai in [72],

see also the alternative implementations of the weighted Arnoldi process in [109].

### 4.2.3 Main results

Theorems 4.11 and 4.12 are proved under the following two conditions, which are the minimal conditions needed for the proof of Theorems 4.11 and 4.12. Therefore, in particular, Condition 4.9 is a very weak condition on the finite-element space  $V_{h,p}$ , since it does not even require convergence (for fixed  $k$ ) as the mesh is refined.

**Condition 4.8** (Nontrapping bound on  $u^{(1)}$ ).  $A^{(1)}, n^{(1)}$ , and  $D$  are such that, given  $f \in L^2(D)$  the solution of Problem 4.1 with

$$L_G(v) = \int_D f \bar{v}, \quad (4.13)$$

$u^{(1)}$ , exists, is unique, and, given  $k_0 > 0$ ,  $u^{(1)}$  satisfies the bound

$$\|u^{(1)}\|_{H_k^1(D)} \leq C_{\text{bound}}^{(1)} \|f\|_{L^2(D)} \quad \text{for all } k \geq k_0, \quad (4.14)$$

where  $C_{\text{bound}}^{(1)}$  is independent of  $k$ , but dependent on  $A^{(1)}, n^{(1)}, D$ , and  $k_0$ .

**Condition 4.9** ( $k$ -independent accuracy of the FE solution for  $a^{(1)}(\cdot, \cdot)$ ).

1. Given  $k_0 > 0$ ,  $h$  and  $p$  are chosen to depend on  $k$  such that for all  $k \geq k_0$ , if  $f = n \sum_j \alpha_j \phi_j$  for some  $\alpha_j \in \mathbb{C}$  and  $n \in L^\infty(D_R)$  (i.e.  $f$  is an arbitrary element of  $V_{h,p}$  multiplied by  $n$ ), then

- The solution  $u_b$  of Problem 4.5 (with  $a_G = a_G^{(1)}$ , and  $L_G(v)$  given by (4.13)) exists and is unique, and
- The error bound

$$\|u - u_b\|_{H_k^1(D)} \leq C_{\text{FEM1}}^{(1)} \|f\|_{L^2(D)} \quad , \quad (4.15)$$

holds, where  $C_{\text{FEM1}}^{(1)}$  is independent of  $k$  and  $h$ , but dependent on  $A^{(1)}, n^{(1)}, D, k_0$ , and  $p$ .

2. Given  $k_0 > 0$ ,  $h$  and  $p$  are chosen to depend on  $k$  such that for all  $k \geq k_0$ , if  $L_G(v) = (A \nabla \tilde{f}, \nabla v)_{L^2(D)}$ , where  $A \in L^\infty(D; \mathbb{R}^{d \times d})$  and  $\tilde{f} := \sum_j \alpha_j \phi_j$  with  $\alpha_j \in \mathbb{C}$  (i.e.  $\tilde{f}$  is an arbitrary element of  $V_{h,p}$ ), then,

- The solution  $u_b$  of Problem 4.5 with  $a_G = a_G^{(1)}$  exists and is unique, and
- The error bound

$$\|u - u_b\|_{H_k^1(D)} \leq C_{\text{FEM2}}^{(1)} k \|L_G\|_{(H_k^1(D))^*} \quad , \quad (4.16)$$

holds, where  $C_{\text{FEM2}}^{(1)}$  is independent of  $k$  and  $h$ , but dependent on  $A^{(1)}, n^{(1)}, D, k_0$ , and  $p$ .

For details of when Conditions 4.8 and 4.9 are satisfied (for Problems 2.10 and 2.12), see Section 2.2 (for Condition 4.8) and Section 2.3.3 (for Condition 4.9 part 1). Conditions 4.8 and 4.9 can be informally stated as

- the obstacle  $D_-$  and the coefficients  $A^{(1)}$  and  $n^{(1)}$  are such that  $u^{(1)}$  exists, is unique, and the problem is *nontrapping* (in the sense described in Section 2.2.3 above), and
- the meshsize  $h$  and polynomial degree  $p$  in the finite-element method are chosen to depend on  $k$  to ensure that the finite-element approximation to the solution of the problem with coefficients  $A^{(1)}$  and  $n^{(1)}$  exists, is unique, and has bounded error in the  $H_k^1$ -norm as  $k \rightarrow \infty$ .

**Remark 4.10** ((4.16) has the same  $k$ -dependence as (4.15)). *Observe that the bound (4.16) has the same  $k$ -dependence as (4.15) despite the fact that a factor  $k$  appears on the right-hand side. If  $L_G(v) = \int_D f \bar{v}$ , then*

$$\begin{aligned} \|L_G\|_{(H_k^1(D))^*} &= \sup_{v \in H_{0,D}^1(D)} \frac{|L_G(v)|}{\|v\|_{H_k^1(D)}} \leq \sup_{v \in H_{0,D}^1(D)} \frac{\|f\|_{L^2(D)} \|v\|_{L^2(D)}}{\|v\|_{H_k^1(D)}} \\ &\lesssim \frac{1}{k} \sup_{v \in H_{0,D}^1(D)} \frac{\|f\|_{L^2(D)} \|v\|_{L^2(D)}}{\|v\|_{L^2(D)}} \\ &= \frac{\|f\|_{L^2(D)}}{k}. \end{aligned}$$

The factor  $k$  appears in (4.16) since we use the weighted norm  $\|\cdot\|_{H_k^1(D)}$  in the definition of  $\|\cdot\|_{(H_k^1(D))^*}$ , rather than the unweighted norm  $\|\cdot\|_{H^1(D)}$ .

**Theorem 4.11** (Answer to Q1:  $k$ -independent weighted GMRES iterations).

Let  $k_0 \geq 0$ ,  $k \geq k_0$ , and assume that  $D_-$ ,  $A^{(1)}$ , and  $n^{(1)}$  satisfy Condition 4.8,  $h$  and  $p$  satisfy Condition 4.9, and  $A^{(2)}$  and  $n^{(2)}$  are as in Problem 4.1. Then there exist constants  $C_1$  and  $C_2$  independent of  $h$  and  $k$  (but dependent on  $D_-$ ,  $A^{(1)}$ ,  $n^{(1)}$ ,  $p$ , and  $k_0$ ) such that if

$$C_1 k \left\| A^{(1)} - A^{(2)} \right\|_{L^\infty(D; \text{op})} + C_2 k \left\| n^{(1)} - n^{(2)} \right\|_{L^\infty(D; \mathbb{R})} \leq \frac{1}{2}, \quad (4.17)$$

then both weighted GMRES working in  $\|\cdot\|_{D_k}$  (and the associated inner product) applied to

$$(A^{(1)})^{-1} A^{(2)} u = f \quad (4.18)$$

and weighted GMRES working in  $\|\cdot\|_{(D_k)^{-1}}$  (and the associated inner product) applied to

$$A^{(2)} (A^{(1)})^{-1} v = f \quad (4.19)$$

converge in a  $k$ -independent number of iterations.

The constants  $C_1$  and  $C_2$  are given explicitly in (4.32) and (4.35) below. The proof of Theorem 4.11 is on page 160 below.

**Theorem 4.12** (Answer to Q1:  $k$ -independent (unweighted) GMRES iterations).

Let  $k_0 \geq 0$ ,  $k \geq k_0$ , and assume that  $D_-$ ,  $A^{(1)}$ , and  $n^{(1)}$  satisfy Condition 4.8,  $h$  and  $p$  satisfy Condition 4.9, and  $A^{(2)}$  and  $n^{(2)}$  are as in Problem 4.1. Let  $C_1$  and  $C_2$  be as in Theorem 4.11, and let

$s_+ > 0$  and  $m_{\pm} > 0$  be as in Lemma 4.6 (note that all these constants are independent of  $k$ ,  $h$ , and  $p$ ). Then if

$$C_1 \left( \frac{s_+}{m_-} \right) \frac{1}{h} \|A^{(1)} - A^{(2)}\|_{L^\infty(D; \text{op})} + C_2 \left( \frac{m_+}{m_-} \right) k \|n^{(1)} - n^{(2)}\|_{L^\infty(D; \mathbb{R})} \leq \frac{1}{2} \quad (4.20)$$

then standard GMRES (working in the Euclidean norm and inner product) applied to either of the equations (4.18) or (4.19) converges in a  $k$ -independent number of iterations.

The proof of Theorem 4.12 is on page 160 below.

Three notes regarding Theorems 4.11 and 4.12: (i) The  $L^\infty$  norms of  $A_1 - A_2$  and  $n_1 - n_2$  in Theorems 4.11 and 4.12 can be replaced by  $L^p$  norms with  $p < \infty$ , at the price of making the conditions (4.17) and (4.20) more restrictive; see Section 4.5 for more details. (ii) The  $\|\cdot\|_{L^\infty(D; \text{op})}$  norm on matrix-valued functions appearing on the right-hand sides of (4.28) and (4.29) is defined by (3.4). (iii) The factor  $1/2$  on the right-hand sides of (4.17) and (4.20) can be replaced by any number  $< 1$  and the result still holds, although the number of GMRES iterations may then be different—but is still independent of  $k$ .

**Remark 4.13.** When  $h \sim k^{-1}$ , the bounds (4.17) and (4.20) are identical in their  $k$ -dependence; however, when  $h \ll k^{-1}$  (as one needs to take to overcome the pollution effect, as discussed in Section 2.3.3) the bound (4.20) for standard GMRES is more pessimistic than the bound (4.17) for weighted GMRES.

**Remark 4.14** (Link to the results of [86]). A result analogous to the Euclidean-norm bounds in Theorem 4.18 was proved in [86] for the case that  $A^{(1)} = A^{(2)} = I$ ,  $n^{(2)} = 1$ , and  $n^{(1)} = 1 + i\varepsilon/k^2$ , with the ‘absorption parameter’ or ‘shift’  $\varepsilon$  satisfying  $0 < \varepsilon \lesssim k^2$ . (Recall from Remark 4.7 that the proof strategy used in this chapter is based on the strategy in [86].) The motivation for proving the results in [86] was that the so-called ‘shifted Laplacian preconditioning’ of the Helmholtz equation is based on preconditioning (with these choices of parameters)  $A^{(2)}$  with an approximation of  $A^{(1)}$ . Similar to Theorem 4.11, bounds on  $\|I - (A^{(1)})^{-1}A^{(2)}\|_2$  and  $\|I - A^{(2)}(A^{(1)})^{-1}\|_2$  then give upper bounds on how large the ‘shift’  $\varepsilon$  can be for GMRES for  $(A^{(1)})^{-1}A^{(2)}$  to converge in a  $k$ -independent number of iterations in the case when the action of  $(A^{(1)})^{-1}$  is computed exactly.

The main differences between [86] and this work are that: (i) [86] focused on the TEDP, not both the TEDP and the EDP, (ii) [86] focused on the particular case that  $D_-$  is star-shaped with respect to a ball, finding a  $k$ - and  $\varepsilon$ -explicit expression for  $C_{\text{bound}}^{(1)}$  in this case using Morawetz identities, whereas we assume the existence of  $C_{\text{bound}}^{(1)}$ , (iii) [86] required a bound on  $(A^{(1)})^{-1}M_n$ , analogous to the bounds in Lemma 4.19 along with one on  $(A^{(1)})^{-1}N$  (in the case that  $T_R$  is approximated by  $ik$ ), but not on  $(A^{(1)})^{-1}S_A$ , and (iv) [86] only proved bounds in the  $\|\cdot\|_2$  norm.

#### 4.2.4 PDE analogues to Theorems 4.11 and 4.12

Numerical experiments in Section 4.3 below indicate that the condition (4.17) is sharp, i.e., that the  $k$  in (4.17) cannot be replaced by  $k^\alpha$  for  $\alpha < 1$ . This indicated sharpness of (4.17) is also supported

by the PDE-result Theorem 4.15 below. Indeed, Theorem 4.15 shows that the condition

$$k \left\| A^{(1)} - A^{(2)} \right\|_{L^\infty(D; \text{op})} \quad \text{and} \quad k \left\| n^{(1)} - n^{(2)} \right\|_{L^\infty(D; \mathbb{R})} \quad \text{both sufficiently small} \quad (4.21)$$

is not only an answer to Q1 (about finite-element discretisations), but is also the natural answer to the analogue of Q1 at the level of PDEs, namely

Q2. How small must  $\left\| A^{(1)} - A^{(2)} \right\|_{L^\infty(D; \text{op})}$  and  $\left\| n^{(1)} - n^{(2)} \right\|_{L^\infty(D; \mathbb{R})}$  be (in terms of their  $k$ -dependence) for the relative error in approximating  $u^{(2)}$  by  $u^{(1)}$  to be bounded independently of  $k$  for arbitrarily-large  $k$ ?

Lemma 4.16 shows that the condition " $k \left\| n^{(1)} - n^{(2)} \right\|_{L^\infty(D; \mathbb{R})}$  sufficiently small" is the *provably-sharp* answer to Q2 when  $A^{(1)} = A^{(2)} = I$ .

To state these PDE results, we use the notation for  $a, b > 0$  that  $a \lesssim b$  when  $a \leq Cb$  for some  $C > 0$ , independent of  $k$ , and  $a \sim b$  if  $a \lesssim b$  and  $b \lesssim a$ .

**Theorem 4.15** (Answer to Q2 (the PDE analogue of Q1)). *Let  $k_0 > 0$  and  $k \geq k_0$ . Let  $D_-, A^{(1)}$ , and  $n^{(1)}$  satisfy Condition 4.8 applied to Problem 2.10 (so that the solution of Problem 2.10  $u^{(1)}$  exists, is unique, and satisfies a  $k$ -independent a priori bound). Let  $D_-, A^{(2)}$ , and  $n^{(2)}$  be such that  $u^{(2)}$  exists for any  $f \in L^2(D)$  such that  $\text{supp } f \subset B_R$ . Then, there exists  $C_3 > 0$ , independent of  $k$  and given explicitly in terms of  $D_-, A^{(1)}$ , and  $n^{(1)}$  in (4.62) below, such that*

$$\frac{\left\| u^{(1)} - u^{(2)} \right\|_{H_k^1(D)}}{\left\| u^{(2)} \right\|_{H_k^1(D)}} \leq C_3 k \max \left\{ \left\| A^{(1)} - A^{(2)} \right\|_{L^\infty(D; \text{op})}, \left\| n^{(1)} - n^{(2)} \right\|_{L^\infty(D; \mathbb{R})} \right\} \quad (4.22)$$

for all  $k \geq k_0$ .

The proof of Theorem 4.15 is on page 160 below.

**Lemma 4.16** (Sharpness of the bound (4.22) when  $A^{(1)} = A^{(2)} = I$ ). *There exist particular choices of  $f$ ,  $n^{(1)}$ , and  $n^{(2)}$  (with  $n^{(1)} \neq n^{(2)}$  both continuous) such that the corresponding solutions  $u^{(1)}$  and  $u^{(2)}$  of Problem 2.1 with  $A^{(1)} = A^{(2)} = I$  exist, are unique, and satisfy*

$$\frac{\left\| u^{(1)} - u^{(2)} \right\|_{H_k^1(D)}}{\left\| u^{(2)} \right\|_{H_k^1(D)}} \sim \frac{\left\| u^{(1)} - u^{(2)} \right\|_{L^2(D)}}{\left\| u^{(2)} \right\|_{L^2(D)}} \sim k \left\| n^{(1)} - n^{(2)} \right\|_{L^\infty(D; \mathbb{R})}. \quad (4.23)$$

The proof of Lemma 4.16 is on page 161 below.

**Remark 4.17** (Physical interpretation for  $k$ -dependence). *It is perhaps unsurprising that the condition (4.21) is a sufficient condition to answer both Q1 and Q2. Recall that  $1/k$  is proportional to the wavelength  $2\pi/k$  of the wave  $u$  (at least when  $A = I$  and  $n = 1$ ). As the wavelength is the natural length scale associated with the wave  $u$ , one expects perturbations of magnitude up to  $\mathcal{O}(1/k)$  to be 'unseen' by the PDE or numerical method. This is exactly what we see; perturbations of size (up to)  $1/k$  give bounded relative difference (in Q2) and bounded GMRES iterations for the nearby-preconditioned*

linear system (in Q1). Also, on a PDE level, perturbations of order  $1/k$  being ‘unseen’ by the PDE can also be seen in bounds proved for  $u$  where  $n = n_1 + \eta$ , with  $n_1$  nontrapping and  $\|\eta\|_{L^\infty(D;\mathbb{R})} \lesssim 1/k$ , see Remark 3.15 above.

### 4.3 NUMERICAL EXPERIMENTS VERIFYING AND INVESTIGATING THE SHARPNESS OF THEOREMS 4.11 AND 4.12

The numerical experiments in this section seek to verify Theorems 4.11 and 4.12 for Problem 2.12, and investigate their sharpness. More specifically, the experiments seek to verify if the condition (4.17) is:

1. sufficient, and
2. necessary

for *standard* GMRES applied to (4.18) to converge in a number of iterations that is independent of  $k$ .

Based on the PDE results Theorem 4.15 and Lemma 4.16 above, we expect that the condition (4.21) is a necessary and sufficient condition for standard GMRES applied to (4.18) to converge in a  $k$ -independent number of iterations, even though we can only prove this is a sufficient condition for *weighted* GMRES. We expect this because (4.21) is a sufficient condition for Q2, the PDE analogue of Q1. Indeed, this is exactly the behaviour we observe in numerical experiments; we see that if (4.21) holds, then standard GMRES applied to (4.18) converges in a  $k$ -independent number of iterations, and moreover, (4.21) may be sharp. We now describe our numerical experiments in more detail.

To verify this expected behaviour, we perform numerical experiments with the setup described in Appendix G with  $A^{(1)} = I$  and  $n^{(1)} = 1$ . We define  $f$  and  $g_I$  to correspond to a plane wave incident from the bottom left passing through a homogeneous medium given by coefficients  $A^{(1)}$  and  $n^{(1)}$ . We perform experiments for  $A$  and  $n$  separately, i.e., first we perform experiments with  $A^{(2)} = I$  and  $n^{(2)}$  varying, and then we perform experiments with  $A^{(2)}$  varying and  $n^{(2)} = 1$ . When we vary  $A^{(2)}$  we measure  $A^{(1)} - A^{(2)}$  in the  $\|\cdot\|_{L^\infty(D;\mathbb{R}^{d \times d})}$  norm, as this norm is easier to control than the  $\|\cdot\|_{L^\infty(D;\text{op})}$  norm. However, these two norms are equivalent on  $L^\infty(D;\mathbb{R}^{d \times d})$  (see the comment above (3.4)).

We define  $A^{(2)}$  and  $n^{(2)}$  to be piecewise constant (matrix-valued and real-valued respectively) on a  $10 \times 10$  square grid, with their values on each square chosen independently at random from a  $\text{Unif}(1 - \alpha, 1 + \alpha)$  distribution, with  $\alpha \in (0, 1)$  chosen as described below. For  $A^{(2)}$ , we impose the restriction that on each square  $A^{(2)}$  is positive-definite almost surely. We solve the linear systems (4.18) for  $k = 20, 40, 60, 80, 100$  using standard GMRES and record the number of GMRES iterations taken to achieve a (relative) tolerance of  $10^{-5}$  (relative to  $\|b\|_2$ ).

We perform experiments taking  $\alpha = 0.5 \times k^{-\beta}$  for  $\beta \in 0, 0.1, \dots, 0.9, 1$ . We expect that when  $\beta \neq 1$  the number of GMRES iterations required for convergence will increase as  $k$  increases,

whereas we expect that when  $\beta = 1$  the number of GMRES iterations required for convergence will remain bounded as  $k$  increases, even though this behaviour for  $\beta = 1$  has only been proved for  $\|A^{(1)} - A^{(2)}\|_{L^\infty(D; \text{op})}$  for weighted GMRES (compare the restrictions on  $\|A^{(1)} - A^{(2)}\|_{L^\infty(D; \text{op})}$  in Theorems 4.11 and 4.12).

In Figures 4.1–4.6, when  $\beta \in \{0, \dots, 0.3\}$  (for  $\|A^{(1)} - A^{(2)}\|_{L^\infty(D; \mathbb{R}^{d \times d})}$ ) and  $\beta \in \{0, \dots, 0.5\}$  (for  $\|n^{(1)} - n^{(2)}\|_{L^\infty(D; \mathbb{R})}$ ) we see growth in the maximum number of GMRES iterations needed (over all realisations) to achieve convergence, otherwise we see that the number of GMRES iterations is bounded as  $k$  increases. This behaviour is better than expected; as the number of GMRES iterations is apparently bounded for a range of  $\beta < 1$ . However, we note that this behaviour could be (i) because we are in a pre-asymptotic regime, and the number of GMRES iterations would grow if we increased  $k$  further, or (ii) the particular structure of  $n^{(2)}$  (being piecewise constant, with the pieces independently, randomly chosen) could result in some kind of ‘averaging’ behaviour, meaning the preconditioner is better than would otherwise be expected. However, we do not investigate these issues further in this thesis.

## 4.4 PROOFS OF THEOREMS 4.11, 4.12, AND 4.15 AND LEMMA 4.16

### 4.4.1 Proof of the main ingredient of the proofs of Theorems 4.11 and 4.12

As the first step towards proving Theorems 4.11 and 4.12, we prove Lemma 4.6, concerning norm equivalences of finite-element functions.

*Proof of Lemma 4.6.* We first show (4.11) by direct computation, before concluding (4.12) from (4.11) and a standard inverse inequality. Throughout this proof, when we use  $\sim$  the hidden constants are independent of  $\tau \in \mathcal{T}_h$ , the mesh size  $h$ , and  $v_b \in V_{h,p}$ , but may depend on  $p$ .

For any  $v_b \in V_{h,p}$  we have (letting  $\mathbf{n}_j$  denote a node of  $\mathcal{T}_h$ )

$$\|v_b\|_{L^2(D)}^2 = \sum_{\tau \in \mathcal{T}_h} \int_{\tau} |v_b|^2 \quad (4.24)$$

$$\sim \sum_{\tau \in \mathcal{T}_h} |\tau| \sum_{\mathbf{n}_j \in \tau} |v_b(\mathbf{n}_j)|^2 \quad (4.25)$$

$$\begin{aligned} &\sim h^d \sum_{\tau \in \mathcal{T}_h} \sum_{\mathbf{n}_j \in \tau} |v_b(\mathbf{n}_j)|^2, \text{ by quasi-uniformity,} \\ &\sim h^2 \|u\|_2, \end{aligned} \quad (4.26)$$

i.e., (4.11). The expression (4.25) follows from (4.24) because the terms defined on  $\tau$  are equivalent norms on  $\tau$  of functions in  $V_{h,p}$ .

To show (4.12) we recall the standard inverse inequality (see, e.g., [29, Theorem 4.5.11 and Remark 4.5.20])

$$\|v_b\|_{H^1(D)} \lesssim h^{-1} \|v_b\|_{L^2(D)}. \quad (4.27)$$

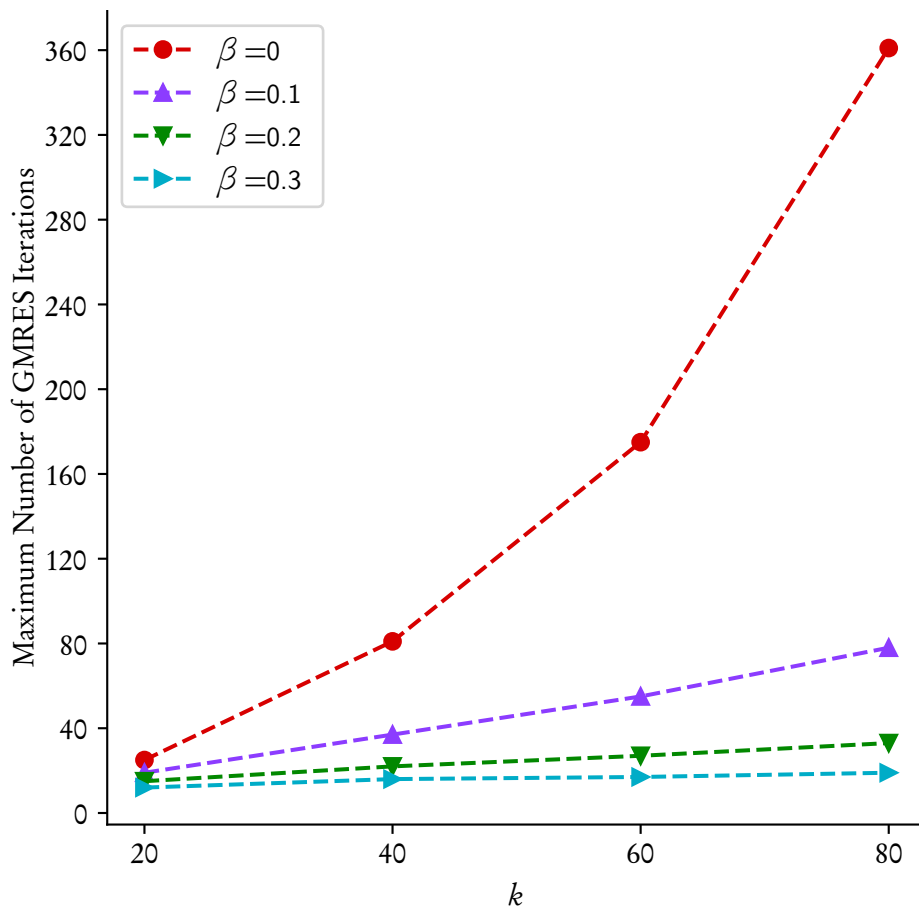


Figure 4.1: Maximum GMRES iteration counts for solving systems with matrix  $(A^{(1)})^{-1}A^{(2)}$ , where  $n^{(1)} = n^{(2)} = 1$  and  $\|A^{(1)} - A^{(2)}\|_{L^\infty(D; \mathbb{R}^{d \times d})} = 0.5 \times k^{-\beta}$  for  $\beta = 0, 0.1, 0.2, 0.3$ .

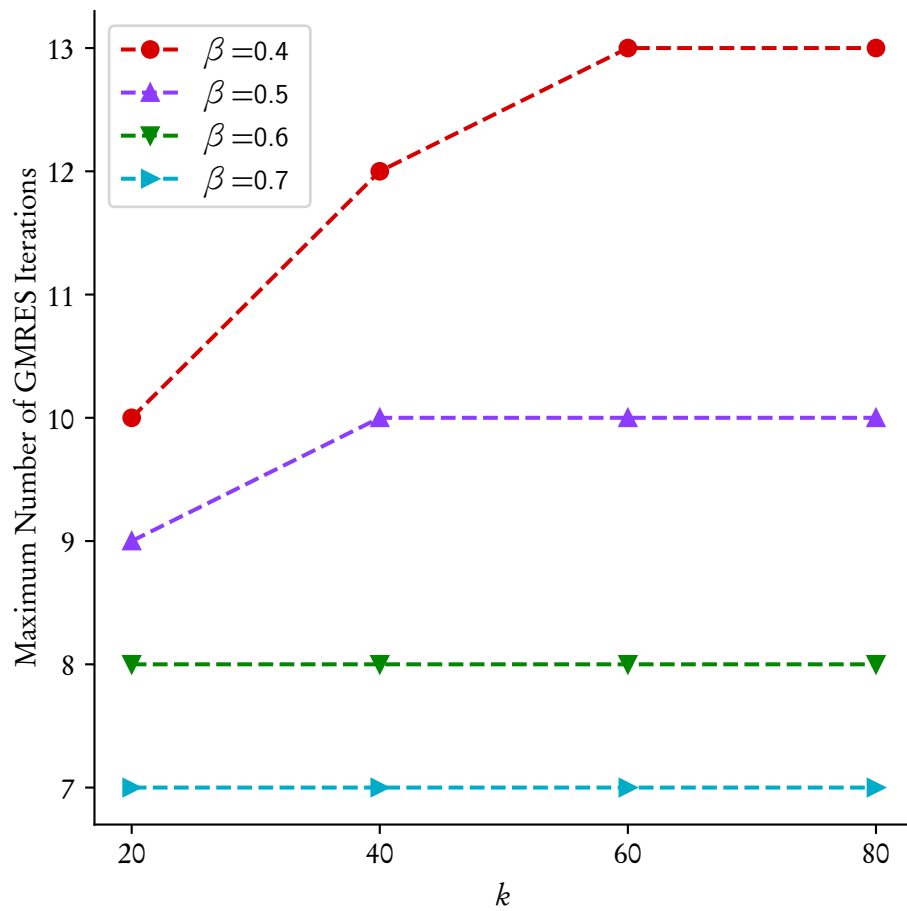


Figure 4.2: Maximum GMRES iteration counts for solving systems with matrix  $(A^{(1)})^{-1}A^{(2)}$ , where  $n^{(1)} = n^{(2)} = 1$  and  $\|A^{(1)} - A^{(2)}\|_{L^\infty(D; \mathbb{R}^{d \times d})} = 0.5 \times k^{-\beta}$  for  $\beta = 0.4, 0.5, 0.6, 0.7$ .

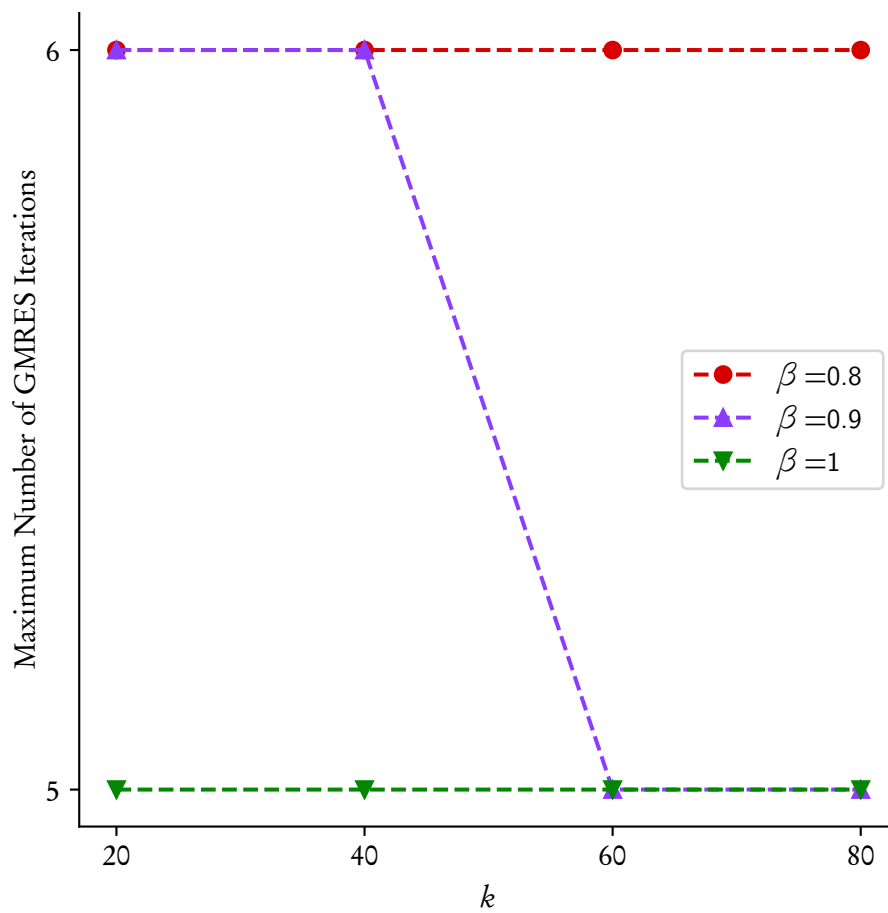


Figure 4.3: Maximum GMRES iteration counts for solving systems with matrix  $(A^{(1)})^{-1}A^{(2)}$ , where  $n^{(1)} = n^{(2)} = 1$  and  $\|A^{(1)} - A^{(2)}\|_{L^\infty(D; \mathbb{R}^{d \times d})} = 0.5 \times k^{-\beta}$  for  $\beta = 0.8, 0.9, 1$ .

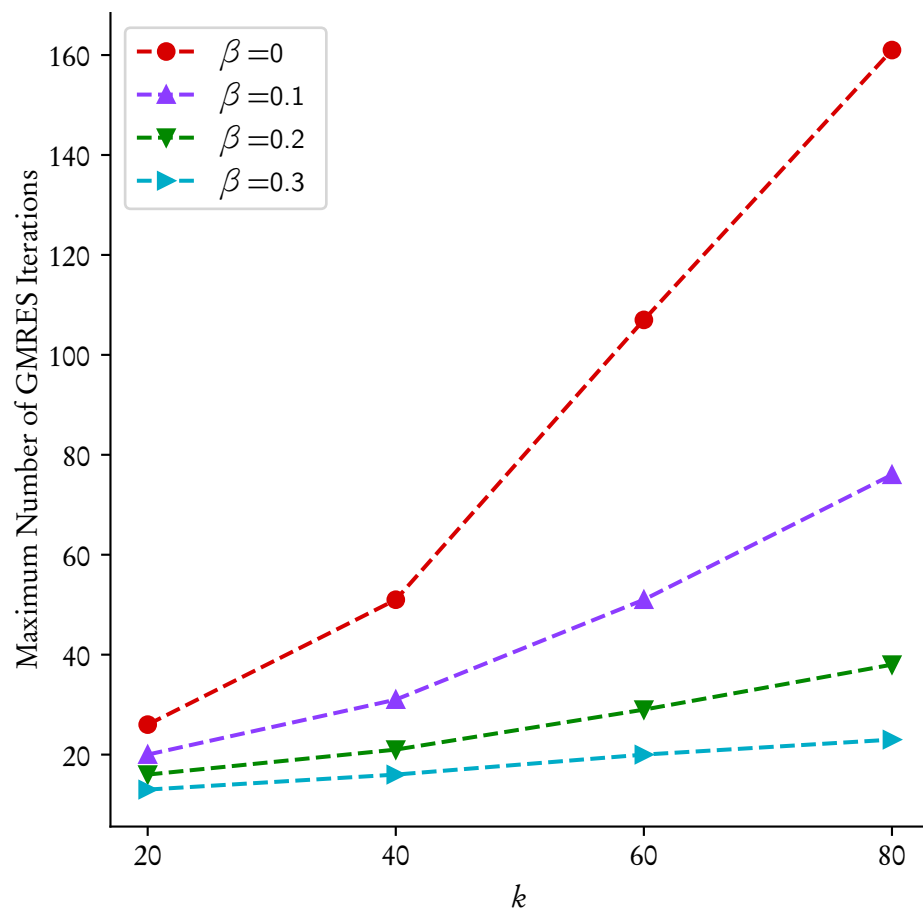


Figure 4.4: Maximum GMRES iteration counts for solving systems with matrix  $(A^{(1)})^{-1}A^{(2)}$ , where  $A^{(1)} = A^{(2)} = 1$  and  $\|n^{(1)} - n^{(2)}\|_{L^\infty(D; \mathbb{R})} = 0.5 \times k^{-\beta}$  for  $\beta = 0, 0.1, 0.2, 0.3$ .

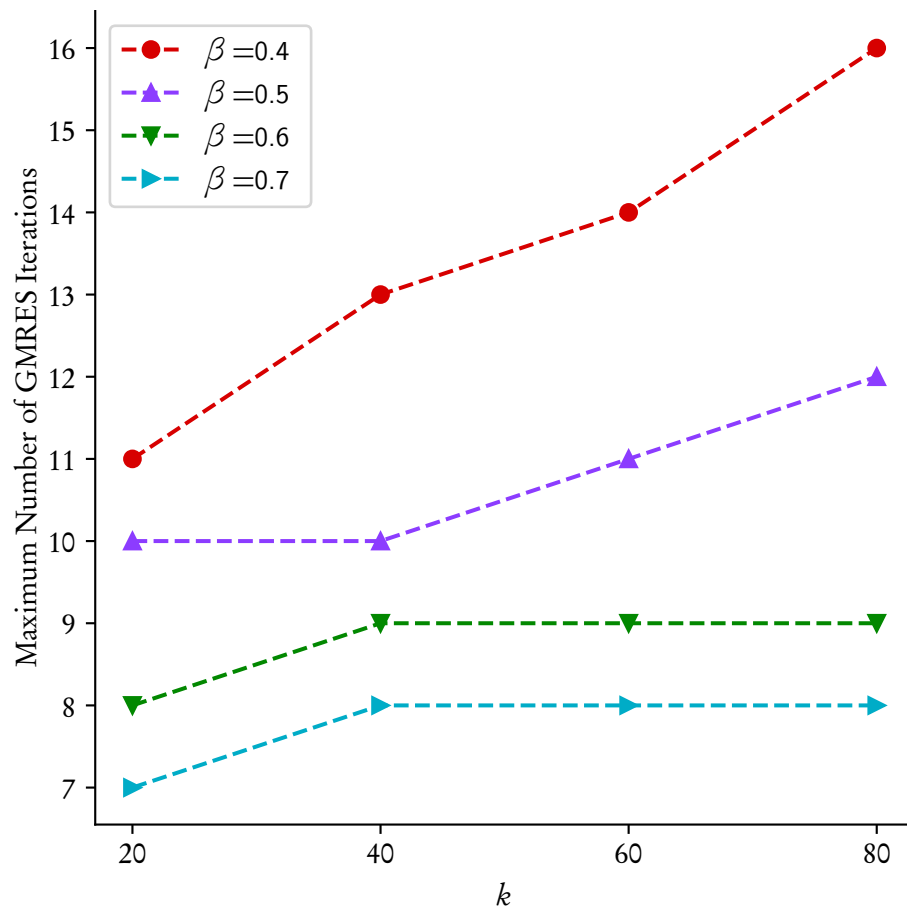


Figure 4.5: Maximum GMRES iteration counts for solving systems with matrix  $(A^{(1)})^{-1}A^{(2)}$ , where  $A^{(1)} = A^{(2)} = 1$  and  $\|n^{(1)} - n^{(2)}\|_{L^\infty(D; \mathbb{R})} = 0.5 \times k^{-\beta}$  for  $\beta = 0.4, 0.5, 0.6, 0.7$ .

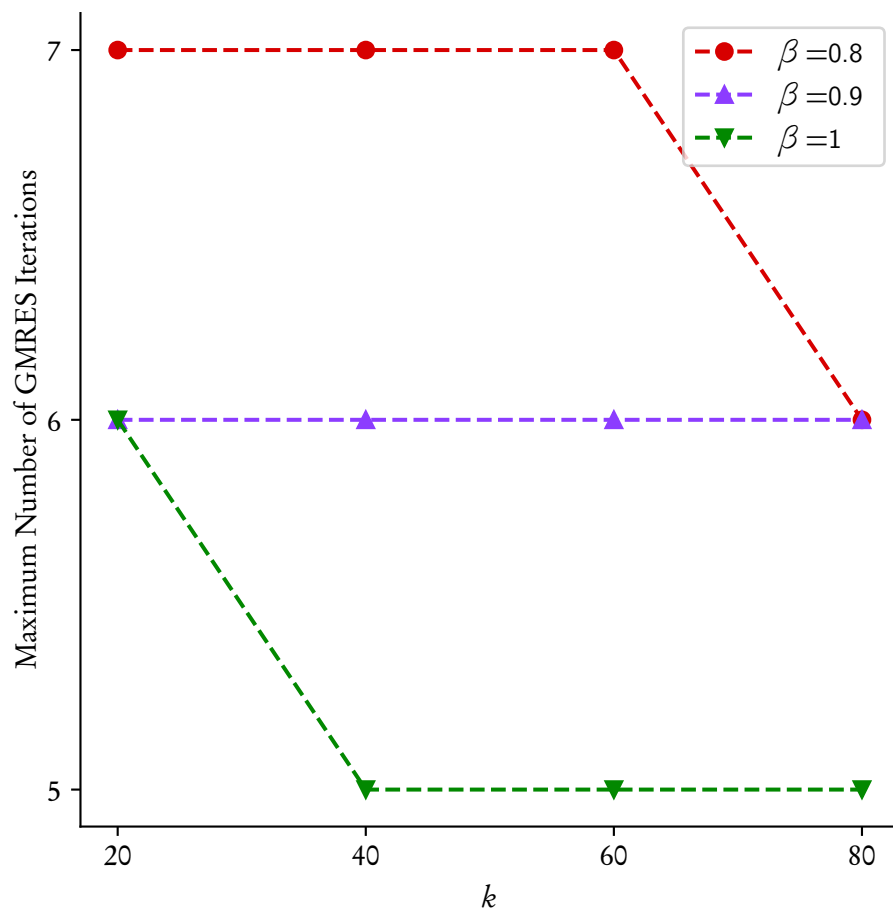


Figure 4.6: Maximum GMRES iteration counts for solving systems with matrix  $(A^{(1)})^{-1}A^{(2)}$ , where  $A^{(1)} = A^{(2)} = 1$  and  $\|n^{(1)} - n^{(2)}\|_{L^\infty(D; \mathbb{R})} = 0.5 \times k^{-\beta}$  for  $\beta = 0.8, 0.9, 1$ .

By combining (4.27) and the right-hand side of (4.11), we obtain (4.12).  $\square$

The main part of the proofs of Theorems 4.11 and 4.12 is the following theorem.

**Theorem 4.18** (Main ingredient of the answer to Q1). *Let  $k_0 \geq 0$ ,  $k \geq k_0$ , and assume  $D_-$ ,  $A^{(1)}$ , and  $n^{(1)}$  satisfy Condition 4.8, and assume that  $h$  and  $p$  satisfy Condition 4.9. Let the  $k$ - and  $h$ -independent constants  $m_{\pm}$  and  $s_{\pm}$  be given as in Lemma 4.6. Then there exist  $C_1, C_2 > 0$ , independent of  $h$  and  $k$  (but dependent on  $D_-, A^{(1)}, n^{(1)}, p$ , and  $k_0$ ) such that*

$$\begin{aligned} \max \left\{ \left\| I - (A^{(1)})^{-1} A^{(2)} \right\|_{D_k}, \left\| I - A^{(2)} (A^{(1)})^{-1} \right\|_{(D_k)^{-1}} \right\} \\ \leq C_1 k \left\| A^{(1)} - A^{(2)} \right\|_{L^\infty(D; \text{op})} + C_2 k \left\| n^{(1)} - n^{(2)} \right\|_{L^\infty(D; \mathbb{R})} \end{aligned} \quad (4.28)$$

and

$$\begin{aligned} \max \left\{ \left\| I - (A^{(1)})^{-1} A^{(2)} \right\|_2, \left\| I - A^{(2)} (A^{(1)})^{-1} \right\|_2 \right\} \\ \leq C_1 \left( \frac{s_+}{m_-} \right) \frac{1}{h} \left\| A^{(1)} - A^{(2)} \right\|_{L^\infty(D; \text{op})} + C_2 \left( \frac{m_+}{m_-} \right) k \left\| n^{(1)} - n^{(2)} \right\|_{L^\infty(D; \mathbb{R})}. \end{aligned} \quad (4.29)$$

The proof of Theorem 4.18 is given after the following two lemmas, that are the heart of the proof of Theorem 4.18.

**Lemma 4.19** (Bounds on  $(A^{(1)})^{-1} M_n$ ). *Assume that Condition 4.8 holds, and assume that part (i) of Condition 4.9 holds. Then, for  $n \in L^\infty(D; \mathbb{R})$ ,*

$$\max \left\{ \left\| (A^{(1)})^{-1} M_n \right\|_{D_k}, \left\| M_n (A^{(1)})^{-1} \right\|_{(D_k)^{-1}} \right\} \leq C_2 \frac{\|n\|_{L^\infty(D; \mathbb{R})}}{k} \quad (4.30)$$

and

$$\max \left\{ \left\| (A^{(1)})^{-1} M_n \right\|_2, \left\| M_n (A^{(1)})^{-1} \right\|_2 \right\} \leq C_2 \left( \frac{m_+}{m_-} \right) \frac{\|n\|_{L^\infty(D; \mathbb{R})}}{k}, \quad (4.31)$$

where

$$C_2 := C_{\text{FEM1}}^{(1)} + C_{\text{bound}}^{(1)}. \quad (4.32)$$

The proof of Lemma 4.19 is on page 156 below.

**Lemma 4.20** (Bounds on  $(A^{(1)})^{-1} S_A$ ). *Assume that Condition 4.8 holds, and assume that part (ii) of Condition 4.9 holds. Then, for  $A \in L^\infty(D; \mathbb{R}^{d \times d})$ ,*

$$\max \left\{ \left\| (A^{(1)})^{-1} S_A \right\|_{(D_k)^{-1}}, \left\| S_A (A^{(1)})^{-1} \right\|_{D_k} \right\} \leq C_1 k \|A\|_{L^\infty(D; \text{op})} \quad (4.33)$$

and

$$\max \left\{ \left\| (A^{(1)})^{-1} S_A \right\|_2, \left\| S_A (A^{(1)})^{-1} \right\|_2 \right\} \leq C_1 \left( \frac{s_+}{m_-} \right) \frac{1}{h} \|A\|_{L^\infty(D; \text{op})}, \quad (4.34)$$

where

$$C_1 := \left[ C_{\text{FEM2}}^{(1)} + \frac{1}{\min\{A_{\min}^{(1)}, n_{\min}^{(1)}\}} \left( \frac{1}{k_0} + 2C_{\text{bound}}^{(1)} n_{\max}^{(1)} \right) \right]. \quad (4.35)$$

The proof of Lemma 4.20 is on page 158 below.

*Proof of Theorem 4.18 using Lemmas 4.19 and 4.20.* Using the definition of the matrices  $A^{(j)}$ ,  $S_A$ , and  $M_n$  in (4.9) and (4.7), we have

$$\begin{aligned} I - (A^{(1)})^{-1}A^{(2)} &= (A^{(1)})^{-1}(A^{(1)} - A^{(2)}) = (A^{(1)})^{-1}(S_{A^{(1)}} - S_{A^{(2)}} - k^2(M_{n^{(1)}} - M_{n^{(2)}})) \\ &= (A^{(1)})^{-1}(S_{A^{(1)}-A^{(2)}} - k^2M_{n^{(1)}-n^{(2)}}), \end{aligned} \quad (4.36)$$

and similarly

$$I - A^{(2)}(A^{(1)})^{-1} = (S_{A^{(1)}-A^{(2)}} - k^2M_{n^{(1)}-n^{(2)}})(A^{(1)})^{-1}. \quad (4.37)$$

The bounds in (4.28) on  $\|I - (A^{(1)})^{-1}A^{(2)}\|_{D_k}$  and  $\|I - A^{(2)}(A^{(1)})^{-1}\|_{D_k^{-1}}$  then follow from using the bounds (4.30) and (4.33) in (4.36) and (4.37). The bounds in (4.29) on  $\|I - (A^{(1)})^{-1}A^{(2)}\|_2$  and  $\|I - A^{(2)}(A^{(1)})^{-1}\|_2$  follow completely analogously, except we use the bounds (4.31) and (4.34) instead of the bounds (4.30) and (4.33).  $\square$

### Proofs of Lemmas 4.19 and 4.20

The proofs of Lemmas 4.19 and 4.20 require the concept of the *adjoint* sesquilinear form to  $a_G(\cdot, \cdot)$ .

**Definition 4.21** (The adjoint sesquilinear form  $a_G^\dagger(\cdot, \cdot)$ ). *Let  $D$ ,  $A$ , and  $n$  be as in Problem 4.1. The adjoint sesquilinear form,  $a_G^\dagger(\cdot, \cdot)$ , to  $a_G(\cdot, \cdot)$  defined in (4.5) is given by*

$$a_G^\dagger(w, v) := \int_D \left( (A\nabla w) \cdot \nabla \bar{v} - k^2 n w \bar{v} \right) - (\gamma_I w, T(\gamma_I v))_{\Gamma_I}. \quad (4.38)$$

It is then straightforward to check that

$$A^\dagger := S_A - k^2 M_n - N^\dagger \quad (4.39)$$

(where  $\dagger$  denotes conjugate transpose) is the Galerkin matrix for the sesquilinear form  $a_G^\dagger(\cdot, \cdot)$ ; i.e.  $(A^\dagger)_{ij} = a_G^\dagger(\phi_j, \phi_i)$ .

The following lemma shows that if  $w$  solves an adjoint Helmholtz problem, then  $\bar{w}$  solves a (standard) Helmholtz problem with a related right-hand side.

**Lemma 4.22** (Link between variational problems involving  $a_G(\cdot, \cdot)$  and  $a_G^\dagger(\cdot, \cdot)$ ).

*If the source term  $L_G$  is as in Problem 4.1,  $w$  is the solution to Problem 4.1, the boundary operator  $T$  satisfies*

$$(T\phi, \bar{\phi})_{\Gamma_I} = (T\phi, \bar{\psi})_{\Gamma_I} \text{ for all } \phi, \psi \in H^{1/2}(\Gamma_I), \quad (4.40)$$

and

$$a_G^\dagger(w, v) = L_G(v) \text{ for all } v \in H_{0,D}^1(D), \quad (4.41)$$

then  $\overline{w}$  satisfies

$$a_G(\overline{w}, v) = \overline{L_G(\overline{v})} \quad \text{for all } v \in H_{0,D}^1(D). \quad (4.42)$$

*Proof of Lemma 4.22.* From (4.41) we have that

$$\overline{a_G^\dagger(w, \overline{v})} = \overline{L_G(\overline{v})} \quad \text{for all } v \in H_{0,D}^1(D).$$

Using the definition of  $a_G^\dagger(\cdot, \cdot)$  and the property (4.40) in the left-hand side of this last equation, we find (4.42).  $\square$

**Corollary 4.23** ((4.42) holds for Problems 2.10 and 2.12). *If Problem 4.1 is chosen to represent either Problem 2.10 or Problem 2.12, then (4.42) holds.*

*Proof of Corollary 4.23.* The only thing we need to check is that (4.40) holds for both Problems 2.10 and 2.12. For Problem 2.12, when  $T = ik$ , the proof is straightforward. For Problem 2.10 when  $T = T_R$  we need the following property of the DtN map  $T_R$ :

$$\left(T_R \psi, \overline{\phi}\right)_{\Gamma_R} = \left(T_R \phi, \overline{\psi}\right)_{\Gamma_R} \quad \text{for all } \phi, \psi \in H^{1/2}(\Gamma_R). \quad (4.43)$$

This property follows from the fact that, if  $u_1$  and  $u_2$  are solutions of the homogeneous Helmholtz equation  $\Delta u + k^2 u = 0$  in  $\mathbb{R}^d \setminus \overline{B_R}$ , both satisfying the Sommerfeld radiation condition (3.3), then

$$\int_{\Gamma_R} (\gamma u_1) \partial_\nu \overline{u_2} = \int_{\Gamma_R} (\gamma u_2) \partial_\nu \overline{u_1};$$

which follows from Green's theorem and, e.g., [200, Lemma 4.10].  $\square$

We can now prove Lemmas 4.19 and 4.20.

*Proof of Lemma 4.19.* We first concentrate on proving (4.30). Given  $f \in \mathbb{C}^N$  and  $n \in L^\infty(D; \mathbb{R})$ , we create a variational problem whose Galerkin discretisation leads to the equation  $A^{(1)} \tilde{u} = M_n f$ . Indeed, let  $\tilde{f} := \sum_j f_j \phi_j \in H_{0,D}^1(D)$ . Define  $\tilde{u}$  to be the solution of the variational problem

$$a^{(1)}(\tilde{u}, v) = \left(n \tilde{f}, v\right)_{L^2(D)} \quad \text{for all } v \in H_{0,D}^1(D), \quad (4.44)$$

and let  $\tilde{u}_b$  be the solution of the finite-element approximation of (4.44), i.e.,

$$a^{(1)}(\tilde{u}_b, v_b) = \left(n \tilde{f}, v_b\right)_{L^2(D)} \quad \text{for all } v_b \in V_{b,p}, \quad (4.45)$$

and let  $\tilde{u}$  be the vector of nodal values of  $\tilde{u}_b$ . The definition of  $\tilde{f}$  then implies that (4.45) is equivalent to the linear system  $A^{(1)} \tilde{u} = M_n f$ , and so to obtain a bound on  $\|(A^{(1)})^{-1} M_n\|_{D_k}$  we need to bound  $\|\tilde{u}\|_{D_k}$  in terms of  $\|f\|_{D_k}$ . (Recall  $f \in \mathbb{C}^N$  was arbitrary.) Because of the definition of  $\|\cdot\|_{D_k}$  in (4.10), this bound is equivalent to bounding  $\|\tilde{u}_b\|_{H_k^1(D)}$  in terms of  $\|\tilde{f}\|_{H_k^1(D)}$ .

Using the triangle inequality and the bounds (4.14) and (4.15) from Conditions 4.8 and 4.9 respectively, we find

$$\|\tilde{u}_b\|_{H_k^1(D)} \leq \|\tilde{u} - \tilde{u}_b\|_{H_k^1(D)} + \|\tilde{u}\|_{H_k^1(D)} \leq C_{\text{FEM1}}^{(1)} \|n\tilde{f}\|_{L^2(D)} + C_{\text{bound}}^{(1)} \|n\tilde{f}\|_{L^2(D)} \quad (4.46)$$

$$\leq (C_{\text{FEM1}}^{(1)} + C_{\text{bound}}^{(1)}) \|n\|_{L^\infty(D;\mathbb{R})} \|\tilde{f}\|_{L^2(D)} \quad (4.47)$$

$$\leq (C_{\text{FEM1}}^{(1)} + C_{\text{bound}}^{(1)}) \|n\|_{L^\infty(D;\mathbb{R})} \frac{\|\tilde{f}\|_{H_k^1(D)}}{k};$$

the bound on  $\|(A^{(1)})^{-1}M_n\|_{D_k}$  in (4.30) then follows from the definition of  $\|\cdot\|_{D_k}$  in (4.10) and the definition of  $C_2$  (4.32).

To prove the bound on  $\|M_n(A^{(1)})^{-1}\|_{(D_k)^{-1}}$  in (4.30), first observe that the definitions of  $\|\cdot\|_{D_k}$  and  $\|\cdot\|_{(D_k)^{-1}}$  in (4.10) imply that, for any matrix  $C \in \mathbb{C}^{N \times N}$  and for any  $v \in \mathbb{C}^N$ ,

$$\frac{\|Cv\|_{(D_k)^{-1}}}{\|v\|_{(D_k)^{-1}}} = \frac{\|C^\dagger w\|_{D_k}}{\|w\|_{D_k}} \quad (4.48)$$

where  $w := (D_k)^{1/2}v$ , and where  $C^\dagger$  is the conjugate transpose of  $C$  (i.e. the adjoint with respect to  $(\cdot, \cdot)_2$ ). Therefore, since  $M_n$  is a real, symmetric matrix,

$$\frac{\|M_n(A^{(1)})^{-1}v\|_{(D_k)^{-1}}}{\|v\|_{(D_k)^{-1}}} = \frac{\|((A^{(1)})^{-1}M_n)^\dagger w\|_{D_k}}{\|w\|_{D_k}} = \frac{\|((A^{(1)})^\dagger)^{-1}M_n w\|_{D_k}}{\|w\|_{D_k}},$$

so that

$$\|M_n(A^{(1)})^{-1}\|_{(D_k)^{-1}} = \|((A^{(1)})^\dagger)^{-1}M_n\|_{D_k}. \quad (4.49)$$

Recall from the text below (4.39) that  $(A^{(1)})^\dagger$  is the Galerkin matrix corresponding to the variational problem (4.41) – the adjoint problem. Lemma 4.22 implies that if the EDP satisfies Conditions 4.8 and 4.9, then so does the adjoint problem. Therefore, the argument above leading to the bound on  $\|(A^{(1)})^{-1}M_n\|_{D_k}$  under Condition 4.8 and Part (i) of Condition 4.9 proves the same bound on  $\|((A^{(1)})^\dagger)^{-1}M_n\|_{D_k}$ , and then, using (4.49), also on  $\|M_n(A^{(1)})^{-1}\|_{(D_k)^{-1}}$ .

To prove the bound on  $\|(A^{(1)})^{-1}M_n\|_2$  in (4.31), we use the bounds

$$m_- b^{d/2} k \|\tilde{u}\|_2 \leq k \|\tilde{u}_b\|_{L^2(D)} \leq \|\tilde{u}_b\|_{H_k^1(D)} \quad \text{and} \quad \|\tilde{f}\|_{L^2(D)} \leq m_+ b^{d/2} \|f\|_2,$$

on either side of the inequality (4.46), with these bounds coming from (4.11). The proof of the bound on  $\|M_n((A^{(1)})^\dagger)^{-1}\|_2$  in (4.31) follows in a similar way to above, using the fact that  $\|M_n(A^{(1)})^{-1}\|_2 = \|((A^{(1)})^\dagger)^{-1}M_n\|_2$  (compare to (4.49)).  $\square$

The proof of Lemma 4.20 uses the following lemma, which one can prove using the Gårding inequality (2.15); see [105, Lemma 5.1].

**Lemma 4.24** (Bound for data in  $(H_{0,D}^1(D))^*$ ). *Given  $\tilde{L}_G \in (H_{0,D}^1(D))^*$ , let  $\tilde{u}$  be the solution of the variational problem*

$$\text{find } \tilde{u} \in H_{0,D}^1(D) \text{ such that } a^{(1)}(\tilde{u}, v) = \tilde{L}_G(v) \text{ for all } v \in H_{0,D}^1(D).$$

*If Condition 4.8 holds, then  $\tilde{u}$  exists, is unique, and satisfies the bound*

$$\|\tilde{u}\|_{H_k^1(D)} \leq \frac{1}{\min\{A_{\min}^{(1)}, n_{\min}^{(1)}\}} \left(1 + 2C_{\text{bound}}^{(1)} n_{\max}^{(1)} k\right) \|\tilde{L}_G\|_{(H_k^1(D))^*} \quad (4.50)$$

for all  $k \geq k_0$ .

*Proof of Lemma 4.20.* In a similar way to the proof of Lemma 4.19, given  $f \in \mathbb{C}^N$  and  $A \in L^\infty(D; \mathbb{R}^{d \times d})$ , let  $\tilde{f} := \sum_j f_j \phi_j$  and observe that  $\tilde{f} \in H_{0,D}^1(D)$ . Define  $\tilde{u}$  to be the solution of the variational problem

$$a^{(1)}(\tilde{u}, v) = \tilde{L}_G(v) \quad \text{for all } v \in H_{0,D}^1(D), \quad \text{where } \tilde{L}_G(v) := ((A \nabla \tilde{f}, \nabla v))_{L^2(D)}. \quad (4.51)$$

Observe that the definition of the norms  $\|\cdot\|_{(H_k^1(D))^*}$  and  $\|\cdot\|_{H_k^1(D)}$  (2.14) and the Cauchy-Schwarz inequality imply that

$$\begin{aligned} \|\tilde{L}_G\|_{(H_k^1(D))^*} &\leq \|A \nabla \tilde{f}\|_{L^2(D)} \\ &\leq \|A\|_{L^\infty(D; \text{op})} \|\nabla \tilde{f}\|_{L^2(D)} \end{aligned} \quad (4.52)$$

$$\leq \|A\|_{L^\infty(D; \text{op})} \|\tilde{f}\|_{H_k^1(D)}. \quad (4.53)$$

Let  $\tilde{u}_b$  be the solution of the finite element approximation of (4.51), i.e.,

$$a^{(1)}(\tilde{u}_b, v_b) = \tilde{L}_G(v_b) \quad \text{for all } v_b \in V_{b,p}, \quad (4.54)$$

and let  $\tilde{u}$  be the vector of nodal values of  $\tilde{u}_b$ . The definition of  $\tilde{f}$  then implies that (4.54) is equivalent to  $A^{(1)}\tilde{u} = S_A f$ .

Similar to the proof of Lemma 4.19, using the triangle inequality, the bound (4.16) from Condition 4.9, the bound (4.50) from Lemma 4.24, the bound (4.53), and the definition of  $C_1$  (4.35), we find

$$\begin{aligned} \|\tilde{u}_b\|_{H_k^1(D)} &\leq \|\tilde{u} - \tilde{u}_b\|_{H_k^1(D)} + \|\tilde{u}\|_{H_k^1(D)}, \\ &\leq \left[ C_{\text{FEM2}}^{(1)} k + \frac{1}{\min\{A_{\min}^{(1)}, n_{\min}^{(1)}\}} \left(1 + 2C_{\text{bound}}^{(1)} n_{\max}^{(1)} k\right) \right] \|\tilde{L}_G\|_{(H_k^1(D))^*}, \\ &\leq C_1 k \|A\|_{L^\infty(D; \text{op})} \|\nabla \tilde{f}\|_{L^2(D)}, \\ &\leq C_1 k \|A\|_{L^\infty(D; \text{op})} \|\tilde{f}\|_{H_k^1(D)}, \end{aligned} \quad (4.55)$$

and the bound on  $\|(A^{(1)})^{-1}S_A\|_{D_k}$  in (4.33) follows.

The bound on  $\|S_A(A^{(1)})^{-1}\|_{(D_k)^{-1}}$  follows in a similar way to how we obtained the bound on  $\|M_n(A^{(1)})^{-1}\|_{(D_k)^{-1}}$  from the bound on  $\|(A^{(1)})^{-1}M_n\|_{D_k}$  in Part (i). Indeed, (4.48) and the fact that  $S_A$  is a real, symmetric matrix imply that

$$\|S_A(A^{(1)})^{-1}\|_{(D_k)^{-1}} = \|((A^{(1)})^\dagger)^{-1}S_A\|_{D_k} \quad (4.56)$$

(c.f. (4.49)), and then the arguments in the proof of part (i) imply that the bound in (4.33) on  $\|(A^{(1)})^{-1}S_A\|_{D_k}$  also holds for  $\|((A^{(1)})^\dagger)^{-1}S_A\|_{D_k}$ .

To prove the bound on  $\|(A^{(1)})^{-1}S_A\|_2$  in (4.34), we use the bounds

$$m_- b^{d/2} k \|\tilde{u}\|_2 \leq k \|\tilde{u}_b\|_{L^2(D)} \leq \|\tilde{u}_b\|_{H_k^1(D)} \quad \text{and} \quad \|\nabla \tilde{f}\|_{L^2(D)} \leq s_+ b^{d/2-1} \|f\|_2,$$

on either side of the inequality (4.55), with these bounds coming from (4.11) and (4.12) respectively. The proof of the bound on  $\|S_A((A^{(1)})^\dagger)^{-1}\|_2$  in (4.34) follows in a similar way to above, using (4.53).  $\square$

#### 4.4.2 Proofs of the finite-element results Theorems 4.11 and 4.12

We first recall properties of (weighted) GMRES that we will use to prove Theorems 4.11 and 4.12.

Let

$$W_D(C) := \left\{ (Cx, x)_D : x \in \mathbb{C}^N, \|x\|_D = 1 \right\}; \quad (4.57)$$

$W_D(C)$  is called the *numerical range* or *field of values* of  $C$  (in the  $(\cdot, \cdot)_D$  inner product).

**Theorem 4.25** (Elman estimate for weighted GMRES). *Let  $C$  be a matrix with  $0 \notin W_D(C)$ . Let  $\beta \in [0, \pi/2)$  be defined such that*

$$\cos \beta := \frac{\text{dist}(0, W_D(C))}{\|C\|_D}. \quad (4.58)$$

*If the matrix equation  $Cx = y$  is solved using weighted GMRES then, for  $m \in \mathbb{N}$ , the GMRES residual  $r_m$  satisfies*

$$\frac{\|r_m\|_D}{\|r_0\|_D} \leq \sin^m \beta. \quad (4.59)$$

The bound (4.59) with  $D = I$  was first proved in [66, Theorem 6.3] (see also [64, Theorem 3.3]) and was written in the above form in [19, Equation 1.2]. The bound (4.59) (for arbitrary Hermitian positive-definite  $D$ ) was stated implicitly (without proof) in [35, p. 247] and proved in [102, Theorem 5.1].

Theorem 4.25 has the following corollary, and the proofs of Theorems 4.11 and 4.12 follow from combining this with Theorem 4.18.

**Corollary 4.26.** *If  $\|I - C\|_D \leq \alpha < 1$ , then, with  $\beta$  defined as in (4.58),*

$$\cos \beta \geq \frac{1 - \alpha}{1 + \alpha}$$

and

$$\sin \beta \leq \frac{2\sqrt{\alpha}}{(1+\alpha)^2}. \quad (4.60)$$

*Proof of Theorem 4.11.* This follows from Theorem 4.18 by applying Corollary 4.26 first with  $C = (A^{(1)})^{-1}A^{(2)}$ ,  $D = D_k$ , and  $\alpha = 1/2$ , and then with  $C = A^{(2)}(A^{(1)})^{-1}$ ,  $D = (D_k)^{-1}$ , and  $\alpha = 1/2$ .  $\square$

*Proof of Theorem 4.12.* This follows from Theorem 4.18 by applying Corollary 4.26 first with  $C = (A^{(1)})^{-1}A^{(2)}$ ,  $D = I$ , and  $\alpha = 1/2$ , and then with  $C = A^{(2)}(A^{(1)})^{-1}$ ,  $D = I$ , and  $\alpha = 1/2$ .  $\square$

**Remark 4.27** (The improvement of the Elman estimate (4.59) in [19]). *A stronger result than (4.59) is given for standard (unweighted) GMRES in [19, Theorem 2.1], and then converted to a result about weighted GMRES in [27, Theorem 5.3]; indeed, the convergence factor  $\sin \beta$  is replaced by a function of  $\beta$  strictly less than  $\sin \beta$  for  $\beta \in (0, \pi/2)$ . Using this stronger result, however, does not improve the  $k$ -dependence of Theorem 4.11.*

#### 4.4.3 Proofs of the PDE results Theorem 4.15 and Lemma 4.16

*Proof of Theorem 4.15.* Because we assumed Condition 4.8 holds for the EDP (Problem 2.10),  $u^{(1)}$  and  $u^{(2)}$  exist, are unique, satisfy  $a^{(1)}(u^{(1)}, v) = L_G(v)$  and  $a^{(2)}(u^{(2)}, v) = L_G(v)$  for all  $v \in H_{0,D}^1(D)$ , respectively, where  $L_G$  is given by (2.18). Subtracting these equations, we find that  $u^{(1)} - u^{(2)}$  satisfies the variational problem

$$a^{(1)}(u^{(1)} - u^{(2)}, v) = \tilde{L}_G(v) \quad \text{for all } v \in H_{0,D}^1(D) \quad (4.61)$$

where

$$\tilde{L}_G(v) := \int_D \left( (A^{(2)} - A^{(1)}) \nabla u^{(2)} \right) \cdot \overline{\nabla v} + k^2 (n^{(1)} - n^{(2)}) u^{(2)} \overline{v}.$$

Now, by the Cauchy-Schwarz inequality and the definition of the norm  $\|\cdot\|_{H_k^1(D)}$  (see (2.14)), we have

$$\begin{aligned} |\tilde{L}_G(v)| &\leq \left\| A^{(1)} - A^{(2)} \right\|_{L^\infty(D; \text{op})} \|\nabla u^{(2)}\|_{L^2(D)} \|\nabla v\|_{L^2(D)} \\ &\quad + k^2 \left\| n^{(1)} - n^{(2)} \right\|_{L^\infty(D; \mathbb{R})} \|u^{(2)}\|_{L^2(D)} \|v\|_{L^2(D)} \\ &\leq \max \left\{ \left\| A^{(1)} - A^{(2)} \right\|_{L^\infty(D; \text{op})}, \left\| n^{(1)} - n^{(2)} \right\|_{L^\infty(D; \mathbb{R})} \right\} \|u^{(2)}\|_{H_k^1(D)} \|v\|_{H_k^1(D)} \end{aligned}$$

(by Cauchy-Schwarz in  $\mathbb{R}^2$ ). Therefore, by the definition of the norm  $\|\cdot\|_{(H_k^1(D))^*}$

$$\|\tilde{L}_G\|_{(H_k^1(D))^*} \leq \max \left\{ \left\| A^{(1)} - A^{(2)} \right\|_{L^\infty(D; \text{op})}, \left\| n^{(1)} - n^{(2)} \right\|_{L^\infty(D; \mathbb{R})} \right\} \|u^{(2)}\|_{H_k^1(D)}.$$

Since Condition 4.8 holds, we can then apply Lemma 4.24, i.e. the bound (4.50), to the solution of the variational problem (4.61) to find that

$$\frac{\|u^{(1)} - u^{(2)}\|_{H_k^1(D)}}{\|u^{(2)}\|_{H_k^1(D)}} \leq \frac{1}{\min\{A_{\min}^{(1)}, n_{\min}^{(1)}\}} \left(1 + 2C_{\text{bound}}^{(1)} n_{\max}^{(1)} k\right) \left(\max\left\{\|A^{(1)} - A^{(2)}\|_{L^\infty(D; \text{op})}, \|n^{(1)} - n^{(2)}\|_{L^\infty(D; \mathbb{R})}\right\}\right),$$

and then the result (4.22) follows with

$$C_3 := \frac{1}{\min\{A_{\min}^{(1)}, n_{\min}^{(1)}\}} \left(\frac{1}{k_0} + 2C_{\text{bound}}^{(1)} n_{\max}^{(1)}\right). \quad (4.62)$$

□

*Proof of Lemma 4.16.* We actually prove the stronger result that given any function  $c(k)$  such that  $c(k) > 0$  for all  $k > 0$ , there exist  $f, n^{(1)}$ , and  $n^{(2)}$  (with  $n^{(1)} \neq n^{(2)}$ ) with

$$\|n^{(1)} - n^{(2)}\|_{L^\infty(D; \mathbb{R})} \sim c(k) \quad (4.63)$$

such that the corresponding solutions  $u^{(1)}$  and  $u^{(2)}$  of Problem 2.10 with  $A^{(1)} = A^{(2)} = I$  exist, are unique, and satisfy (4.23).

The heart of the proof is the equation

$$(\Delta + k^2)(e^{ikr} \chi(r)) = e^{ikr} \left( \Delta \chi(r) + 2ik \frac{\partial \chi}{\partial r}(r) + ik \frac{d-1}{r} \chi(r) \right) =: -\tilde{f}(r), \quad (4.64)$$

where  $\chi(r)$  is chosen to have  $\text{supp } \chi \subset D$ . Observe that (4.64) is the Helmholtz operator applied to a circular wave  $e^{ikr}$ , with the added factor  $\chi$  which can be chosen to have compact support. The equation (4.64) can be proved using the formula for the Laplacian in  $d$ -dimensional spherical coordinates

$$\Delta \chi = \frac{1}{r^{d-1}} \frac{\partial}{\partial r} \left( r^{d-1} \frac{\partial \chi}{\partial r} \right) + \frac{1}{r^2} \Delta_{\mathbb{S}^{d-1}} \chi, \quad (4.65)$$

where  $\Delta_{\mathbb{S}^{d-1}}$  is the Laplace–Beltrami operator on the  $d - 1$ -dimensional sphere (see, e.g., [184, Equations (17.23) and (17.25)] for (4.65) in  $d = 2$  and  $3$ ). Observe that  $\Delta_{\mathbb{S}^{d-1}} e^{ikr} \chi(r) = 0$ .

We expect that (4.64) will be key in the proof of the sharpness of (4.22), for the following reasons. Observe that (4.64) proves the sharpness of the nontrapping resolvent estimate (4.14), since  $\|\tilde{f}\|_{L^2(D)} \sim k$  and  $\|e^{ikr} \chi(r)\|_{H_k^1(D)} \sim k$  and hence  $\|e^{ikr} \chi(r)\|_{H_k^1(D)} \sim \|\tilde{f}\|_{L^2(D)}$  (see, e.g., [38, Lemma 3.10], [199, Lemma 4.12]).

Also, recall that the nontrapping resolvent estimate (4.14) was used in the proof of the PDE bound (4.22) applied to  $u^{(1)} - u^{(2)}$ . Therefore we expect that if we set things up so that

$$u^{(1)} - u^{(2)} = e^{ikr} \chi(r), \quad (4.66)$$

then combining (4.64) and (4.66) will show the sharpness of the PDE bound (4.22). Moreover, the function  $e^{ikr} \chi(r)$  was used to prove the sharpness of resolvent estimates in [38, Discussion on p. 1445 and Lemma 3.10] and [199, Lemma 4.12], and so we can expect it will also be effective for proving sharpness in our setting.

We now set things up so that (4.66) holds. We define  $n^{(1)} = 1$  and

$$n^{(2)} = n^{(1)} + c(k) \tilde{\chi}(r), \quad (4.67)$$

for some function  $\tilde{\chi}(r)$  such that  $\tilde{\chi} \in C^\infty(D)$ ,  $\tilde{\chi} \neq 1$  (so that  $n^{(2)} \neq n^{(1)}$ ),  $\text{supp } \tilde{\chi} \subset\subset D$ , and  $\|\tilde{\chi}\|_{L^\infty(D;\mathbb{R})} = 1$  (so that  $\|n^{(1)} - n^{(2)}\|_{L^\infty(D;\mathbb{R})} = c(k)$ ). As above, let  $\chi = \chi(r)$  with  $\chi \in C^\infty(D)$  and  $\text{supp } \chi \subset\subset D$ . We will specify  $\tilde{\chi}$  and  $\chi$  in more detail later.

Let  $\tilde{f}(r)$  be as defined in (4.64), and define

$$u^{(2)}(\mathbf{x}) := -\frac{1}{k^2 c(k)} \frac{\tilde{f}(r)}{\tilde{\chi}(r)} \quad (4.68)$$

and

$$f(\mathbf{x}) := -(\Delta + k^2 n^{(2)}(\mathbf{x})) u^{(2)}(\mathbf{x}). \quad (4.69)$$

I.e.,  $u^{(2)}$  solves Problem 2.10 with coefficients  $A^{(2)} = I$  and  $n^{(2)}$  given by (4.67), and right-hand side  $f$ . We will define  $\chi$  and  $\tilde{\chi}$  below in such a way that  $u^{(2)} \in H^1(D)$  and  $f \in L^2(D)$ . In particular, we choose  $\chi$  and  $\tilde{\chi}$  so that if  $\tilde{\chi} = 0$ , then  $\chi = 0$ . Since  $\tilde{f}$  depends on  $\chi$ , this relation means we understand the right-hand side of (4.68) to be zero if  $\tilde{\chi}$  is zero. In addition, since  $\tilde{\chi}(r)$  has compact support and  $\tilde{f}$  depends on  $\chi$ , we need to tie both the support of  $\tilde{\chi}$  and how fast  $\tilde{\chi}$  vanishes in a neighbourhood of its support to the definition of  $\chi$  for both  $u^{(2)}$  and  $f$  to be well defined. As the final part of the setup, let  $u^{(1)}$  solve

$$(\Delta + k^2) u^{(1)} = -f.$$

I.e.,  $u^{(1)}$  solves Problem 2.10 with coefficients  $A^{(1)} = I$  and  $n^{(1)} = 1$  and right-hand side  $f$ .

Now observe that by construction (since  $n^{(2)}$  is given by (4.67))

$$\begin{aligned} (\Delta + k^2)(u^{(1)} - u^{(2)}) &= (\Delta + k^2)u^{(1)} - (\Delta + k^2 n^{(2)} - k^2(n^{(2)} - 1))u^{(2)} \\ &= -f + f + k^2(n^{(2)} - 1)u^{(2)} \\ &= k^2(n^{(2)} - 1)u^{(2)} \\ &= k^2 c(k) \tilde{\chi} \frac{-1}{k^2 c(k)} \frac{\tilde{f}}{\tilde{\chi}} \\ &= -\tilde{f} \\ &= (\Delta + k^2)(e^{ikr} \chi(r)). \end{aligned}$$

Therefore, by uniqueness of the solution of Problem 2.10 (with constant coefficients)

$$u^{(1)}(\mathbf{x}) - u^{(2)}(\mathbf{x}) = e^{ikr} \chi(r). \quad (4.70)$$

Therefore, by (4.70) and the properties of  $e^{ikr} \chi(r)$  discussed above, we have

$$\|u^{(1)} - u^{(2)}\|_{L^2(D)} \sim 1 \quad \text{and} \quad \|u^{(1)} - u^{(2)}\|_{H_k^1(D)} \sim k. \quad (4.71)$$

Furthermore, the definitions of  $u^{(2)}$  and  $\tilde{f}$  imply that

$$\|u^{(2)}\|_{L^2(D)} \sim \frac{1}{k c(k)} \quad \text{and} \quad \|u^{(2)}\|_{H_k^1(D)} \sim \frac{1}{c(k)}, \quad (4.72)$$

and, since  $\|n^{(1)} - n^{(2)}\|_{L^\infty(D)} = c(k)$ , by combining (4.71) and (4.72), we see (4.63) holds, as required.

Therefore, to complete the proof, we only need to show that there exists a choice of  $\chi$  and  $\tilde{\chi}$  for which  $u^{(2)}$  and  $f$  defined by (4.68) and (4.69) are in  $H^1(D)$  and  $L^2(D)$  respectively (in fact, we prove that they are in  $W^{1,\infty}(D)$  and  $L^\infty(D)$  respectively). Because  $\chi$  and  $\tilde{\chi}$  are in  $C^\infty(D)$  and we choose  $\chi$  and  $\tilde{\chi}$  so that if  $\tilde{\chi} = 0$ , then  $\chi = 0$ , the only issue is what happens at the boundary of the support of  $\tilde{\chi}$ , where  $u^{(2)}$  has the potential to be singular. Since  $\overline{D_-} \subset B_R$ , there exist  $0 < R_1 < R_2 < R$  such that  $\overline{D_-} \subset B_{R_2} \setminus B_{R_1} \subset B_R$ . Let  $\text{supp } \chi = B_{R_2} \setminus B_{R_1}$  and let  $\chi$  vanish to order  $m$  at  $r = R_1$  and  $r = R_2$ ; i.e.  $\chi(r) \sim (r - R_1)^m$  as  $r \downarrow R_1$  and  $\chi(r) \sim (R_2 - r)^m$  as  $r \uparrow R_2$ . The definition of  $\tilde{f}$  (4.64) then implies that  $\tilde{f}$  vanishes to order  $m - 2$ . Let  $\tilde{\chi}(r)$  vanish to order  $\tilde{m}$  at  $r = R_1$  and  $r = R_2$ . We now claim that if  $m > \tilde{m} + 4$ , then  $u^{(2)} \in W^{1,\infty}(D)$  and  $f \in L^\infty(D)$ . Indeed, straightforward calculation from (4.68) shows that  $u^{(2)}(r) \sim (r - R_1)^{m - \tilde{m} - 2}$ ,  $\nabla u^{(2)}(r) \sim (r - R_1)^{m - \tilde{m} - 3}$ , and  $\Delta u^{(2)}(r) \sim (r - R_1)^{m - \tilde{m} - 4}$  as  $r \downarrow R_1$ , with analogous behaviour at  $r = R_2$ . The assumption  $m > \tilde{m} + 4$  therefore implies that  $u^{(2)}$ ,  $\nabla u^{(2)}$ , and  $\Delta u^{(2)}$  vanish (and hence are finite) at  $r = R_1$  and  $r = R_2$ .  $\square$

**Remark 4.28** (Why doesn't Lemma 4.16 cover the case  $A^{(1)} \neq A^{(2)}$ ?). When  $n^{(j)} := 1$ ,  $j = 1, 2$ ,  $A^{(1)} := I$ , and  $A^{(2)} := I + c(k)\tilde{\chi}$ , the variational problem (4.61) implies that

$$\Delta(u^{(1)} - u^{(2)}) + k^2(u^{(1)} - u^{(2)}) = c(k)\nabla \cdot (\tilde{\chi}\nabla u^{(2)}). \quad (4.73)$$

It is now much harder than in (4.73) to set things up so that  $u^{(1)}(\mathbf{x}) - u^{(2)}(\mathbf{x}) = e^{ikr} \chi(r)$  (so that one can then use (4.64)).

## 4.5 EXTENSION OF THE NEARBY PRECONDITIONING RESULTS TO WEAKER NORMS

Recall from Sections 4.2 and 4.3 that GMRES applied to  $(A^{(1)})^{-1}A^{(2)}$  converges in a  $k$ -independent number of iterations if  $k\|n^{(1)} - n^{(2)}\|_{L^\infty(D;\mathbb{R})}$  is sufficiently small (with an analogous result for  $A^{(1)} - A^{(2)}$ ). This result (and the related numerics) shows that  $1/k$  may be a sharp threshold when

we consider the maximum norm of the difference between  $n^{(1)}$  and  $n^{(2)}$ . However, this result does not say anything if  $n^{(1)} - n^{(2)}$  is merely small in some integral norm. For example if  $n^{(1)}$  and  $n^{(2)}$  (defined on the unit square) are given by

$$n^{(1)}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x}_1 \leq \frac{1}{2} \\ 2 & \text{if } \mathbf{x}_1 > \frac{1}{2} \end{cases} \quad (4.74)$$

and

$$n^{(2)}(\mathbf{x}) = \begin{cases} 1 & \text{if } \mathbf{x}_1 \leq \frac{1}{2} + \alpha \\ 2 & \text{if } \mathbf{x}_1 > \frac{1}{2} + \alpha \end{cases} \quad (4.75)$$

for some  $0 < \alpha < 1/2$ , then  $\|n^{(1)} - n^{(2)}\|_{L^\infty(D;\mathbb{R})} = 1$  for all  $\alpha$ , but one would expect that for small  $\alpha$  the corresponding solutions of Problem 2.1 would satisfy  $u^{(1)} \approx u^{(2)}$ . In addition, one might expect that GMRES applied to  $(A^{(1)})^{-1}A^{(2)}$  would converge in a  $k$ -independent number of iterations. Therefore, in this section we seek to obtain analogues of Theorems 4.11, 4.12, and 4.18 with the difference in  $n^{(1)} - n^{(2)}$  and  $A^{(1)} - A^{(2)}$  measured in weaker norms than the  $L^\infty$  norm.

The (realistic) best-case result we could obtain would be that GMRES applied to  $(A^{(1)})^{-1}A^{(2)}$  converges in a  $k$ -independent number of iterations if  $\|n^{(1)} - n^{(2)}\|_{L^1(D;\mathbb{R})} \lesssim 1/k$ . This result is ‘best’ in the sense that it depends optimally on  $k$ ; recall the discussion in Remark 4.17 that  $1/k$  is the length scale governing the behaviour of Helmholtz problems. In addition, we measure  $n^{(1)} - n^{(2)}$  in the  $L^\infty$  norm as above, we are able to control the magnitude of  $n^{(1)} - n^{(2)}$ , but not the spatial variability; if  $n^{(1)} - n^{(2)} \neq 0$  only on a set of small (but nonzero) measure, and  $n^{(1)} - n^{(2)} = 1$  on this small set, then  $\|n^{(1)} - n^{(2)}\|_{L^\infty(D;\mathbb{R})} = 1$ , regardless of the measure of the set. In contrast, the  $L^1$  norm allows us to control both the magnitude of  $n^{(1)} - n^{(2)}$  and the measure of the sets on which it is nonzero.

We will give numerical results indicating that this theoretical best-case result can be achieved (our numerical results actually indicate that we can obtain  $k$ -independent convergence when  $\|n^{(1)} - n^{(2)}\|_{L^q(D;\mathbb{R})} \sim 1/k$  for any  $1 \leq q < \infty$ ). We will also provide theory results that are, to our knowledge, the best one can prove, although they are sub-optimal in both  $q$  and the dependence on  $k$ .

### 4.5.1 Theory in weaker norms

Before we prove results analogous to Theorems 4.11 and 4.12 in weaker norms (using a result analogous to Theorem 4.18 in weaker norms), we first recap why the terms  $\|A^{(1)} - A^{(2)}\|_{L^\infty(D;\text{op})}$  and  $\|n^{(1)} - n^{(2)}\|_{L^\infty(D;\mathbb{R})}$  appear in Theorem 4.18. These terms appear in Theorem 4.18 because the terms  $\|n\|_{L^\infty(D;\mathbb{R})}$  and  $\|A\|_{L^\infty(D;\text{op})}$  appear in Lemmas 4.19 and 4.20, respectively. These terms

appear in these lemmas because in (4.47) and (4.52) we use the bounds

$$\|nf\|_{L^2(D)} \leq \|n\|_{L^\infty(D;\mathbb{R})} \|f\|_{L^2(D)} \quad (4.76)$$

and

$$\|A\nabla\tilde{f}\|_{L^2(D)} \leq \|A\|_{L^\infty(D;\text{op})} \|\nabla\tilde{f}\|_{L^2(D)} \quad (4.77)$$

respectively, for an arbitrary function  $\tilde{f} \in V_{b,p}$ , and these bounds are carried through the rest of the proof.

However, we observe that we have the following generalisation of Hölder's inequality: If  $q, s > 2$  such that  $1/2 = 1/q + 1/s$ , then

$$\|v_1 v_2\|_{L^2(D)} \leq \|v_1\|_{L^q(D)} \|v_2\|_{L^s(D)}. \quad (4.78)$$

If we instead use (4.78) to bound (4.76) and (4.77) we obtain

$$\|nf\|_{L^2(D)} \leq \|n\|_{L^q(D,\mathbb{R})} \|f\|_{L^s(D)} \quad (4.79)$$

and

$$\|A\nabla\tilde{f}\|_{L^2(D)} \leq \|A\|_{L^q(D;\text{op})} \|\nabla\tilde{f}\|_{L^s(D)}. \quad (4.80)$$

As  $\tilde{f} \in V_{b,p}$ , we can apply an inverse inequality to bound  $\|f\|_{L^s(D)}$  by  $\|f\|_{L^2(D)}$ . The required inverse inequality is (see [29, Theorem 4.5.11 and Remark 4.5.20])

$$\|f\|_{L^s(D)} \leq C_{\text{inv},s} b^{d(\frac{1}{s}-\frac{1}{2})} \|f\|_{L^2(D)}. \quad (4.81)$$

If we then apply (4.81) to (4.79) and (4.80) we obtain

$$\|nf\|_{L^2(D)} \leq C_{\text{inv},s} \|n\|_{L^q(D,\mathbb{R})} b^{d(\frac{1}{s}-\frac{1}{2})} \|f\|_{L^2(D)} = C_{\text{inv},s} \|n\|_{L^q(D,\mathbb{R})} b^{-\frac{d}{q}} \|f\|_{L^2(D)} \quad (4.82)$$

and

$$\|A\nabla\tilde{f}\|_{L^2(D)} \leq C_{\text{inv},s} \|A\|_{L^q(D;\text{op})} b^{d(\frac{1}{s}-\frac{1}{2})} \|\nabla\tilde{f}\|_{L^2(D)} = C_{\text{inv},s} \|A\|_{L^q(D;\text{op})} b^{-\frac{d}{q}} \|\nabla\tilde{f}\|_{L^2(D)}. \quad (4.83)$$

Replacing (4.47) and (4.52) with (4.82) and (4.83) in the proofs of Lemmas 4.19 and 4.20, and proceeding as in those proofs, we can obtain the following theorems, the analogues of Theorems 4.11 and 4.12.

**Theorem 4.29** (Alternative answer to Q1:  $k$ -independent weighted GMRES iterations).

Let the assumptions of Theorem 4.11 hold. Given  $q > 2$ , there exist  $\tilde{C}_1, \tilde{C}_2 > 0$ , independent of  $h$  and  $k$  (but dependent on  $d, D_-, A^{(1)}, n^{(1)}, p, q$ , and  $k_\circ$ ) such that if

$$\tilde{C}_1 k b^{-\frac{d}{q}} \|A^{(1)} - A^{(2)}\|_{L^q(D;\text{op})} + \tilde{C}_2 k b^{-\frac{d}{q}} \|n^{(1)} - n^{(2)}\|_{L^q(D,\mathbb{R})} \leq \frac{1}{2} \quad (4.84)$$

then both weighted GMRES working in  $\|\cdot\|_{D_k}$  (and the associated inner product) applied to (4.18) and weighted GMRES working in  $\|\cdot\|_{(D_k)^{-1}}$  (and the associated inner product) applied to (4.19) converge in a  $k$ -independent number of iterations.

**Theorem 4.30** (Alternative answer to Q1:  $k$ -independent (unweighted) GMRES iterations).

Let the assumptions of Theorem 4.12 hold. Given  $q > 2$ , there exist  $\tilde{C}_1, \tilde{C}_2 > 0$ , independent of  $h$  and  $k$  (but dependent on  $d, D_-, A^{(1)}, n^{(1)}, p, q$ , and  $k_\circ$ ) such that if

$$\tilde{C}_1 \left( \frac{s_+}{m_-} \right) h^{-\frac{d}{q}-1} \|A^{(1)} - A^{(2)}\|_{L^q(D; \text{op})} + \tilde{C}_2 \left( \frac{m_+}{m_-} \right) k h^{-\frac{d}{q}} \|n^{(1)} - n^{(2)}\|_{L^q(D; \mathbb{R})} \leq \frac{1}{2}, \quad (4.85)$$

then standard GMRES (working in the Euclidean norm and inner product) applied to either of the equations (4.18) or (4.19) converges in a  $k$ -independent number of iterations.

A sketch proof of Theorems 4.29 and 4.30 is on page 168 below.

**Remark 4.31** (Trade off between the type of norm and powers of  $h$  and  $k$ ). Observe that in Theorems 4.29 and 4.30 there is a trade-off between the norm that one uses to measure  $n^{(1)} - n^{(2)}$  (or  $A^{(1)} - A^{(2)}$ ) and the restriction on the magnitude of this norm. E.g., the condition on  $n^{(1)} - n^{(2)}$  in both Theorems 4.29 and 4.30 can be summarised as

$$\|n^{(1)} - n^{(2)}\|_{L^q(D; \mathbb{R})} k h^{-\frac{d}{q}} \text{ is sufficiently small.} \quad (4.86)$$

with analogous conditions on  $A^{(1)} - A^{(2)}$ . Observe that as  $q \downarrow 2$ , we measure  $n^{(1)} - n^{(2)}$  in a weaker norm, but the condition (4.86) becomes more restrictive; the power of  $h$  increases. Conversely, as  $q \uparrow \infty$ , we measure  $n^{(1)} - n^{(2)}$  in a stronger norm, but the condition (4.86) becomes less restrictive; the power of  $h$  decreases. (Also observe that in the  $q \uparrow \infty$  limit we recover the condition (4.21) we previously proved for  $\|n^{(1)} - n^{(2)}\|_{L^\infty(D; \mathbb{R})}$ .)

**Remark 4.32** (Theorems 4.11 and 4.12 are a special case of Theorems 4.29 and 4.30). Observe that in the case  $q = \infty$  Theorems 4.29 and 4.30 become our previous results in the  $L^\infty$  norm, Theorems 4.11 and 4.12.

The numerical experiments in Section 4.5.2 below suggest that, at least in certain cases, a sufficient condition for nearby preconditioning to be effective is

$$\|n_1 - n_2\|_{L^q(D; \mathbb{R})} k \text{ is sufficiently small,} \quad (4.87)$$

for any  $q \geq 1$ , and moreover (4.87) appears sharp in its  $k$ -dependence. (This requirement would fit with our previous observation about  $1/k$  being the length scale below which perturbations cannot be seen—see Remark 4.17 above.) However, we do not say that (4.87) is sufficient for all cases; recall that for transmission problems, very small perturbations in  $n$  can lead to very different behaviour in the solution  $u$  if  $k$  is a quasi-resonance for  $n_1$  or  $n_2$ ; see the discussion at the end of Section 2.2.3 above.

**Proof of Theorems 4.29 and 4.30**

We first state analogues of Lemmas 4.19 and 4.20 in weaker norms; these lemmas are the key to the proofs of Theorems 4.29 and 4.30 above. The essence of the proofs of Lemmas 4.33 and 4.34 are the discussion at the start of Section 4.5.1.

**Lemma 4.33** (Alternative bounds on  $(A^{(1)})^{-1}M_n$ ). *Under the assumptions of Lemma 4.19, for  $n \in L^\infty(D; \mathbb{R})$  and for any  $q > 2$ ,*

$$\max \left\{ \left\| (A^{(1)})^{-1}M_n \right\|_{D_k}, \left\| M_n(A^{(1)})^{-1} \right\|_{D_k^{-1}} \right\} \leq \tilde{C}_2 h^{-\frac{d}{q}} \frac{\|n\|_{L^q(D, \mathbb{R})}}{k} \quad (4.88)$$

and

$$\max \left\{ \left\| (A^{(1)})^{-1}M_n \right\|_2, \left\| M_n(A^{(1)})^{-1} \right\|_2 \right\} \leq \tilde{C}_2 \left( \frac{m_+}{m_-} \right) h^{-\frac{d}{q}} \frac{\|n\|_{L^q(D, \mathbb{R})}}{k}, \quad (4.89)$$

where

$$\tilde{C}_2 := C_{\text{inv},s} C_2, \quad (4.90)$$

where  $C_2$  is defined by (4.32) and  $1/s = 1/2 - 1/q$ .

**Lemma 4.34** (Alternative bounds on  $(A^{(1)})^{-1}S_A$ ). *Under the assumptions of Lemma 4.20, for  $A \in L^\infty(D, \mathbb{R}^{d \times d})$  and for any  $q > 2$*

$$\max \left\{ \left\| (A^{(1)})^{-1}S_A \right\|_{D_k}, \left\| S_A(A^{(1)})^{-1} \right\|_{D_k^{-1}} \right\} \leq \tilde{C}_1 h^{-\frac{d}{q}} k \|A\|_{L^q(D; \text{op})} \quad (4.91)$$

and

$$\max \left\{ \left\| (A^{(1)})^{-1}S_A \right\|_2, \left\| S_A(A^{(1)})^{-1} \right\|_2 \right\} \leq \tilde{C}_1 \left( \frac{s_+}{m_-} \right) h^{-\frac{d}{q}-1} \|A\|_{L^q(D, \mathbb{R})}, \quad (4.92)$$

where

$$\tilde{C}_1 := C_{\text{inv},s} C_1, \quad (4.93)$$

where  $C_1$  is given by (4.35) and  $1/s = 1/2 - 1/q$ .

The proofs of Lemmas 4.33 and 4.34 are virtually identical to the proofs of Lemmas 4.19 and 4.20, with the modifications for  $L^q$  norms detailed at the beginning of Section 4.5.1.

**Remark 4.35** (Reduction to Lemmas 4.19 and 4.20). *Observe that in the case  $s = 2$  and  $q = \infty$  Lemmas 4.33 and 4.34 reduce to our previous results Lemmas 4.19 and 4.20.*

We can use Lemmas 4.33 and 4.34 in place of Lemmas 4.19 and 4.20 to obtain the following analogue of Theorem 4.18 in weaker norms.

**Theorem 4.36** (Alternative main ingredient to answer to Q1). *If all the assumptions of Theorem 4.18 hold, then there exist  $\tilde{C}_1, \tilde{C}_2 > 0$ , independent of  $h$  and  $k$  (but dependent on  $d, D_-, A^{(1)}, n^{(1)}, p$ ,*

$q$ , and  $k_0$ ) such that

$$\begin{aligned} & \max \left\{ \left\| I - (A^{(1)})^{-1} A^{(2)} \right\|_{D_k}, \left\| I - A^{(2)} (A^{(1)})^{-1} \right\|_{D_k^{-1}} \right\} \\ & \leq \tilde{C}_1 k b^{-\frac{d}{q}} \left\| A^{(1)} - A^{(2)} \right\|_{L^q(D; \text{op})} + \tilde{C}_2 k b^{-\frac{d}{q}} \left\| n^{(1)} - n^{(2)} \right\|_{L^q(D, \mathbb{R})} \end{aligned} \quad (4.94)$$

and

$$\begin{aligned} & \max \left\{ \left\| I - (A^{(1)})^{-1} A^{(2)} \right\|_2, \left\| I - A^{(2)} (A^{(1)})^{-1} \right\|_2 \right\} \\ & \leq \tilde{C}_1 \left( \frac{s_+}{m_-} \right) b^{-\frac{d}{q}-1} \left\| A^{(1)} - A^{(2)} \right\|_{L^q(D; \text{op})} + \tilde{C}_2 \left( \frac{m_+}{m_-} \right) k b^{-\frac{d}{q}} \left\| n^{(1)} - n^{(2)} \right\|_{L^q(D, \mathbb{R})}. \end{aligned} \quad (4.95)$$

The proof of Theorem 4.36 is identical to the proof of Theorem 4.18, with Lemmas 4.19 and 4.20 replaced by Lemmas 4.33 and 4.34.

*Sketch proof of Theorems 4.29 and 4.30.* The proofs of Theorems 4.29 and 4.30 are completely analogous to the proofs of Theorems 4.11 and 4.12, with the exception that we use Theorem 4.36 in place of Theorem 4.18.  $\square$

#### 4.5.2 Numerics in weaker norms

For our computations, we use the computational setup as in Appendix G, with  $f$  and  $g_I$  corresponding to a plane wave passing through homogeneous media. We let  $A^{(1)} = A^{(2)} = I$ , and we define  $n^{(1)}$  and  $n^{(2)}$  by (4.74) and (4.75). For  $\alpha = 0.2k^{-\beta}$ ,  $\beta = 0, 0.1, \dots, 0.9, 1$  and for  $k = 10, 20, \dots, 100$  we used GMRES to solve  $(A^{(1)})^{-1} A^{(2)} = (A^{(1)})^{-1} f$  (for  $f$  given by the Helmholtz problem), and we record the number of GMRES iterations taken to achieve convergence.

Our results in Figures 4.7–4.9 (also displayed in Table 4.1) indicate the following conclusions for  $\left\| n^{(1)} - n^{(2)} \right\|_{L^q(D, \mathbb{R})} \sim 0.1/k^{-\beta}$ , for all  $1 \leq q < \infty$ :

- For  $\beta \in (0, 0.6)$  there is clear growth of the number of GMRES iterations with  $k$ ,
- For  $\beta = 1$  there is clear boundedness of the number of GMRES iterations with  $k$ , and
- for  $\beta \in (0.7, 0.9)$  it is unclear if the number of GMRES iterations grows with  $k$ .

We note that the results in Figures 4.7–4.9 are the analogues of those in Figures 4.4–4.6.

If we compare our numerical results with the theory results in Theorem 4.30, we see that the theory (if  $h \sim k^{-3/2}$  and  $d = 2$ , as in our computational experiments) predicts that the number of iterations will remain bounded if  $\left\| n^{(1)} - n^{(2)} \right\|_{L^q(D, \mathbb{R})} k^{1+3/q}$  is sufficiently small, for any  $q > 2$ . Our computed results indicate that this result is not sharp. The computed results indicate that if  $\left\| n^{(1)} - n^{(2)} \right\|_{L^q(D, \mathbb{R})} \sim k^{-1}$  for any  $q \geq 1$ , then the number of GMRES iterations is bounded as  $k$  increases. Observe again that the ‘best case’  $1/k$  condition is only predicted by the theory in the  $q \rightarrow \infty$  limit.

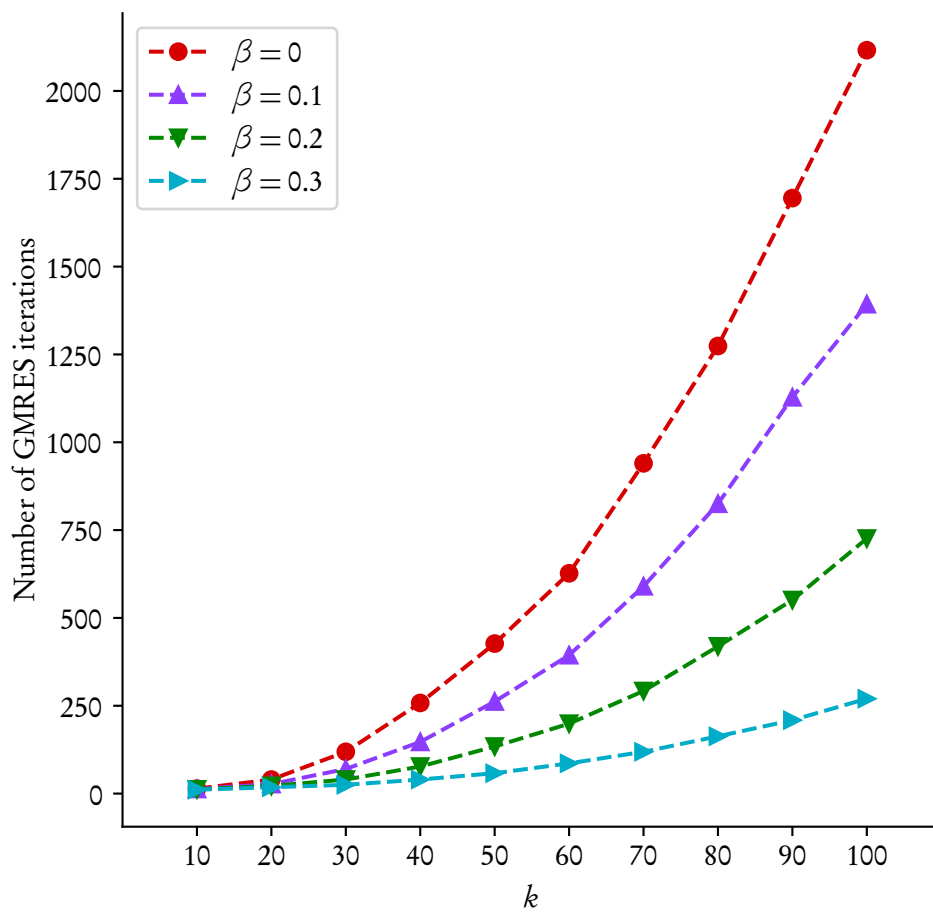


Figure 4.7: GMRES iteration counts for  $(A^{(1)})^{-1}A^{(2)}$  given by (4.74) and (4.75), where  $\alpha = 0.2 \times k^{-\beta}$ , for  $\beta = 0, 0.1, 0.2, 0.3$ .

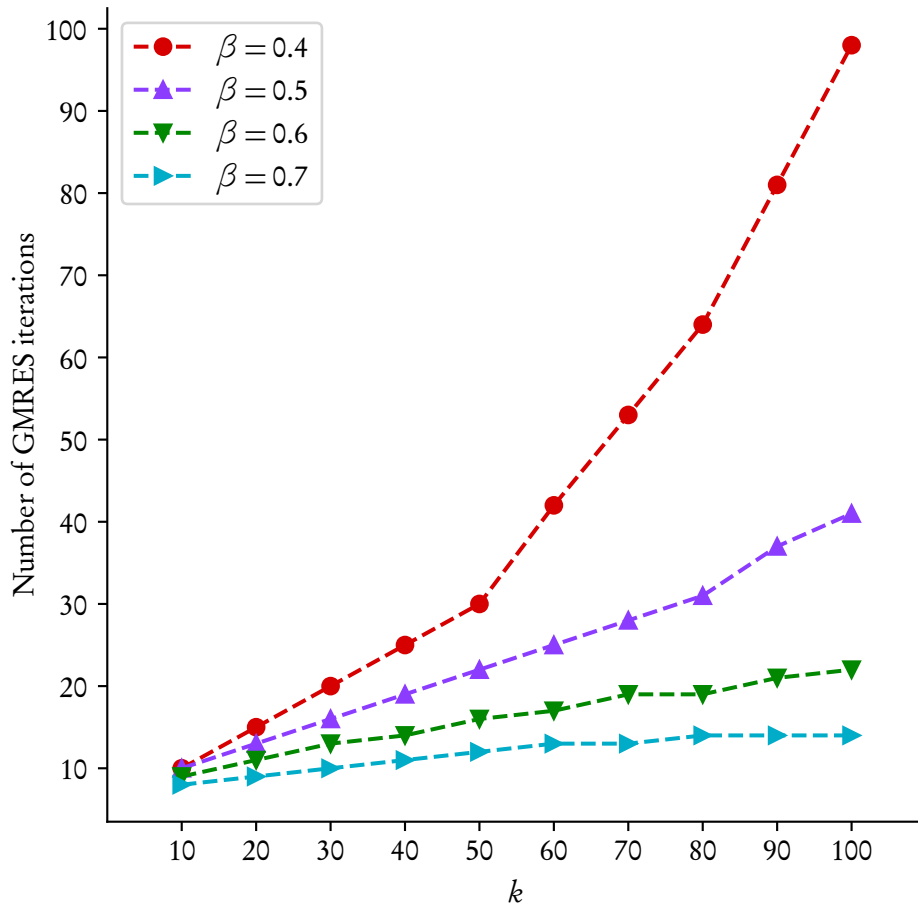


Figure 4.8: GMRES iteration counts for  $(A^{(1)})^{-1}A^{(2)}$  given by (4.74) and (4.75), where  $\alpha = 0.2 \times k^{-\beta}$ , for  $\beta = 0.4, 0.5, 0.6, 0.7$ .

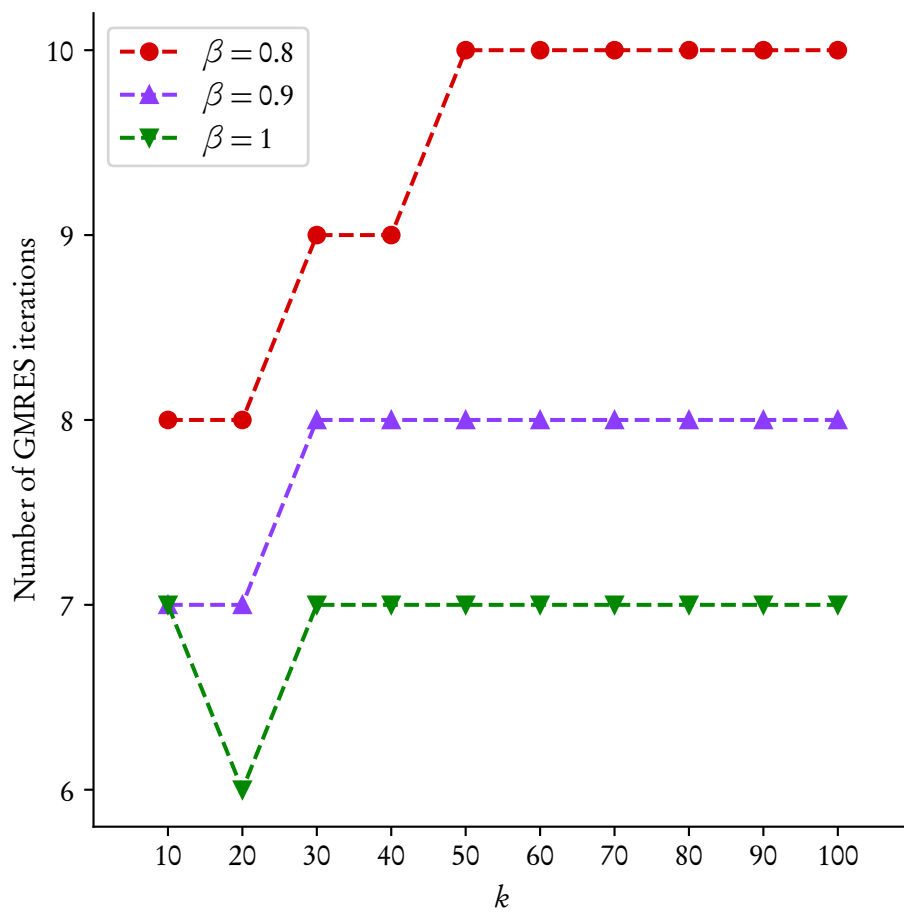


Figure 4.9: GMRES iteration counts for  $(A^{(1)})^{-1}A^{(2)}$  given by (4.74) and (4.75), where  $\alpha = 0.2 \times k^{-\beta}$ , for  $\beta = 0.8, 0.9, 1$ .

| $\beta \backslash k$ | 10 | 20 | 30  | 40  | 50  | 60  | 70  | 80   | 90   | 100  |
|----------------------|----|----|-----|-----|-----|-----|-----|------|------|------|
| 0.0                  | 14 | 40 | 119 | 258 | 427 | 627 | 940 | 1274 | 1695 | 2116 |
| 0.1                  | 13 | 27 | 70  | 147 | 262 | 394 | 590 | 825  | 1128 | 1393 |
| 0.2                  | 12 | 22 | 40  | 77  | 134 | 199 | 292 | 419  | 551  | 726  |
| 0.3                  | 11 | 18 | 25  | 40  | 58  | 86  | 119 | 163  | 209  | 270  |
| 0.4                  | 10 | 15 | 20  | 25  | 30  | 42  | 53  | 64   | 81   | 98   |
| 0.5                  | 10 | 13 | 16  | 19  | 22  | 25  | 28  | 31   | 37   | 41   |
| 0.6                  | 9  | 11 | 13  | 14  | 16  | 17  | 19  | 19   | 21   | 22   |
| 0.7                  | 8  | 9  | 10  | 11  | 12  | 13  | 13  | 14   | 14   | 14   |
| 0.8                  | 8  | 8  | 9   | 9   | 10  | 10  | 10  | 10   | 10   | 10   |
| 0.9                  | 7  | 7  | 8   | 8   | 8   | 8   | 8   | 8    | 8    | 8    |
| 1.0                  | 7  | 6  | 7   | 7   | 7   | 7   | 7   | 7    | 7    | 7    |

Table 4.1: GMRES iteration counts for  $(A^{(1)})^{-1}A^{(2)}$  given by (4.74) and (4.75), where  $\alpha = 0.2 \times k^{-\beta}$ .

## 4.6 APPLYING NEARBY PRECONDITIONING TO A QUASI-MONTE-CARLO METHOD FOR THE HELMHOLTZ EQUATION

We now apply nearby preconditioning in the implementation of a Quasi-Monte-Carlo (QMC) method for the Helmholtz equation. We begin with a brief description of QMC methods, before detailing two ways in which we apply nearby preconditioning to these methods. Finally, we give computational results illustrating this application.

### 4.6.1 Brief description of QMC

QMC methods (or rules) are high-dimensional quadrature rules, designed to give rates of convergence (with respect to the number of integration points) which are superior to those of Monte-Carlo methods, under certain conditions. Suppose one wants to approximate  $\mathbb{E}[Q]$ , where  $Q$  is some random variable (later in this section,  $Q$  will be a function of the solution  $u(\omega)$  of a stochastic Helmholtz equation). By definition, the expectation is

$$\mathbb{E}[Q] = \int_{\Omega} Q(\omega) d\mathbb{P}(\omega). \quad (4.96)$$

If we now suppose  $Q$  depends on the sample space  $\Omega$  via a finite set of random variables  $U_1, \dots, U_j$ , then we can rewrite (4.96) as

$$\mathbb{E}[Q] = \int_{\Omega} Q((U_1(\omega), \dots, U_j(\omega))) d\mathbb{P}(\omega). \quad (4.97)$$

If, for example, the  $U_j$  are all independent uniform random variables on  $[-1/2, 1/2]$ , then (4.97)

can be rewritten as

$$\mathbb{E}[Q] = \int_{[-1/2, 1/2]^J} Q(\mathbf{y}) d\lambda(\mathbf{y}), \quad (4.98)$$

where  $\lambda$  denotes Lebesgue measure.

Any quadrature rule, or method for approximating  $\mathbb{E}[Q]$ , can then be seen as a method for approximating the  $J$ -dimensional integral on the right-hand side of (4.98) and vice-versa. Equal-weight quadrature rules choose points  $\mathbf{y}_1, \dots, \mathbf{y}_{N_{\text{points}}} \in [-1/2, 1/2]^J$  and use the approximation

$$\mathbb{E}[Q] \approx \frac{1}{N_{\text{points}}} \sum_{l=1}^{N_{\text{points}}} Q(\mathbf{y}_l).$$

Monte-Carlo and Quasi-Monte-Carlo rules correspond to different choices of the points  $\mathbf{y}_l$ . In a Monte-Carlo rule the points are chosen at random in accordance with the associated probability distribution. For example, in the case that the  $U_j$  are  $\text{Unif}(-1/2, 1/2)$  random variables, the points  $\mathbf{y}_l$  are chosen according to the Uniform distribution on  $[-1/2, 1/2]^J$ . Observe that Monte-Carlo rules do not need the dependence of  $Q$  on  $\omega$  to take the form prescribed in (4.97), indeed, they apply to any random variable.

Quasi-Monte-Carlo rules, in contrast to Monte-Carlo rules, do require the dependence on  $\omega$  to be via finitely- or countable-many random variables. This is because QMC rules are high-dimensional quadrature rules (in the simplest case performing quadrature on the high-dimensional cube  $[-1/2, 1/2]^J$ ). In pure QMC rules the points  $\mathbf{y}_l$  are chosen deterministically, unlike Monte-Carlo rules.

The main advantage of QMC rules is that they can exhibit higher rates of convergence compared to Monte-Carlo rules; Monte Carlo rules typically converge with rate  $N_{\text{points}}^{-1/2}$  (see, e.g., [95, Section 1.1]), whereas QMC rules can converge with rates up to  $N_{\text{points}}^{-1}$  or with even higher rates for higher-order QMC rules, see, e.g., [131, Penultimate paragraph of Section 1.2].

In applying QMC rules to stochastic PDEs, we assume that the random coefficient ( $n$  in our case) is defined via finitely many (or countably many) random variables, as in (4.97) above, and we then use QMC rules to estimate expectations of quantities of interest of the solution  $u$ , i.e.,  $Q = Q(u)$ . We note that applying QMC rules to stochastic PDEs is a vibrant and active research area. For recent overviews of this field, see [131, 133] (and the associated tutorial [132]). We note that there is currently no rigorous study of how QMC methods behave for the Helmholtz equation, although we understand some such work is currently underway by Ganesh, Kuo, and Sloan [88].

#### 4.6.2 Methods for applying nearby preconditioning to QMC

In all of our previous uses of nearby preconditioning, we have fixed  $n^{(1)}$ , the value for which we calculate the preconditioner, and have then used  $A^{(1)}$  to precondition  $A^{(2)}$  for different values of  $n^{(2)}$ . However, the key idea for applying nearby preconditioning to QMC methods for the Helmholtz equation is to choose *a number* of different realisations of  $n^{(1)}$  and use each realisation of  $n^{(1)}$  as a preconditioner only for those realisations of  $n^{(2)}$  for which  $A^{(1)}$  is a *good* preconditioner

for  $A^{(2)}$ . We adopt this approach because it is highly unlikely that a single realisation of  $n^{(1)}$  will be a good preconditioner for every realisation of  $n^{(2)}$ .

Therefore, the algorithms presented in this section seek to answer the two questions:

1. For which realisations of  $n$  should a preconditioner be *calculated*?
2. To which realisations of  $n$  should each preconditioner be *applied*?

We now detail two methods for using nearby preconditioning to speed up QMC methods for the Helmholtz equation. To apply these methods, we use the following model problem: We consider the Interior Impedance Problem in 2-d with  $f = 1$  and  $g_I = 0$ ,  $A = I$ , and  $n$  given by

$$n(\omega, \mathbf{x}) = 1 + \sum_{j=1}^{10} U_j(\omega) \sqrt{\lambda_j} \psi_j(\mathbf{x}), \quad (4.99)$$

where

$$\sqrt{\lambda_j} = j^{-2} \quad (4.100)$$

and

$$\psi_j(\mathbf{x}) = \cos\left(\frac{j\pi}{4}x\right) \cos\left(\frac{(j+1)\pi}{4}y\right). \quad (4.101)$$

Observe that  $\|\psi_j\|_{L^\infty(D_R)} = 1$  for all  $j$ , and  $\sqrt{\lambda_j} \rightarrow 0$  as  $j \rightarrow \infty$ . Also note that  $n_{\min} = 1 - (\sum_{j=1}^{10} j^{-2})/2 \approx 0.225$ . This expansion is based on the random-field expansion used in [92, Section 5.1], although the main change we make from [92] is to introduce the factors  $1/4$  in (4.101). We introduce this factor to ensure that the oscillations in the medium  $n$  are ‘low frequency’ compared to the frequency  $k$  of the waves passing through the medium<sup>1</sup>. Expansions similar to (4.99) are often described as ‘artificial Karhunen–Loève expansions’ due to their similarity with the Karhunen–Loève expansion of a random field. In a Karhunen–Loève expansion the  $U_j$  are independent random variables whose distribution is determined by the distribution of the random field, and the  $\lambda_j$  and  $\psi_j$  are the eigenvalues and eigenvectors of the covariance operator, see, e.g., [143, Section 7.4].

In the remainder of this section we will be using QMC methods to approximate  $\mathbb{E}[Q(u)]$  (for some quantities of interest  $Q$ ). Observe that this expectation can be written

$$\mathbb{E}[Q(u)] = \int_{\Omega} Q(u(\omega)) d\mathbb{P}(\omega) = \int_{[-\frac{1}{2}, \frac{1}{2}]^{10}} Q(u(U_1, \dots, U_{10})) dU_1 \cdots dU_{10},$$

where we consider  $n$  (and therefore  $u$ ) as depending on each of the Uniform random variables  $U_j$  individually. Therefore, because of this correspondence between  $n$  as function on  $\Omega$ , and  $n$  as a function on  $[-1/2, 1/2]^{10}$  we will sometimes instead write  $n(\mathbf{y})$  for  $\mathbf{y} \in [-1/2, 1/2]^{10}$ , by which

<sup>1</sup>The highest ‘frequency’ associated with the oscillations in the medium is  $(10+1)\pi/4 \approx 26$ , whereas we consider waves with frequencies  $k = 10, \dots, 60$ . Therefore (for  $k > 26$ ) the waves are of a ‘higher frequency’ than the medium. Moreover, we would see if there is any change in the behaviour of our algorithm as the frequency of the waves increases past the ‘frequency’ of the medium. However, we do not see any such change.

we mean

$$n(\mathbf{y}) = 1 + \sum_{j=1}^{10} \mathbf{y}_j \sqrt{\lambda_j} \psi_j.$$

There is no a priori reason that one must have such an affine dependence of the random field on the randomness in order to apply nearby preconditioning to QMC methods. One could, for example, take  $n$  to be a lognormal random field, in which case  $n$  would take the form  $n(\mathbf{y}) = \exp(n_0 + \sum_j N_j \sqrt{\lambda_j} \psi_j)$  where the  $N_j$  are Normal(0, 1) random variables. However, in the case of affine dependence there is a ‘parallelisable’ nearby-preconditioning-QMC algorithm which we present below.

We stress that the results in this section are strictly numerical; there is no current theory to support these calculations. In particular, we observe in Section 4.6.3 below that in these experiments, for the QMC error for Helmholtz problems to remain bounded as  $k$  increases, one must increase the number of QMC points with  $k$ . We again remark that there is currently no theoretical justification for this behaviour.

*Terminology* Before we describe the nearby-preconditioning-QMC algorithms in detail, we establish two pieces of terminology that will be of use in describing them. Firstly, we will use the word ‘point’ to refer to a point in the parameter space  $[-1/2, 1/2]^J$ , and use phrases such as ‘calculate a preconditioner at the point  $\mathbf{y}$ ’ as shorthand for ‘calculate the LU decomposition of the system matrix  $A$  corresponding to the finite-element discretisation of the Helmholtz IIP (as described above) with coefficient  $n(\mathbf{y})$ ’.

We also use the words ‘nearby’ and ‘nearest’ (when referring to QMC points) to mean: nearest in the metric

$$d_{\text{approx}}(\mathbf{y}_1, \mathbf{y}_2) := \sum_{j=1}^J \sqrt{\lambda_j} |y_{1j} - y_{2j}|. \quad (4.102)$$

*The approximate metric* The metric  $d_{\text{approx}}$  is an approximation of the metric

$$d_{\text{QMC}}(\mathbf{y}_1, \mathbf{y}_2) = \|n(\mathbf{y}_1) - n(\mathbf{y}_2)\|_{L^\infty(D; \mathbb{R})}, \quad (4.103)$$

i.e., the metric on  $[-1/2, 1/2]^J$  induced by the spatial  $L^\infty$  norm. The metric  $d_{\text{approx}}$  is an approximation of  $d_{\text{QMC}}$  in the sense made precise in the following lemma.

**Lemma 4.37** ( $d_{\text{approx}}$  approximates  $d_{\text{QMC}}$ ). *For all  $\mathbf{y}_1, \mathbf{y}_2 \in [-1/2, 1/2]^J$ ,*

$$d_{\text{QMC}}(\mathbf{y}_1, \mathbf{y}_2) \leq d_{\text{approx}}(\mathbf{y}_1, \mathbf{y}_2).$$

The proof of Lemma 4.37 is straightforward and omitted.

Observe further that the structure of  $d_{\text{approx}}$  is similar to that of  $d_{\text{QMC}}$  and  $d_{\text{approx}}$  is a weighted  $L^1$  metric on  $[-1/2, 1/2]^J$ , with the weights corresponding to the terms in (4.99). Recall that  $\sqrt{\lambda_j} \rightarrow 0$  as  $j \rightarrow \infty$ ; therefore the higher dimensions contribute less to the value of  $d_{\text{approx}}$  (or, informally, points are ‘closer’ in higher dimensions, or higher dimensions are ‘smaller’ than lower

dimensions).

Ideally, for the purposes of utilising nearby preconditioning, we would use the metric  $d_{\text{QMC}}$  when describing the geometry of the QMC points (and computing the nearest QMC point), since the best rigorous results on the behaviour of nearby preconditioning (in terms of their  $k$ -dependence) are proved in Section 4.4 for the spatial  $L^\infty$ -norm<sup>2</sup>. However, computing  $d_{\text{QMC}}$  exactly is, in principle, complicated. In contrast, it is easy to compute with  $d_{\text{approx}}$ , since  $d_{\text{approx}}$  enables one to think of  $[-1/2, 1/2]^J$  as the high-dimensional rectangle  $[0, \sqrt{\lambda_1}] \times \cdots \times [0, \sqrt{\lambda_J}]$  equipped with the standard  $L^1$  metric. Moreover, as discussed above,  $d_{\text{approx}}$  is an approximation of  $d_{\text{QMC}}$ , and therefore we expect that it will induce a similar geometry on  $[-1/2, 1/2]^J$ .

*Computational complexity of calculating the nearest point* At various points in the two nearby-preconditioning-QMC algorithms we present below, given a point  $\mathbf{y} \in [-1/2, 1/2]^J$  and a subset  $S$  of  $[-1/2, 1/2]^J$  we must calculate  $\text{nearest}(\mathbf{y}, S) \in [-1/2, 1/2]^J$ , that is the element of  $S$  that is closest to  $\mathbf{y}$  in the metric  $d_{\text{approx}}$ . In all of the numerical results we present below, we calculate  $\text{nearest}(\mathbf{y}, S)$  by brute force, i.e., we calculate  $d_{\text{approx}}(\mathbf{y}, \tilde{\mathbf{y}})$  for all  $\tilde{\mathbf{y}} \in S$ , and choose the element of  $S$  that minimises  $d_{\text{approx}}(\mathbf{y}, \tilde{\mathbf{y}})$ . Since calculating  $d_{\text{approx}}(\mathbf{y}, \tilde{\mathbf{y}})$  involves  $\mathcal{O}(J)$  operations, the brute force approach to calculating  $d_{\text{approx}}(\mathbf{y}, \tilde{\mathbf{y}})$  involves  $\mathcal{O}(J|S|)$  operations. Clearly, this method of calculating  $\text{nearest}(\mathbf{y}, S)$  does not scale in  $J$ , the stochastic dimension, although it is computationally feasible for our numerical experiments (with  $J = 10$ ) below. See Section 4.9.2 below for a suggestion of an alternative, scalable way to calculate  $\text{nearest}(\mathbf{y}, S)$ .

### A sequential algorithm

We first describe a straightforward algorithm that uses nearby preconditioning to speed up a QMC calculation. We call this a ‘sequential’ algorithm because, unlike the ‘parallel’ algorithm that we describe below, it is intrinsically sequential and cannot be parallelised, i.e., finite-element solves for different realisations of the random field  $n$  cannot be treated in parallel. Although, when performing the individual finite-element solves, one is not restricted to a single core, i.e., one can use parallelisation for each finite-element solve if the linear systems  $A$  are large enough to warrant this.

An overview of the algorithm is:

1. Choose a QMC point  $\mathbf{y}$  for which to calculate a preconditioner
2. Find the nearest QMC point  $\mathbf{y}'$  to  $\mathbf{y}$  and attempt a GMRES solve of the problem at  $\mathbf{y}'$  using the LU decomposition of the system at  $\mathbf{y}$  as a preconditioner.
3. If GMRES converges quickly (i.e., in fewer than a preset number of iterations), return to Step 2.
4. If GMRES takes too long to converge, recalculate the preconditioner at  $\mathbf{y}'$ , set  $\mathbf{y} = \mathbf{y}'$ , and return to Step 2.

The algorithm is written in more formal pseudocode in Algorithm 4.1.

<sup>2</sup>Although, in line with the results in Section 4.5, we could instead use a spatial  $L^q$  norm, for some  $q \geq 1$  in (4.103).

```

Input:  $Its_{\max}$ ,  $S_{\text{QMC}}$ 
Choose starting point  $\mathbf{y}^{(\text{start})}$ 
 $\mathbf{y}^{(\text{pre})} \leftarrow \mathbf{y}^{(\text{start})}$ 
 $S_{\text{remaining}} \leftarrow S_{\text{QMC}} \setminus \{\mathbf{y}^{(\text{pre})}\}$ 
Calculate and store preconditioner  $\text{LU} = (\mathbf{A}^{\text{pre}})^{-1}$ 
 $\mathbf{y}^{(\text{current})} \leftarrow \text{nearest}(\mathbf{y}^{(\text{pre})}, S_{\text{remaining}})$ 
while  $S_{\text{remaining}} \neq \emptyset$  do
  if GMRES applied to  $(\mathbf{U})^{-1}(\mathbf{L})^{-1}\mathbf{A}\mathbf{y}^{(\text{current})} = (\mathbf{U})^{-1}(\mathbf{L})^{-1}\mathbf{f}$  converges in fewer than  $Its_{\max}$ 
  iterations then
     $S_{\text{remaining}} \leftarrow S_{\text{remaining}} \setminus \{\mathbf{y}^{(\text{current})}\}$ 
     $\mathbf{y}^{(\text{current})} \leftarrow \text{nearest}(\mathbf{y}^{(\text{pre})}, S_{\text{remaining}})$ 
  else
     $\mathbf{y}^{(\text{pre})} \leftarrow \mathbf{y}^{(\text{current})}$ 
    Calculate and store preconditioner  $\text{LU} = (\mathbf{A}^{\text{pre}})^{-1}$ 
  end
end

```

**Algorithm 4.1:** The sequential nearby-preconditioning-Quasi-Monte-Carlo algorithm.  $Its_{\max}$  is the maximum allowed number of GMRES iterations and  $S_{\text{QMC}}$  is the set of all QMC points.  $\text{nearest}(\mathbf{y}^{(\text{pre})}, S_{\text{remaining}})$  denotes the point in  $S_{\text{QMC}}$  nearest to  $\mathbf{y}^{(\text{pre})}$  in the  $d_{\text{QMC}}$  metric.

### A parallel algorithm

The main disadvantage of the ‘sequential’ algorithm described above is that the points at which preconditioners are calculated are identified as the algorithm progresses. The algorithm cannot be parallelised by sending different collections of QMC points to different processors (as one does not know a priori which preconditioner to use for each QMC point). Therefore, we now suggest an alternative algorithm that allows one to specify the number of preconditioning points *before* the algorithm begins. The algorithm then calculates which points to use as preconditioning points, before performing the linear solves. Because the preconditioners are known in advance, the solves can be computed in parallel if required. The most complicated part of the algorithm is deciding at which points to calculate the preconditioners, and so we describe this part of the algorithm in more detail here. A more formal pseudocode description of the algorithm is given in Algorithm 4.2.

Suppose we are given a set  $S_{\text{QMC}} = \{\mathbf{y}_1, \dots, \mathbf{y}_{N_{\text{QMC}}}\}$  of QMC points and a number  $N_{\text{pre,target}}$ ; the target number of preconditioners to compute. The aim of this algorithm is to select (approximately)  $N_{\text{pre,target}}$  QMC points that are (approximately) equally spaced with respect to the  $d_{\text{QMC}}$  metric defined above. If such a goal is achieved, then one expects that the preconditioning points are best located to minimise the total number of GMRES iterations across the solves for all of the QMC points.

The algorithm contains two key ideas:

1. Use an approximate metric in place of  $d_{\text{QMC}}$ , and

2. Locate the preconditioning points according to a tensor-product rule.

We now describe each of these two ideas in turn, before describing our final algorithm.

*Approximate metric* Whilst the metric  $d_{\text{QMC}}$  is the metric in which nearby preconditioning is analysed (as described in Section 4.1 above), in practice  $d_{\text{QMC}}$  is difficult to work with. This difficulty is because the geometry  $d_{\text{QMC}}$  induces on  $[-1/2, 1/2]^J$  is nontrivial, since the geometry is dependent on the interaction between the functions  $\psi_j$  in the expansion (4.99). Therefore, we use the approximate metric  $d_{\text{approx}}$ , defined by (4.102) above.

*Tensor-product algorithm for locating preconditioning points* We first describe the intuition behind our use of a tensor-product rule to locate the preconditioning points (even though we do not use this intuition in the final algorithm). Once we have described this intuition, we will then show how it can be adapted to provide the final algorithm. To understand why we use locate the preconditioning points using a tensor-product rule, we first describe the heuristic we use. Let us assume we want to cover  $[-1/2, 1/2]^J$  with ‘balls’ of radius  $r$ . Observe that these balls are measured in the  $d_{\text{approx}}$  metric, and therefore have a similar geometry to balls on  $[-1/2, 1/2]^J$  in the  $L^1$  metric. Therefore, given the centres  $\mathbf{c}_1$  and  $\mathbf{c}_2$  of two adjacent balls, we will have

$$d_{\text{approx}}(\mathbf{c}_1, \mathbf{c}_2) = 2r. \quad (4.104)$$

The question now arises of how we choose  $\mathbf{c}_1$  and  $\mathbf{c}_2$  so that (4.104) holds. We observe that, by the definition of  $d_{\text{approx}}$ , if we choose  $\mathbf{c}_1$  and  $\mathbf{c}_2$  such that

$$\sqrt{\lambda_j} |\mathbf{c}_{1j} - \mathbf{c}_{2j}| = \frac{2r}{J} \text{ for all } j = 1, \dots, J,$$

then we will have (4.104) by construction, because

$$d_{\text{approx}}(\mathbf{c}_1, \mathbf{c}_2) = \sum_{j=1}^J \frac{2r}{J} = 2r.$$

Therefore, in dimension  $j$  we choose the centres of the balls to be spaced

$$\min \left\{ \frac{2r}{J\sqrt{\lambda_j}}, 1 \right\}$$

apart (where we include the minimum so that, for high dimensions, we include at least one centre). That is, in dimension  $j$ , we take

$$N_j := \max \left\{ 1, \frac{J\sqrt{\lambda_j}}{2r} \right\} \quad (4.105)$$

equally spaced points in the sets  $\mathcal{C}_j = \{c_{j,1}, \dots, c_{j,N_j}\}$ , and then we form the centres  $\mathbf{c}_1, \dots, \mathbf{c}_{N_{\text{pre}}}$  by taking all possible tensor products of the points in  $\mathcal{C}_1, \dots, \mathcal{C}_J$ , giving a total of

$$N_{\text{pre}} = N_1 \times \dots \times N_J \quad (4.106)$$

preconditioning points.

However, we face three immediate difficulties with the above approach:

1. The above procedure assumes we know the radius  $r$ , and then returns the total number of preconditioning points, and their locations. However, we only know in advance the ideal total number of preconditioning points.
2. There is no guarantee that the numbers of points  $N_j$  calculated above are integers.
3. There is no guarantee the preconditioning points given by the above procedure are QMC points.

These questions are all completely valid, and so we slightly modify the above procedure to deal with them.

*Definition of the parallel algorithm* Recall that we assume that we are given a target number of preconditioners  $N_{\text{pre,target}}$ . The above procedure (amongst other things) defines a map  $N_{\text{pre,ideal}} : \mathbb{R}^+ \rightarrow \mathbb{R}^+$  given by  $r \mapsto N_{\text{pre}}$ , where  $N_{\text{pre}}$  is defined by (4.106) and the number of preconditioners in each dimension is given by (4.105). Therefore we can numerically invert this map (or more precisely, calculate numerically the value  $r_{\text{ideal}}$  such that  $N_{\text{pre,ideal}}(r_{\text{ideal}}) = N_{\text{pre,target}}$ ). (In our computations, we do this calculation via interval bisection.)

Given we expect that the size of the balls over which nearby preconditioning is effective decreases with  $\mathcal{O}(1/k)$  (in line with Theorem 4.11), and the number of QMC points needed to keep the error bounded increases with  $k$  (see Section 4.6.3 below), it is not obvious that we should know  $N_{\text{pre,ideal}}$  in advance. See page 184 for how we use the sequential algorithm to determine how  $N_{\text{pre,ideal}}$  scales with  $k$  for the parallel algorithm. We assume for now that we know  $N_{\text{pre,ideal}}$  and hence  $r_{\text{ideal}}$ .

Once we know the value of  $r_{\text{ideal}}$ , we can then calculate the numbers of centres in each dimension  $N_1(r_{\text{ideal}}), \dots, N_J(r_{\text{ideal}})$  as above (recalling that the  $N_j(r_{\text{ideal}})$  are not necessarily integers). We then obtain integers  $N_{\text{pre,actual},j} = \text{round}(N_j(r_{\text{ideal}}))$ , where  $\text{round}(\cdot)$  denotes rounding to the nearest integer. (Recall  $N_j(r_{\text{ideal}}) \geq 1$  for all  $j$  by construction, so  $N_{\text{pre,actual},j}$  will be a positive integer for all  $j$ .)

We then use  $N_{\text{pre,actual},j}$  centres in each dimension and define the sets  $\mathcal{C}_j$  as described above, with  $N_j = N_{\text{pre,actual},j}$ . We then obtain a total of  $N_{\text{pre,actual}} = N_{\text{pre,actual},1} \times \dots \times N_{\text{pre,actual},J}$  preconditioning points.

These points may not be QMC points. We could simply calculate the preconditioners at these non-QMC points. However we instead replace each calculated centre with its nearest QMC point (calculated using brute-force) and calculate the preconditioners at these QMC points. We denote

the set of preconditioning points by  $S_{\text{pre}}$ . Finally, we calculate the map  $\text{Pre}_{\text{nearest}} : S_{\text{QMC}} \rightarrow S_{\text{pre}}$ , i.e., for each QMC point we find its nearest preconditioning point, and use the corresponding preconditioner for the linear solve.

This algorithm is summarised more formally in Algorithm 4.2.

**Remark 4.38** (Is calculating  $\text{Pre}_{\text{nearest}}$  computationally expensive?). *We note that calculating the map  $\text{Pre}_{\text{nearest}} : S_{\text{QMC}} \rightarrow S_{\text{pre}}$  is an  $\mathcal{O}(N_{\text{QMC}}N_{\text{pre}})$  operation, because for each QMC point we must find the nearest preconditioning point. Given that  $S_{\text{pre}} \subseteq S_{\text{QMC}}$ , it is possible that calculating  $\text{Pre}_{\text{nearest}}$  could actually be an  $\mathcal{O}(N_{\text{QMC}}^2)$  operation.*

*However, we expect that  $N_{\text{pre}}$  will be small relative to  $N_{\text{QMC}}$  (and this is borne out in the numerical experiments summarised in Table 4.5 below) and therefore we expect  $\mathcal{O}(N_{\text{QMC}}N_{\text{pre}}) \approx \mathcal{O}(N_{\text{QMC}})$ . Hence calculating  $\text{Pre}_{\text{nearest}}$  should not be an expensive computational task.*

*A similar line of reasoning shows that calculating the nearest QMC point to each of calculated tensor-product points (as outlined above) should also be an  $\mathcal{O}(N_{\text{QMC}})$  task.*

**Input** :  $N_{\text{pre,target}} \in \mathbb{N}$

**Output** : The set  $S_{\text{pre}}$ , the map  $\text{Pre}_{\text{nearest}} : S_{\text{QMC}} \rightarrow S_{\text{pre}}$

Solve (numerically)  $N_{\text{pre,ideal}}(r_{\text{ideal}}) = N_{\text{pre,target}}$  for  $r_{\text{ideal}}$

**for**  $j = 1$  **to**  $J$  **do**

    Calculate  $N_{\text{pre,actual},j} = \text{round}(N_{\text{pre,ideal},j}(r_{\text{ideal}}))$

    Define  $S_{\text{pre},j}$  to be set of  $N_{\text{pre,actual},j}$  equally spaced points in  $[-1/2, 1/2]$

**end**

Define  $N_{\text{pre}} = \prod_{j=1}^J N_{\text{pre,actual},j}$

Define  $S_{\text{pre}}$  by taking all possible tensor products of points in  $S_{\text{pre},j}$ , and then finding the nearest QMC point to each one

**for**  $l = 1$  **to**  $N_{\text{QMC}}$  **do**

    Calculate  $\text{Pre}_{\text{nearest}}(\mathbf{y}^{(l)})$

**end**

**Algorithm 4.2:** The main part of the parallel nearby-preconditioning-Quasi-Monte-Carlo algorithm. This part of the algorithm determines  $S_{\text{pre}}$  and  $\text{Pre}_{\text{nearest}}$ .  $S_{\text{pre}}$  is the set of preconditioning points, and  $\text{Pre}_{\text{nearest}} : S_{\text{QMC}} \rightarrow S_{\text{pre}}$  maps each QMC points to its nearest preconditioning point.

## Comparing and Constrasting the two algorithms

We now briefly list the main differences in the two algorithms given above.

*Complexity* The sequential algorithm is simple and intuitive to describe, given that it mainly revolves around ‘finding the nearest point’. However, the parallel algorithm is much more complicated, both in the underlying ideas, but also in its technical definition.

*Heuristics* The sequential algorithm has very minimal heuristics; one only needs to specify the maximum number of GMRES iterations and this could be determined, for example, by the memory constraints of the machine one is using. In contrast, for the parallel algorithm one needs a heuristic for how many preconditioning points to choose, as this is not given by the algorithm. (In our numerical experiments below, we obtain this heuristic by using the sequential algorithm for low  $k$ , and then extrapolating the proportion of preconditioning points used for low values of  $k$  to larger values of  $k$ .)

*Parallelisability* Unsurprisingly (given the name) the sequential algorithm is inherently serial; one must see whether a given solve converges in the required number of GMRES iterations before knowing whether we must recalculate the preconditioner for subsequent solves. (In principle one could parallelise the algorithm by splitting the QMC points up onto different groups of processors, and then use the sequential algorithm on each group of processors. However, there is no guarantee one would split the QMC points up in a way that grouped nearby points, therefore this approach could lead to a substantial increase in computational work.) In contrast, the parallel algorithm is fully parallelisable; once the preconditioning points and the map  $\text{Pre}_{\text{nearest}} : S_{\text{QMC}} \rightarrow S_{\text{pre}}$  have been calculated, one can send different linear solves to different groups of processors as one chooses. (Although note that, unless one sends all of the QMC points corresponding to a single preconditioner to the *same* group of processors, one may need to calculate the same preconditioner several times, on different groups of processors<sup>3</sup> However, the decrease in computational time gained from parallelisation should more than offset this increase in computational effort.)

*Choice of preconditioning points* Neither algorithm will necessarily pick the optimal set of preconditioning points (optimal in the sense of the minimal number of preconditioning points needed). In the sequential algorithm, there is no guarantee that this method for exploring the sample space and choosing the preconditioning points will give an optimal collection of preconditioning points. Also, whilst for the parallel algorithm the preconditioning points should fill the parameter space ‘well’ (given the points are chosen a priori to be well spaced according to the  $d_{\text{approx}}$  metric), the number of preconditioning points generated is not exactly  $N_{\text{pre,target}}$  due to rounding the ‘ideal’ number of centres in each dimension to the nearest integer. Therefore, even in the parallel case, one may not end up with an optimal set of preconditioning points.

### 4.6.3 Numerical Experiments

We now describe numerical experiments that demonstrate the effectiveness of the above algorithms. Our main result is that, for a particular QMC model problem, nearby preconditioning gives a substantial speedup, with around 98% of solves being computed using a previously-calculated LU decomposition.

For the computational setup, including the algorithm we use to generate our QMC points, see Appendix G.

---

<sup>3</sup>In our code, we split up the points with respect to the order they are generated by the QMC code. This was purely to make the code simpler.

Before we perform our numerical experiments, we need to determine:

- How the number of QMC points should scale with  $k$ , and
- How many preconditioners we should choose.

Throughout this section we use the model problem detailed in (4.99)–(4.101) above.

### QMC error estimators

To determine how the number of QMC points should scale with  $k$ , we first estimate how the QMC error grows as  $k$  increases. The QMC rule we use is a randomly shifted QMC rule, we use such a rule because there exists an error estimator for this rule, see (4.107) below. Our exposition below follows that in [100, Section 4.2].

Suppose our QMC points are  $\mathbf{y}_1, \dots, \mathbf{y}_{N_{\text{QMC}}}$ , and the resulting QMC rule is

$$\mathcal{Q}_{N_{\text{QMC}}}(Q) = \frac{1}{N_{\text{QMC}}} \sum_{l=1}^{N_{\text{QMC}}} Q(u(\mathbf{y}_l)).$$

For a ‘shift’  $\mathbf{s} \in [-1/2, 1/2]^J$  we define the shifted QMC rule

$$\mathcal{Q}_{N_{\text{QMC}}, \mathbf{s}}(Q) = \frac{1}{N_{\text{QMC}}} \sum_{l=1}^{N_{\text{QMC}}} Q(u(\mathbf{y}_l \oplus \mathbf{s})),$$

where  $\mathbf{y} \oplus \mathbf{s}$  denotes  $\mathbf{y} + \mathbf{s}$  ‘wrapped around’ onto the hypercube  $[-1/2, 1/2]^J$ . (Formally  $\mathbf{y} \oplus \mathbf{s} = \text{frac}((\mathbf{y} + \frac{1}{2}) + \mathbf{s}) - \frac{1}{2}$ , where  $\text{frac}(\cdot)$  denotes the fractional part and  $\frac{1}{2}$  denotes the  $J$ -dimensional vector with every entry  $1/2$ .)

We can then define the randomly-shifted QMC rule (with multiple randomly-chosen shifts  $\mathbf{s}_1, \dots, \mathbf{s}_{N_{\text{shifts}}}$ )

$$\mathcal{Q}_{N_{\text{QMC}}, N_{\text{shifts}}}^{\text{rand}}(Q) = \frac{1}{N_{\text{shifts}}} \sum_{s=1}^{N_{\text{shifts}}} \mathcal{Q}_{N_{\text{QMC}}, \mathbf{s}_s}(Q) = \frac{1}{N_{\text{QMC}} N_{\text{shifts}}} \sum_{s=1}^{N_{\text{shifts}}} \sum_{l=1}^{N_{\text{QMC}}} Q(u(\mathbf{y}_l \oplus \mathbf{s}_s)).$$

Having defined the randomly shifted QMC rule, one can use the standard statistical estimator of the standard deviation of the statistical error in  $\mathcal{Q}_{N_{\text{QMC}}, N_{\text{shifts}}}^{\text{rand}}(Q)$  [100, Equation (4.6)]

$$\text{Err}_{\text{QMC}}(N_{\text{QMC}}, N_{\text{shifts}}) = \left( \frac{1}{N_{\text{shifts}}(N_{\text{shifts}} - 1)} \sum_{s=1}^{N_{\text{shifts}}} \left( \mathcal{Q}_{N_{\text{QMC}}, \mathbf{s}_s}(Q) - \mathcal{Q}_{N_{\text{QMC}}, N_{\text{shifts}}}^{\text{rand}}(Q) \right)^2 \right)^{\frac{1}{2}}. \quad (4.107)$$

(See Appendix D for proof that  $\text{Err}_{\text{QMC}}(N_{\text{QMC}}, N_{\text{shifts}})^2$  is an unbiased estimator of the variance of  $\mathcal{Q}_{N_{\text{QMC}}, N_{\text{shifts}}}^{\text{rand}}(Q)$ ; recall that it does *not* then follow that  $\text{Err}_{\text{QMC}}(N_{\text{QMC}}, N_{\text{shifts}})$  is an *unbiased* estimator of the standard deviation of  $\mathcal{Q}_{N_{\text{QMC}}, N_{\text{shifts}}}^{\text{rand}}(Q)$ .)

### $k$ -dependence of the number of QMC points

We first sought to determine how  $\text{Err}_{\text{QMC}}(N_{\text{QMC}}, N_{\text{shifts}})$  depends on  $k$ . We estimated the error  $\text{Err}_{\text{QMC}}(N_{\text{QMC}}, N_{\text{shifts}})$  for the setup described in Appendix G with  $N_{\text{QMC}} = 2048$  and  $N_{\text{shifts}} = 20$  (i.e., 40,960 PDE solves in total) for  $k = 10, 20, 30, 40, 50, 60$ . We set  $h = 0.002$  for all of the computations (as  $0.002 \approx 60^{-3/2}$ ), as then by Theorem 2.39 the finite-element error is of the order  $h^2 k^3 \sim (k/60)^3 \lesssim 1$  for all the values of  $k$  we consider<sup>4</sup>. The quantities of interest (QoIs) we considered were:

- The integral of  $u$  over the whole domain  $[0, 1]^2$ ,
- The value of  $u$  at the origin,
- The value of  $u$  at the top-right corner of the domain, and
- The  $x$ -component of  $\nabla u$  at the top-right corner of the domain.

Observe that these QoIs require a certain amount of regularity of the solution. (The integral is defined for functions in  $L^1(D)$ , point evaluation for functions in  $H^{3/2+\varepsilon}(D)$  and point evaluation of the gradient for functions in  $H^{5/2+\varepsilon}(D)$  (in 3-d - the corresponding function spaces are  $H^{1+\varepsilon}(D)$  and  $H^{2+\varepsilon}(D)$  in 2-d) for any  $\varepsilon > 0$ .) Therefore computing for this range of QoIs will give a good insight into the behaviour of QMC applied to the Helmholtz equation<sup>5</sup>.

Motivated by QMC theory for other applications, e.g., [100, Equation 4.2], we test experimentally the assumption that the QMC error satisfies

$$\text{Err}_{\text{QMC}}(Q, N_{\text{shifts}}) = C N_{\text{QMC}}^{-\alpha}, \quad (4.108)$$

for some  $C, \alpha > 0$ . Using data for the values of  $k$  listed above, Figures 4.10–4.13 plot the computed values of  $C$  and  $\alpha$  against  $k$ . (In Appendix E, we plot the QMC error for increasing  $N_{\text{QMC}}$  for each  $k \in \{10, 20, 30, 40, 50, 60\}$  and for each QoI—these plots allow us to determine the values of  $C$  and  $\alpha$  for each value of  $k$ .) For the QoIs that are point evaluations (Figures 4.11 and 4.12),  $C$  appears not to vary very much; thus we assume  $C$  is constant in all of the following calculations.

Figures 4.10–4.13 (bottom panes) show  $\alpha$  decreasing at a rate proportional to  $\log k$ . Therefore we conjecture

$$\alpha(k) = \alpha_0 - \alpha_1 \ln(k), \quad (4.109)$$

for some constants  $\alpha_0, \alpha_1 > 0$ . (Throughout this section,  $\ln$  denotes the natural logarithm.) We fitted  $\alpha_0$  and  $\alpha_1$  numerically, and have plotted the resulting line of best fit on Figures 4.10–4.13. (Observe that the conjectured form (4.109) cannot hold for  $k$  very large, as then  $\alpha(k)$  would be negative, and there would be no convergence as the number of QMC points is increased.

<sup>4</sup>Observe that we do not let  $h$  depend on  $k$ , in contrast to the rest of this thesis. This decision means we do not have to consider the effect of changing the mesh on the resulting interpolation of the random field  $n$ , and how this interpolation may affect the overall error. In addition, since  $k \leq 60$ , our particular choice of mesh ensures that the finite-element error is small for all the values of  $k$  we consider.

<sup>5</sup>We can evaluate point values of  $u_h$  because  $u_h$  is continuous, and we use the constant value of  $\nabla u_h$  on the upper-rightmost mesh element as a proxy for  $\nabla u_h((1, 1))$ ; such a use is possible due to the structure of our mesh, see Figure G.1, and the fact that we use first-order finite elements.

Nevertheless, for the range of  $k$  we consider in these numerical experiments, the form (4.109) seems to give a good fit with the data.) The values of  $C$  and  $\alpha$  for the different QoIs are given in Figures 4.10–4.13.

Having understood how the QMC error increases with  $k$  for fixed  $N_{\text{QMC}}$ , we now use this knowledge to determine how one should increase  $N_{\text{QMC}}$  with  $k$  in order to keep the QMC error bounded. Recalling that we assume  $C$  in (4.108) is constant, if we take

$$N_{\text{QMC}}(k) = \exp(\tilde{C} \alpha(k)^{-1}), \quad (4.110)$$

for some constant  $\tilde{C} > 0$ , then substituting (4.110) into (4.108), we see that the QMC error should remain bounded, with

$$\text{Err}_{\text{QMC}}(Q, N_{\text{shifts}}) = C \exp(-\tilde{C}).$$

Observe that, since  $\alpha(k)$  decreases as  $k$  increases, (4.110) will increase as  $k$  increases.

In our numerical experiments with increasing  $N_{\text{QMC}}(k)$  below, we set  $\tilde{C}$  so that  $N_{\text{QMC}}(10) = 2048$ , because in our numerical experiments to determine the behaviour of the QMC error, we used  $N_{\text{QMC}} = 2048$  (with 20 shifts). Also in our numerical experiments below we take the number of QMC points to be a power of 2, because the lattice rule we use to generate the points is a complete lattice rule if  $N_{\text{QMC}}$  is a power of 2 (see [162]). We choose  $N_{\text{QMC}}$  to be a power of 2 by setting  $N_{\text{QMC}}(k) = 2^{M(k)}$ , where

$$M(k) = \text{round}(\log_2(\exp(\tilde{C} \alpha(k)^{-1}))).$$

Based on the results for the QoIs in Table 4.2 (excluding the results for the QoI being the integral of  $u$  and  $\nabla u((1, 1))$ ), as these seem to display slightly different convergence characteristics), in our numerical experiments below we take  $\alpha(k) = 1.38 - 0.19 \ln(k)$ . The resulting values of  $N_{\text{QMC}}$  are summarised in Table 4.3.

|            | $Q = \int_D u$ | $Q = u(\mathbf{0})$ | $Q = u((1, 1))$ | $Q = \nabla u((1, 1))$ |
|------------|----------------|---------------------|-----------------|------------------------|
| $\alpha_0$ | 1.34           | 1.38                | 1.51            | 1.51                   |
| $\alpha_1$ | 0.16           | 0.19                | 0.21            | 0.21                   |

Table 4.2: The value of  $\alpha_0$  and  $\alpha_1$  for different QoIs, where the QMC error  $\approx C N_{\text{QMC}}^{-(\alpha_0 - \alpha_1 \ln(k))}$ .

### Numerical results for nearby preconditioning applied to QMC

Now that we have an estimate of how the number of QMC points should scale with  $k$  in order to keep the QMC error bounded, we apply nearby preconditioning to QMC (with the number of points chosen as in Table 4.3) and observe how the computational work of this nearby-preconditioning-QMC (NP-QMC) algorithm scales with  $k$ .

As outlined above, we combine our sequential- and parallel-NPQMC algorithms:

| $k$ | $\exp(\tilde{C}\alpha(k)^{-1})$ | $N_{\text{QMC}}$ |
|-----|---------------------------------|------------------|
| 10  | $2^{11}$                        | $2^{11}$         |
| 20  | $2^{12.78}$                     | $2^{13}$         |
| 30  | $2^{14.12}$                     | $2^{14}$         |
| 40  | $2^{15.26}$                     | $2^{15}$         |
| 50  | $2^{16.28}$                     | $2^{16}$         |
| 60  | $2^{17.21}$                     | $2^{17}$         |

Table 4.3: The ideal and actual number of QMC points  $N_{\text{QMC}}$  used in the numerical experiments summarised in Tables 4.4 and 4.5, chosen so that the QMC error is empirically bounded for all  $k$ .

- We first use the sequential algorithm for low  $k$  (fixing the maximum number of GMRES iterations) and observe how the number of preconditioners (as a proportion of the number of QMC points) changes with  $k$ . We thus obtain an empirical relationship between  $k$  and the proportion of QMC points used to construct preconditioners.
- We then use the parallel algorithm (with the above proportion of preconditioners) for higher values of  $k$ .

We remark that, in principle, one could use the sequential algorithm for all values of  $k$ , however, this would take an incredibly long time— we see in Table 4.3 that for  $k = 60$  we must perform  $2^{17}$  Helmholtz solves; if we performed these solves sequentially, and each solve took 10 seconds, this computation would take over 2 weeks to complete.

The results for the sequential algorithm are summarised in Table 4.4, for  $k = 10, 20, 30$ . The results show that nearby preconditioning is effective, with the number of preconditioners growing (approximately) linearly in  $k$ , but at a very low percentage of the total number of solves. Also, observe that nearby preconditioning is much more effective than mean-based preconditioning, where we use a single preconditioner, corresponding to the mean of  $n$ , to precondition all the realisations.

Performing a linear fit for the percentage of LU-factorisations used in the nearby-preconditioning algorithm, we obtain that the percentage of LU-factorisations grows like  $-0.04 + 0.02k$  (see Figure 4.14). This result indicates that although the radius of the balls in which nearby preconditioning is effective decreases with  $\mathcal{O}(1/k)$ , the fact that the number of QMC points increases with  $k$  means that a large proportion of the solves are computed using a previously-calculated LU decomposition. Observe that if the number of QMC points remained constant in  $k$ , we would expect the number of preconditioners to (potentially) increase like  $k^J$ , because the number of balls of radius  $\sim 1/k$  in  $[-1/2, 1/2]^J$  is  $\sim k^J$ .

Based on these sequential results, we then used the parallel algorithm with a target proportion of preconditioners of  $(-0.04 + 0.02k)\%$ . (Although recall from our discussion above that the actual proportion of preconditioners used can vary due to rounding in the algorithm.) The results of

these computations are summarised in Table 4.5. We observe that the fraction of preconditioners is approximately  $-0.04 + 0.02k$ , but the maximum (and average) number of GMRES iterations appears to grow slowly with  $k$ . This growth (which did not occur with the sequential algorithm) may be because the placement of the preconditioning points is not optimal with respect to the  $d_{\text{QMC}}$  metric; we conjecture that oversampling the number of preconditioners needed (for example, taking a proportion of  $(0.05k)\%$ ) may result in a bounded number of GMRES iterations. Nevertheless, we see that nearby preconditioning gives considerable speedup, drastically reducing the number of preconditioners that must be calculated.

In conclusion, we see that nearby preconditioning gives a significant speedup when applied to a QMC model problem.

## 4.7 REVIEW OF RELATED TECHNIQUES IN THE LITERATURE

Having proved rigorous results on the effectiveness of nearby preconditioning, and also applied it to a UQ algorithm, we now review similar computational techniques (applied to other problems) which can be found in the literature. Whilst the idea of *nearby* preconditioning introduced here is, as far as we are aware, novel, there has been a body of work on the closely-related idea of *mean-based* preconditioning. In mean-based preconditioning a *single* preconditioner is calculated corresponding to the mean of the random coefficient. This is in contrast to nearby preconditioning, where *multiple* preconditioners are calculated, corresponding to each realisation in a particular subset of all the realisations. Mean-based preconditioning has been most extensively studied for the stationary diffusion equation

$$\nabla \cdot (x \nabla u) = -f,$$

with a small number of works analysing other PDEs, including two works on the Helmholtz equation. We will first explain the idea of mean-based preconditioning before we review the literature applying it to the stationary diffusion equation and other PDEs, and finally turning our attention to mean-based preconditioning for the Helmholtz equation. In general, the computational and mathematical results in the literature show that mean-based preconditioning is effective if the variance of the random parameters is small enough, i.e., if most of the samples are sufficiently close to the mean.

Mean-based preconditioning was first developed for the stationary diffusion equation in the context of so-called Stochastic Spectral Finite-Element Methods (SSFEMs). In these methods, the random field  $a$  is given by a series expansion, such as a Karhunen–Loève expansion, and the dependence of  $u$  on the random parameters is computed using a Polynomial Chaos expansion (see, e.g., [91, Section 2.4.2]). The resulting problem is then discretised in the whole space  $D \times \Omega$ , where  $D$  is the spatial domain and  $\Omega$  the probability space. The resulting discrete problems involve very large matrices of the form

$$A \otimes G, \tag{4.111}$$

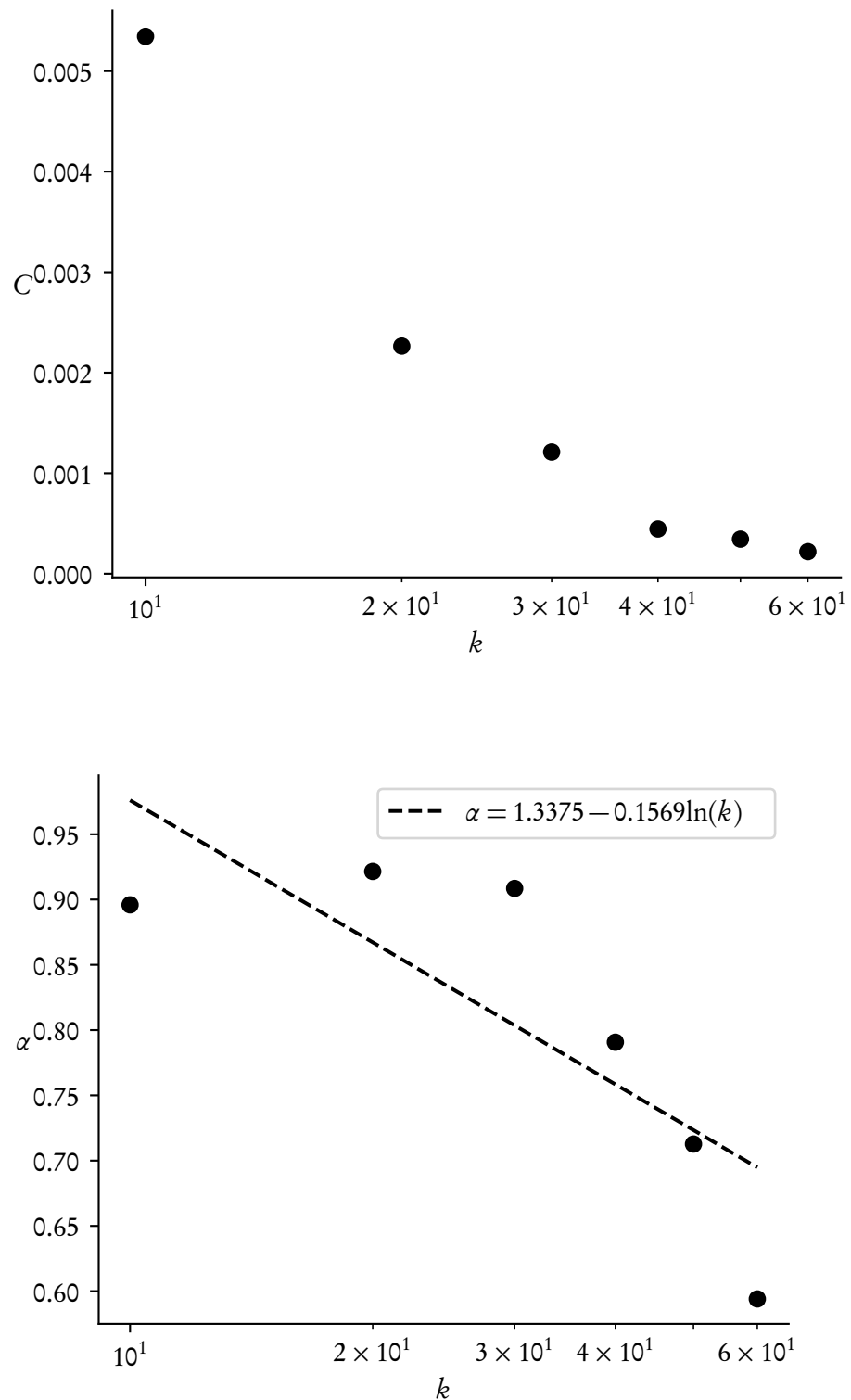


Figure 4.10: Plots of the computed values of  $C$  (top) and  $\alpha$  (bottom) against  $k$  in (4.108) for  $Q(u) = \int_D u$ . Observe the  $x$ -axes are on a  $\log_{10}$  scale, but  $\ln$  is the natural logarithm.

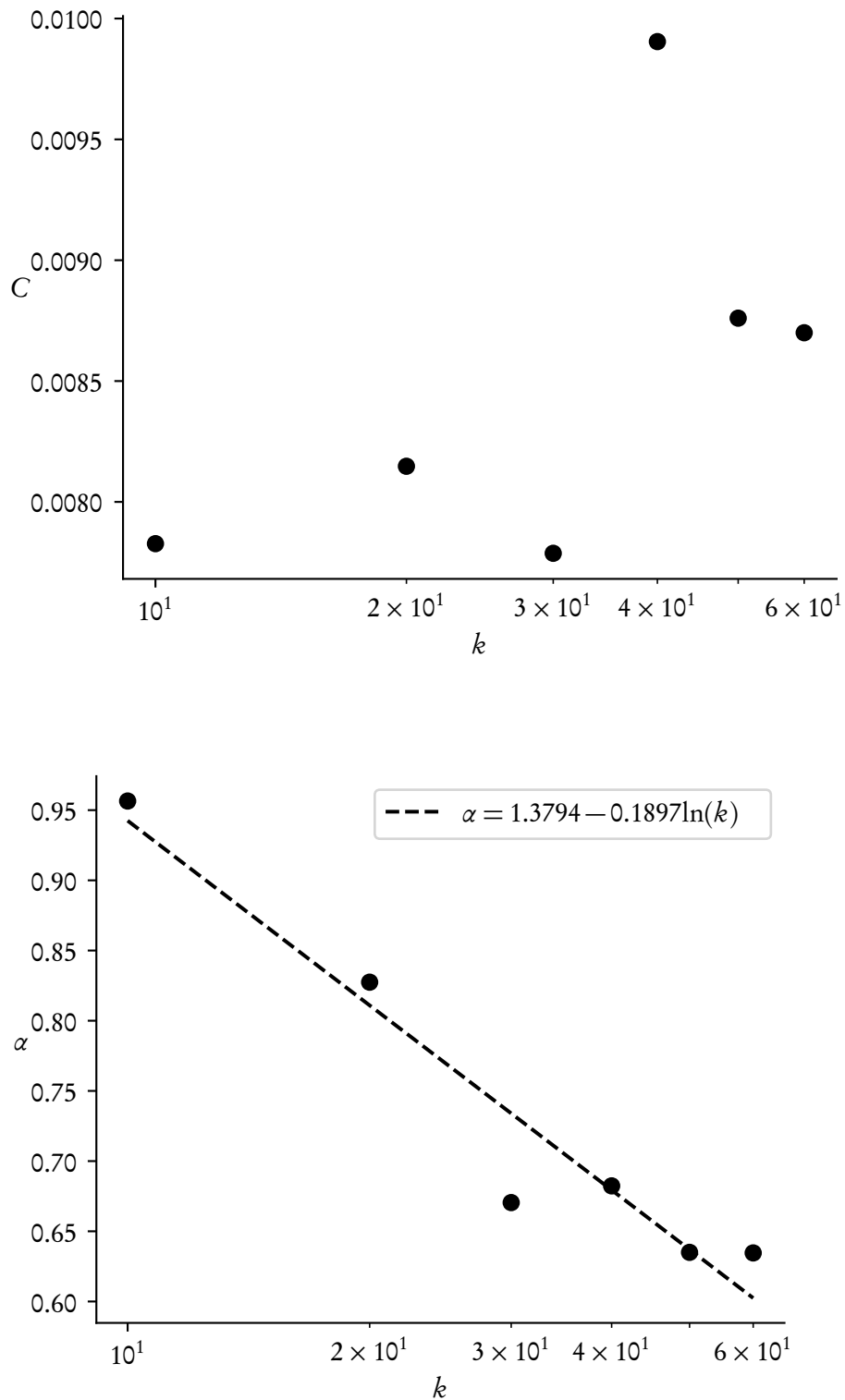


Figure 4.11: The computed values of  $C$  (top) and  $\alpha$  (bottom) against  $k$  in (4.108) for  $Q(u) = u(0)$ . Observe the  $x$ -axes are on a  $\log_{10}$  scale, but  $\ln$  is the natural logarithm.

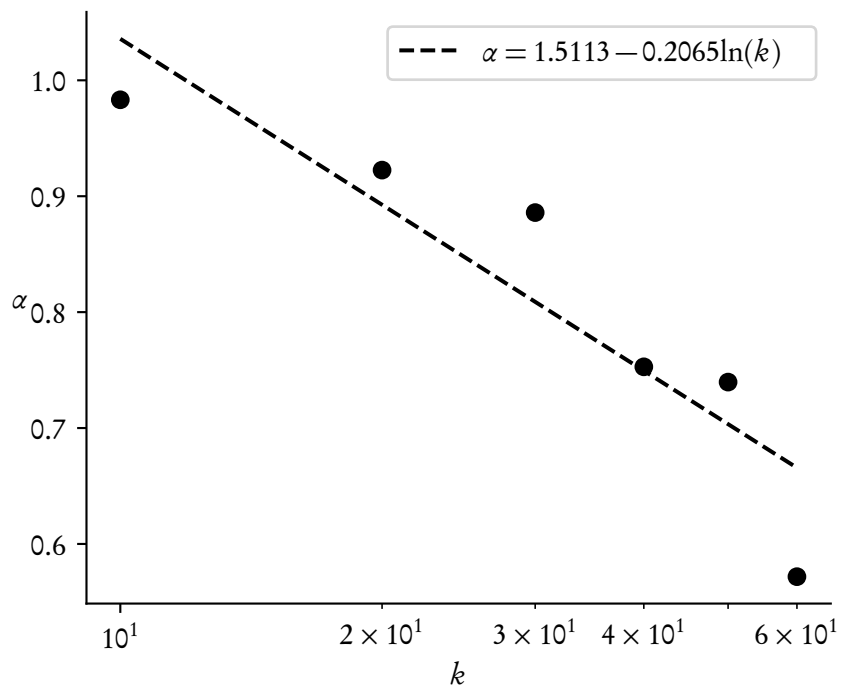
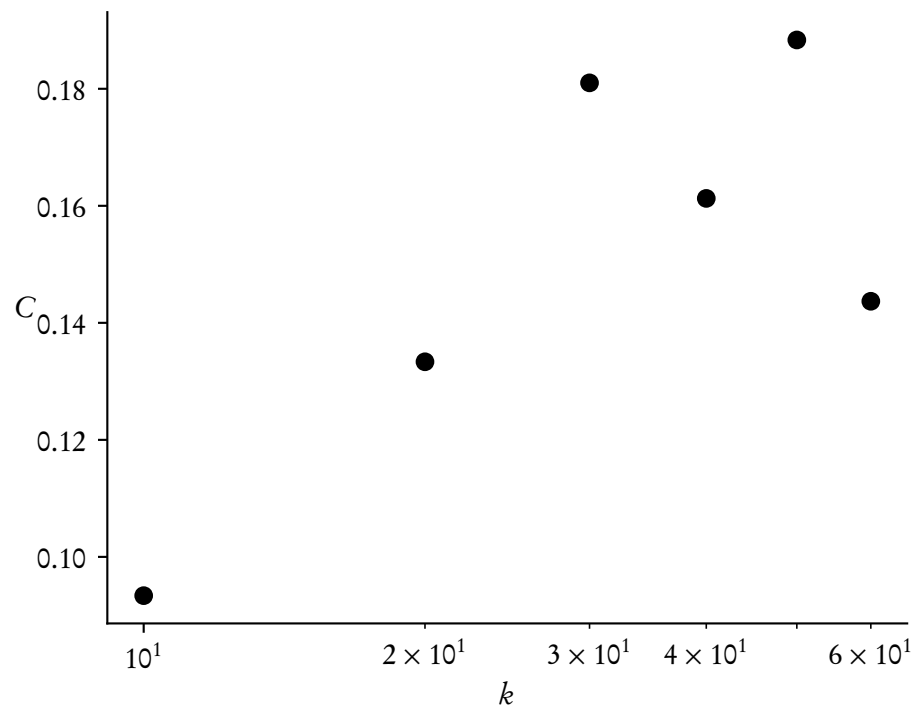


Figure 4.12: The computed values of  $C$  (top) and  $\alpha$  (bottom) against  $k$  in (4.108) for  $Q(u) = u((1, 1))$ . Observe the  $x$ -axes are on a  $\log_{10}$  scale, but  $\ln$  is the natural logarithm.

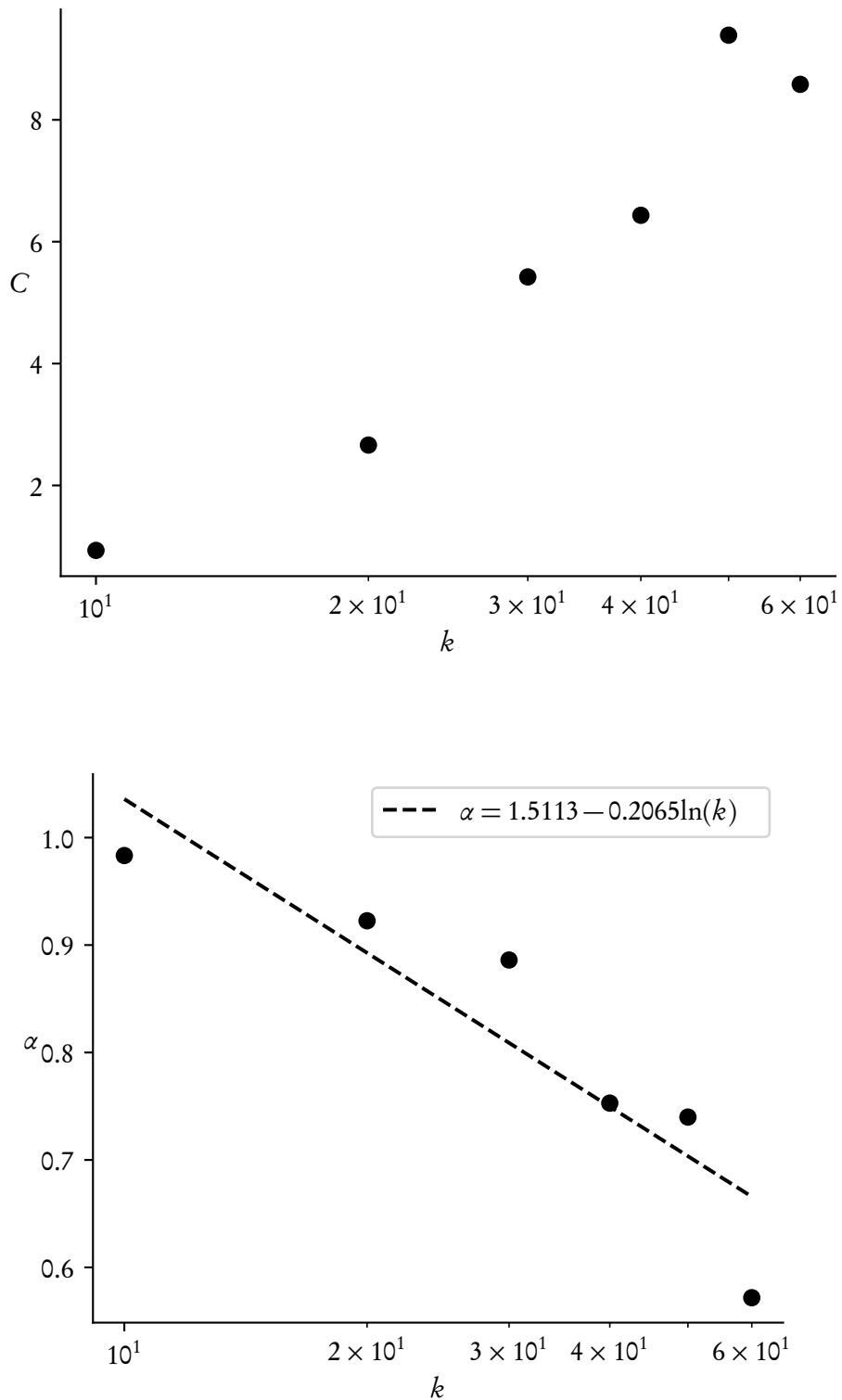


Figure 4.13: The computed values of  $C$  (top) and  $\alpha$  (bottom) against  $k$  in (4.108) for  $Q(u) = \nabla u((1, 1))$ . Observe the  $x$ -axes are on a  $\log_{10}$  scale, but  $\ln$  is the natural logarithm.

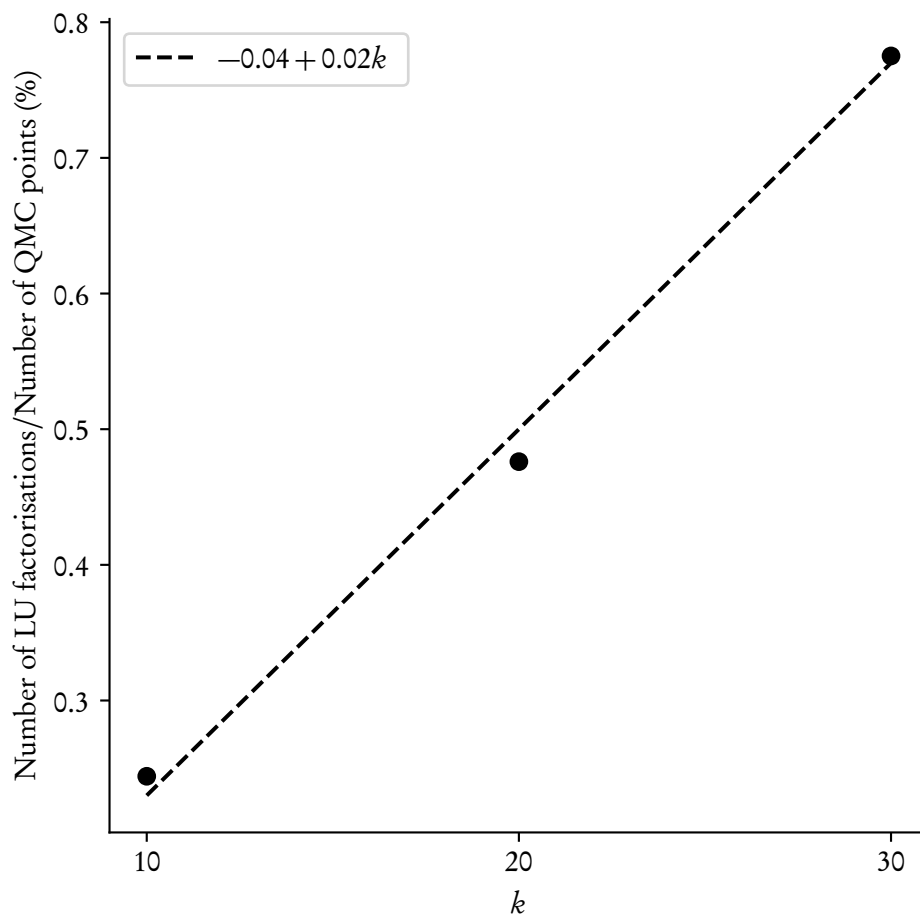


Figure 4.14: The number of LU factorisations in the sequential algorithm as a percentage of the total number of solves.

| $k$ | # LU factorisations | Total # linear systems | # LU factorisations/<br># linear systems (%) | Average # GMRES iterations | Max. # GMRES iterations | Average # GMRES iterations using mean-based preconditioning | Max. # GMRES iterations using mean-based preconditioning |
|-----|---------------------|------------------------|----------------------------------------------|----------------------------|-------------------------|-------------------------------------------------------------|----------------------------------------------------------|
| 10  | 5                   | 2048                   | 0.24                                         | 7.13                       | 10                      | 9.47                                                        | 15                                                       |
| 20  | 39                  | 8192                   | 0.48                                         | 7.26                       | 10                      | 14.74                                                       | 40                                                       |
| 30  | 127                 | 16384                  | 0.78                                         | 7.47                       | 10                      | 21.00                                                       | 80                                                       |

Table 4.4: Results applying our sequential nearby-preconditioning-Quasi-Monte-Carlo algorithm with the maximum number of GMRES iterations = 10, alongside results for mean-based preconditioning.

| $k$ | # LU factorisations | Total # linear systems | # LU factorisations/<br># linear systems (%) | Average # GMRES iterations | Max. # GMRES iterations |
|-----|---------------------|------------------------|----------------------------------------------|----------------------------|-------------------------|
| 10  | 4                   | 2048                   | <b>0.20</b>                                  | 6.46                       | 10                      |
| 20  | 33                  | 8192                   | <b>0.40</b>                                  | 6.42                       | 11                      |
| 30  | 127                 | 16384                  | <b>0.78</b>                                  | 6.66                       | 13                      |
| 40  | 207                 | 32768                  | <b>0.63</b>                                  | 7.16                       | 15                      |
| 50  | 1027                | 65536                  | <b>1.57</b>                                  | 7.07                       | 14                      |
| 60  | 1444                | 131072                 | <b>1.10</b>                                  | 7.41                       | 16                      |

Table 4.5: Results applying our parallel nearby-preconditioning-Quasi-Monte-Carlo algorithm with the target proportion of preconditioners as  $(-0.04 + 0.02k)\%$ .

where  $A$  is a standard finite-element matrix,  $G$  is a matrix corresponding to the discretisation in  $\Omega$ , and  $\otimes$  is the Kronecker product. For SSFEMs (and the closely-related stochastic-Galerkin FEMs, see, e.g. [8], which also have discretisations of the form (4.111)) a mean-based preconditioner is a matrix of the form

$$A^{(\text{mean})} \otimes I_{\Omega}, \quad (4.112)$$

where  $A^{(\text{mean})}$  is the standard finite-element matrix corresponding to the mean of  $\chi$  and  $I_{\Omega}$  is the identity matrix associated with the discretisation on  $\Omega$ . Using a mean-based preconditioner of the form (4.112) gives considerable computational savings, as only one preconditioner of a standard finite-element matrix needs to be calculated.

When stochastic Galerkin methods are used with so-called ‘doubly-orthogonal bases’ (see, e.g., [70, Section 3.2]), then the linear system (4.111) decouples into many distinct standard finite-element matrices; mean-based preconditioning has also been investigated in this context (and in the context of stochastic collocation methods (see, e.g., [7]), where one similarly obtains many different standard finite-element matrices) as will be discussed below.

The main insight gleaned from studies of mean-based preconditioning is that, as stated above, if the variance of  $\chi$  (or any other stochastic coefficients) is sufficiently small, then mean-based preconditioning is effective.

The initial computational work on mean-based preconditioning for the stationary diffusion equation was carried out by Ghanem and Kruger [90], Pellissetti and Ghanem [167], and Keese [127], with theory (proving bounds on the eigenvalues of the preconditioned matrices) following from Powell and Elman [180] and Ernst, Powell, Silvester, and Ullmann [70]. These eigenvalue bounds are analogous to results in Section 4.2 above, as they allow one to infer convergence properties of the iterative method used. All of the above results were for  $\chi$  given by a (real or artificial) Karhunen–Loève expansion; that is, in the case where  $\chi$  depends linearly on the random parameter. In the case where  $\chi$  is a lognormal random field (and so the dependence is no longer linear), Powell and Ullmann [182] declared mean-based preconditioners to be ineffective, and so developed more advanced preconditioners; in contrast, Ullmann, Elman and Ernst [210] transformed a stationary diffusion problem with lognormal coefficient into a stationary convection-diffusion problem with a random coefficient depending linearly on the noise, before proving eigenvalue bounds as before. With a more computational slant, Tipireddy, Phipps, and Ghanem [207] and Rosseel and Vandewalle [185] compared the computational properties of several mean-based preconditioners and Elman, Miller, Phipps, and Tuminaro [67] compared the computational cost of mean-based preconditioners for stochastic Galerkin and stochastic collocation methods.

Seeking to apply mean-based preconditioning to more challenging problems, Powell and Silvester [181] performed computational investigations for mean-based preconditioners applied to stochastic Galerkin discretisations of the steady-state Navier–Stokes equations, and Sousedík and Elman [198] introduced a Gauss–Seidel-type preconditioner, using mean-based ideas, for the steady-state Navier–Stokes equations. Finally, Khan, Powell, and Silvester [128] applied mean-based preconditioning to stochastic Galerkin discretisations of the equations for nearly-

incompressible elasticity.

The works applying mean-based preconditioning to many individual systems (for the stationary diffusion equation) are those of Eiermann, Ernst and Ullmann [63]; Ernst, Powell, Silvester, and Ullmann [70]; and Gordon and Powell [98]. [63] contained computational results in a (decoupled) stochastic Galerkin setting; [70] proved eigenvalue bounds in the same setting, and [98] proved rigorous eigenvalue bounds in a stochastic collocation setting. All these works assume linear dependence on the noise, and show that mean-based preconditioning works well when the variance is sufficiently small.

We now turn our attention to mean-based preconditioning for the Helmholtz equation. The first work we discuss is the recent work of Wang and Liao [213]. They discretise a stochastic Helmholtz problem with  $k = 10$  and  $n$  given by a truncated Karhunen–Loève expansion (with either 4 terms or 1 term) and use a generalised polynomial chaos (gPC) expansion (see, e.g., [219]) for the solution  $u$ . Whilst they use mean-based preconditioning (in the ‘Kronecker product’ sense) they are more interested in investigating the effect of the number of terms in the gPC expansion on the accuracy of the discrete solution. Nonetheless, they see convergence using the mean-based preconditioner, although more iterations are needed when the random field is ‘close to’ exciting a resonant frequency (see [213, Example 4.2]).

The work most similar to ours is the work of Jin and Cai [125], who use a stochastic Galerkin discretisation with a doubly-orthogonal basis for a stochastic Helmholtz equation, resulting in around 5000 linear systems. They take  $k = 225$  and a Karhunen–Loève expansion with 4 terms for both (scalar-valued)  $A$  and  $n$ . The random variables in the Karhunen–Loève expansions are  $\text{Unif}(-\sqrt{3}, \sqrt{3})$  and  $\text{Unif}(-45\sqrt{3}, 45\sqrt{3})$  for  $A$  and  $n$  respectively. Their mean-based preconditioner is a 1-level additive Schwarz preconditioner, and they compare reusing the preconditioner with reusing the Krylov subspaces (an idea first introduced by Parks, De Sturler, Mackey, Johnson, and Maiti in [166]), as well as combining both techniques. Intriguingly, they see no additional benefit from reusing the preconditioner, but considerable benefit from recycling the Krylov subspaces. Based on our results in this chapter, we conjecture that they see no benefit from a single mean-based preconditioner because  $k$  is reasonably large, and therefore for most of the realisations,  $k \left\| \mathbb{E}[n] - n^{(j)} \right\|_{L^\infty(D; \mathbb{R})}$  and  $k \left\| \mathbb{E}[A] - A^{(j)} \right\|_{L^\infty(D; \mathbb{R}^{d \times d})}$  are not sufficiently small, and so there is little-to-no effect on the number of GMRES iterations from mean-based preconditioning. We conjecture that if they had used multiple preconditioners distributed around the stochastic parameter space, they would have seen computational improvements, as described in this chapter.

## 4.8 PROBABILISTIC NEARBY PRECONDITIONING RESULTS

We now briefly overview how one can prove probabilistic results on the effectiveness of nearby preconditioning. All of the results in Sections 4.1 and 4.3–4.5 above have been for deterministic (as opposed to stochastic) coefficients  $A$  and  $n$  (and we then applied these deterministic results to QMC methods for the Helmholtz equation in Section 4.6). Therefore we now turn our attention to obtaining probabilistic results on the effectiveness of nearby preconditioning for stochastic Helmholtz problems, i.e., Problems 3.1–3.3 from Chapter 3. Firstly, in Corollary 4.40 below, we

prove an ‘essentially deterministic’ result on the effectiveness of nearby preconditioning, before proving probabilistic results on the effectiveness of nearby preconditioning applied to stochastic problems. However, we will see that our efforts to prove probabilistic results are restricted by the applicability of the Elman estimate (Theorem 4.25 above).

Throughout this section we consider Problem 3.1 from Chapter 3 but with  $A = I$ , i.e., for simplicity we only consider the case of random  $n$ , although everything we say could be easily extended to include random  $A$ . To maintain consistent notation with the rest of this chapter we will use a superscript <sup>(2)</sup> to refer to the stochastic problem (e.g., the random coefficient will be  $n^{(2)}(\omega)$ , the solution will be  $u^{(2)}(\omega)$ , the matrices arising from the finite-element discretiation will be  $A^{(2)}(\omega)$ , etc.). We let  $n^{(1)} \in L^\infty(D; \mathbb{R})$  define a *deterministic* Helmholtz problem. We will use the discretisation of this deterministic Helmholtz problem to precondition the discretisations of the realisations of the stochastic Helmholtz problem. I.e., we will consider the performance of GMRES applied to

$$\left(A^{(1)}\right)^{-1} A^{(2)}(\omega) u = \left(A^{(1)}\right)^{-1} f. \quad (4.113)$$

For simplicity, in all that follows we will measure  $n_1 - n_2$  in the  $L^\infty$  norm, although one could use any of the weaker norms discussed in Section 4.5 above, and obtain analogous results.

### 4.8.1 Probabilistic theory for nearby preconditioning

**Definition 4.39** (Number of GMRES iterations required for convergence).

Let  $\text{GMRES}(\varepsilon, n^{(1)}, n^{(2)})$  denote the number of iterations required for GMRES in the unweighted norm  $\|\cdot\|_2$  with  $\|r_0\|_2 = 1$ , applied to

$$\left(A^{(1)}\right)^{-1} A^{(2)} u = \left(A^{(1)}\right)^{-1} f$$

to converge to within a tolerance  $\varepsilon$ , i.e., to achieve

$$\frac{\|r_m\|_2}{\|f\|_2} < \varepsilon.$$

Note that  $\text{GMRES}(\varepsilon, n^{(1)}, n^{(2)})$  is a random variable, see Lemma 4.41 below.

If we apply Theorem 4.12 to the problem (4.113) we can straightforwardly conclude the following corollary.

**Corollary 4.40** (Almost-sure nearby preconditioning). *Let  $0 < \varepsilon < 1$ ,  $n^{(2)} : \Omega \rightarrow L^\infty(D; \mathbb{R})$  satisfy the assumptions at the start of Section 3.1.1,  $n_1$ ,  $D_-$ , and  $f$  be as in Problem 4.1, and let the assumptions of Theorem 4.12 hold. Then  $\text{GMRES}(\varepsilon, n^{(1)}, n^{(2)})$  is bounded independently of  $k$  almost surely if*

$$\left\|n^{(1)} - n^{(2)}(\omega)\right\|_{L^\infty(D; \mathbb{R})} \leq \frac{1}{2C_2 k} \quad (4.114)$$

*almost surely.*

**Lemma 4.41** (GMRES( $\varepsilon, n_1, n_2$ ) is a random variable). *Under the assumptions of Corollary 4.40, GMRES( $\varepsilon, n^{(1)}, n^{(2)}$ ) is a random variable, i.e., GMRES( $\varepsilon, n^{(1)}, n^{(2)}$ ):  $\Omega \rightarrow \mathbb{R}$  is measurable.*

*Sketch Proof of Lemma 4.41.* All of the operations used in constructing the vectors  $x_m$  in the GMRES algorithm are measurable functions of  $x_{m-1}$  and  $(A^{(1)})^{-1}A^{(2)}$  (see, e.g., [97, Algorithms 11.4.2 and 5.1.3]), therefore  $(r_m)_{m=1}^N$  is a sequence of random variables, i.e., a stochastic process (see, e.g., [164, Definition 2.1.4]). The stopping criterion  $\|r_m\|_2/\|f\|_2 < \varepsilon$  is an exit time for the stochastic process  $x_m$  from the set  $\mathbb{C}^N \setminus B_{x^*}^{\mathbb{C}^N}(\varepsilon\|f\|_2)$ , where  $x^*$  is the true solution. Therefore, because we assume  $(\Omega, \mathcal{F}, \mathbb{P})$  is a complete probability space, it follows from, e.g., [164, Example 7.2.2] that GMRES( $\varepsilon, n_1, n_2$ ) is a stopping time (see [164, Definition 7.2.1]). Because GMRES( $\varepsilon, n_1, n_2$ ) is a stopping time, it is measurable with respect to the associated filtration (see, e.g., [164, Definition 3.2.2]), and so is measurable with respect to  $\mathcal{F}$ ; i.e., GMRES( $\varepsilon, n_1, n_2$ ) is a random variable.  $\square$

The numerical results in Section 4.3 above can be seen (in part) as confirming Corollary 4.40. Recall that in Section 4.3 we let  $n^{(1)} - n^{(2)}$  be a piecewise-constant random field, and we fixed  $\alpha = \left\|n^{(1)} - n^{(2)}\right\|_{L^\infty(D; \mathbb{R})}$  or  $\left\|A^{(1)} - A^{(2)}\right\|_{L^\infty(D; \mathbb{R}^{d \times d})}$  almost surely. When we fixed  $\alpha = 0.5/k$  almost surely (see Figures 4.3 and 4.6) we saw that the number of GMRES iterations was bounded independently of  $k$ . This behaviour is precisely that given in Corollary 4.40.

**Remark 4.42** (Drawbacks of Corollary 4.40). *There are two drawbacks of Corollary 4.40:*

1. *The condition (4.114) must hold almost surely, and*
2. *Corollary 4.40 does not give any explicit information on how the distribution of the number of GMRES iterations depends on the distribution of  $\left\|n^{(1)} - n^{(2)}\right\|_{L^\infty(D; \mathbb{R})}$ .*

*Drawback 1 is not ideal because in many physically realistic problems  $\|n_1 - n_2(\omega)\|_{L^\infty(D; \mathbb{R})}$  may be unbounded (e.g., if  $n_2$  is a lognormal random field) or even if bounded may not satisfy the condition (4.114) almost surely.*

To correct the deficiencies described in Remark 4.42 one would aim to prove a bound on the number of GMRES iterations depending explicitly on  $\left\|n^{(1)} - n^{(2)}(\omega)\right\|_{L^\infty(D; \mathbb{R})}$ , and then use this bound to prove a probabilistic estimate for the number of GMRES iterations. Such a bound is given in Lemma F.1 in Appendix F. However, such a bound will be highly pessimistic, and will impart little useful information. The reason for this lack of information is that the Elman estimate (Corollary 4.26 above) when applied to the nearby-preconditioned system  $(A^{(1)})^{-1}A^{(2)}$  only applies when  $k\left\|A^{(1)} - A^{(2)}\right\|_{L^\infty(D; \text{op})}$  and  $k\left\|n^{(1)} - n^{(2)}\right\|_{L^\infty(D; \mathbb{R})}$  are sufficiently small (as we saw in Theorem 4.11 above). Therefore one can only obtain detailed information on how the number of GMRES iterations depends on  $\left\|A^{(1)} - A^{(2)}\right\|_{L^\infty(D; \text{op})}$  and  $\left\|n^{(1)} - n^{(2)}\right\|_{L^\infty(D; \mathbb{R})}$  when these quantities are small (informally, when they are  $\lesssim 1/k$ ). In all other cases (again, informally, when these quantities are  $\gtrsim 1/k$ ) the only statement one can make about the convergence of GMRES is that there will be at most  $N$  iterations, where  $N$  is the number of degrees of freedom (this result is recalled in Corollary F.2 below). In summary, current results on GMRES convergence

will only allow us to prove what are likely to be very pessimistic bounds on how the number of GMRES iterations for  $(A^{(1)})^{-1}A^{(2)}$  depends on  $\|A^{(1)} - A^{(2)}\|_{L^\infty(D; \text{op})}$  and  $\|n^{(1)} - n^{(2)}\|_{L^\infty(D; \mathbb{R})}$ . For completeness, we record these results in Appendix F.

### 4.8.2 Numerical probabilistic results for nearby preconditioning

Notwithstanding the fact that we are limited in the probabilistic results that we can *prove* about nearby preconditioning, we will now see that we *observe* reasonable probabilistic behaviour when we perform numerical experiments. We again recall that (informally) Corollary 4.40 states that we obtain almost-surely bounded GMRES iterations if  $\|n^{(1)} - n^{(2)}\|_{L^\infty(D; \mathbb{R})} \lesssim 1/k$ . A plausible probabilistic analogue of this result would be that we have bounded *average* number of GMRES iterations if the standard deviation of  $\|n^{(1)} - n^{(2)}\|_{L^\infty(D; \mathbb{R})}$  is of the order  $1/k$ . We expect this result because the standard deviation of a random variable is a (probabilistic) measure of its variation. In Corollary 4.40 we show that the number of GMRES iterations is bounded almost surely if the variation in  $\|n^{(1)} - n^{(2)}\|_{L^\infty(D; \mathbb{R})}$  is bounded (of the order  $1/k$ ) almost surely. Therefore, it is reasonable to assume that the probabilistic analogue of the number of GMRES iterations (the average) is bounded if the probabilistic analogue of the variation in  $\|n^{(1)} - n^{(2)}\|_{L^\infty(D; \mathbb{R})}$  (the standard deviation) is bounded (of the order  $1/k$ ). We will see exactly this behaviour in our numerical experiments.

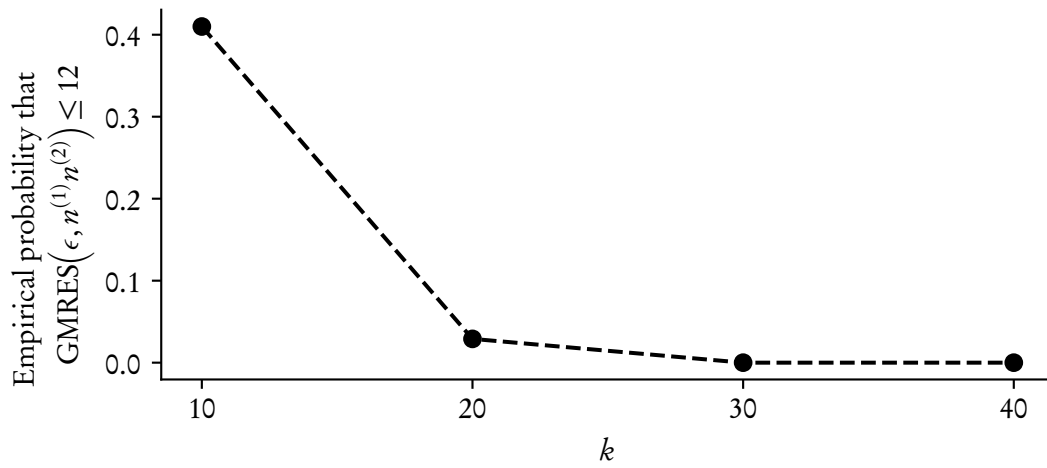
In our numerical experiments we use the computational setup described in Appendix G, with  $f = 1$  and  $g_I = 0$ ,  $A^{(1)} = A^{(2)} = I$ ,  $n^{(1)} = 1$ , and  $\|n^{(1)} - n^{(2)}\|_{L^\infty(D; \mathbb{R})}$  given by an exponential random variable with standard deviation  $\sigma$ . We consider three cases:

1.  $\sigma = 1$ ,
2.  $\sigma = \frac{1}{k}$ , and
3.  $\sigma = \frac{1}{k^2}$ .

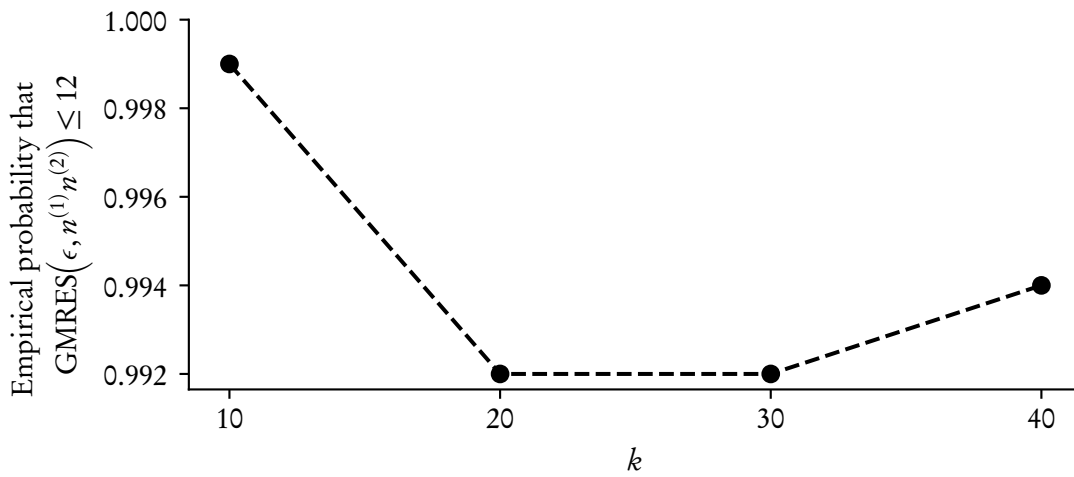
For each of these cases we calculate

$$\mathbb{P}(\text{GMRES}(\varepsilon, n_1, n_2) \leq 12). \quad (4.115)$$

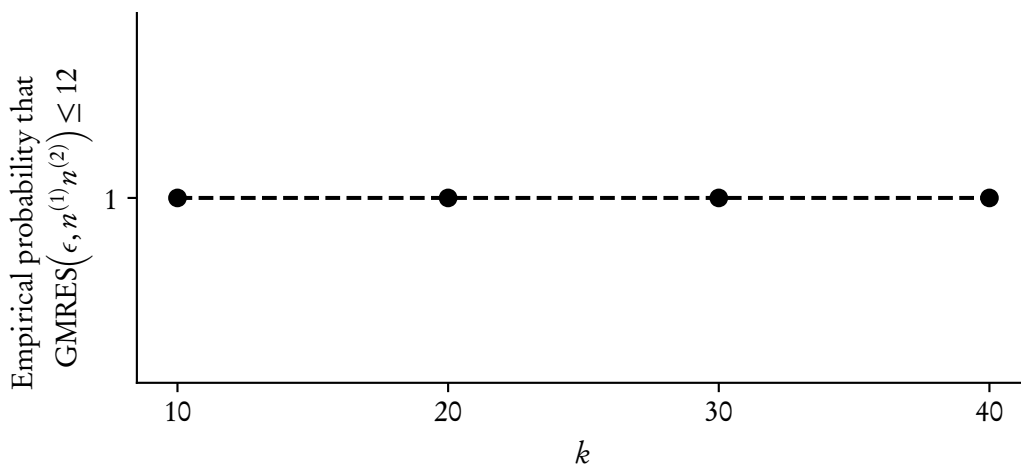
Based on the reasoning above we expect that in case 2 the probability (4.115) is *constant* as  $k$  increases, and using similar reasoning, we expect that in case 1 the probability (4.115) *decreases* as  $k$  increases and in case 3 the probability (4.115) *increases* as  $k$  increases. This is approximately the behaviour we observe in Figures 4.15a–4.15c. This behaviour demonstrates that whilst the theory developed in the rest of this chapter does not allow us to easily prove useful results about the probabilistic behaviour of nearby preconditioning, the theory does give us *intuition* as to what the probabilistic behaviour will be.



(a) The empirical probability that  $\text{GMRES}(\epsilon, n_1, n_2) \leq 12$  for  $\sigma = 1$ .



(b) The empirical probability that  $\text{GMRES}(\epsilon, n_1, n_2) \leq 12$  for  $\sigma = 1/k$



(c) The empirical probability that  $\text{GMRES}(\epsilon, n_1, n_2) \leq 12$  for  $\sigma = 1/k^2$ .

Figure 4.15: The empirical probability (calculated from 1000 realisations) that  $\text{GMRES}(\epsilon, n_1, n_2) \leq 12$  for  $k = 10, 20, 30, 40$ , where  $R = 12$ ,  $\epsilon = 10^{-5}$ ,  $n^{(1)} = 1$ , and  $\|n_1 - n_2\|_{L^\infty(D_R)} \sim \text{Exp}(\sigma)$  for different functional forms of  $\sigma$ .

## 4.9 SUMMARY AND FUTURE WORK

### 4.9.1 Summary

In this chapter we introduced, studied, and applied a nearby-preconditioning technique for multiple realisations of finite-element discretisations of heterogeneous Helmholtz problems motivated by Uncertainty Quantification (UQ). In particular:

- In Section 4.2 we gave rigorous results on the effectiveness of nearby preconditioning, giving  $k$ -explicit sufficient conditions (in terms of the  $L^\infty$ -norm of the difference in the coefficients) for nearby-preconditioned linear systems to achieve  $k$ -independent numbers of GMRES iterations. These results were confirmed by numerics, and supported by analogous PDE results.
- In Section 4.5 we extended the results in Section 4.2, by giving alternative  $k$ -explicit conditions in terms of the  $L^p$ -norms of the difference in the coefficients, for a range of exponents  $p$ . Numerical experiments indicated these conditions were not sharp in their  $k$ -dependence.
- In Section 4.8 we proved probabilistic analogues of the results in Section 4.2 and showed the results of some numerical experiments into the probabilistic behaviour of nearby preconditioning.
- In Section 4.6 we applied nearby preconditioning to a Quasi-Monte-Carlo (QMC) method for the Helmholtz equation, requiring thousands of individual PDE solves. We gave numerical evidence for how the number of QMC points must scale with  $k$ , and showed that nearby preconditioning applied to this problem is very effective, with around 98% of PDE solves using a previously-calculated preconditioner.

### 4.9.2 Future work

There are many possibilities for extending, improving, and applying the work in this chapter:

- Applying the idea of nearby preconditioning to other problems for which it is computationally intensive to construct preconditioners and, where possible, proving results on the effectiveness of nearby preconditioning. For linear problems, e.g., the time-harmonic Maxwell's equations, we would expect the behaviour and proofs of effectiveness to be analogous to that for the Helmholtz equation. For nonlinear problems (e.g., the steady-state Navier-Stokes equations, see, e.g., [181]), it is less clear how effective nearby preconditioning would be, and if it is possible to prove results on its effectiveness, but this could be a profitable line of future research.
- Investigating stochastic-dimension-independent methods for choosing the preconditioning points when applying nearby preconditioning to QMC methods. E.g., one may be able to use the nestedness of QMC points (as is the case with embedded lattice rules, see, e.g., [51, Property 3, p.2169]) to choose preconditioning points in a stochastic-dimension-independent way.

- Applying nearby preconditioning to other UQ methods for the Helmholtz equation. For example, applying nearby preconditioning to Multi-Level Monte-Carlo methods for the Helmholtz equation (see Chapter 5), where a preconditioner could be calculated on one ‘level’ (one discretisation), and then transferred other ‘levels’, perhaps using multigrid smoothing/prolongation. Alternatively, applying nearby preconditioning Markov Chain Monte-Carlo methods for Bayesian inverse problems for the Helmholtz equation. Nearby preconditioning is a natural fit for such problems, where realisations are chosen one at a time, with the next realisation typically being close to the current one.
- Analysing rigorously the behaviour of QMC methods applied to the Helmholtz equation. We understand that such work is already underway in [88], but there is clearly scope to develop the theory; in particular, in understanding how the number of QMC points should scale with  $k$  in order to obtain bounded QMC error.

# Monte-Carlo and Multi-Level Monte-Carlo methods for the Helmholtz equation

## 5.1 INTRODUCTION

In Section 4.6 we considered how to speed up solving the individual linear systems in UQ algorithms for the stochastic Helmholtz equation via nearby preconditioning. We now consider how, using a Multi-Level Monte-Carlo method, one can reduce the total number of linear systems we must solve. In particular, we prove bounds on the computational effort needed for Monte Carlo (MC) and Multi-Level Monte-Carlo (MLMC) methods for the stochastic Helmholtz equation. We compare and contrast the behaviour of these methods for different wavenumbers and tolerances and we show that Multi-Level Monte-Carlo methods asymptotically (as the prescribed tolerance goes to 0) require less work than Monte-Carlo methods.

We highlight that, in contrast to our empirical analysis of Quasi-Monte-Carlo (QMC) methods in Section 4.6.3, in this chapter we provide a rigorous analysis of Monte-Carlo and Multi-Level Monte Carlo methods. We prove how Monte-Carlo and Multi-Level Monte-Carlo methods must be adapted for increasing  $k$  to ensure the overall error (both numerical and statistical) remains bounded, and we prove bounds on the expected computational cost of both the Monte-Carlo and Multi-Level Monte-Carlo methods.

We now provide a brief overview of this chapter. In Section 5.2 we give an introduction to Monte-Carlo and Multi-Level Monte-Carlo methods, and discuss some of the challenges in applying them to the Helmholtz equation. We then review literature on Multi-Level Monte-Carlo methods, focusing only on those works that are relevant for our study of Multi-Level Monte-Carlo methods applied to the stochastic Helmholtz equation. In Section 5.3 we give an abstract setting for a  $k$ -dependent analysis of Multi-Level Monte-Carlo methods. In Section 5.4 we prove a bound on the computational work for the Monte-Carlo method in this abstract setting and in Section 5.5 we prove an analogous result for the Multi-Level Monte-Carlo method. Finally, in Section 5.6 we show that the stochastic Helmholtz equation fits into this abstract setting, and we then compare and contrast the behaviour of Monte-Carlo and Multi-Level Monte-Carlo methods for the stochastic Helmholtz equation.

## 5.2 BACKGROUND ON BOTH MONTE-CARLO AND MULTI-LEVEL MONTE-CARLO METHODS

### 5.2.1 The ideas of Monte-Carlo and Multi-Level Monte-Carlo methods

Throughout this section we assume our goal is to compute an approximation of  $\mathbb{E}[Q]$ , where  $Q : \Omega \rightarrow \mathbb{R}$  is a random variable. We assume we have access to a family of random variables  $Q_b : \Omega \rightarrow \mathbb{R}$ , indexed by  $b > 0$ , where  $Q_b$  approximates  $Q$  in a sense made precise in Assumption 5.1 below. We assume we can compute samples of  $Q_b$ , for any  $b > 0$ . When we consider quantities of interest corresponding to the solution of a stochastic PDE,  $Q$  will be a function of the true solution  $u$ , and  $Q_b$  will be a function of the finite-element approximation  $u_b$  of  $u$ . However, to explain the ideas behind Monte-Carlo and Multi-Level Monte-Carlo methods we will only occasionally need to mention  $u$  and  $u_b$ . Therefore, for most of this chapter we will instead work with  $Q$  and  $Q_b$ . Our exposition throughout this chapter is based on that of Cliffe, Giles, Scheichl, and Teckentrup [49], who proved the first results for Multi-Level Monte-Carlo methods for elliptic PDEs.

#### Monte-Carlo Estimators

The *Monte-Carlo estimator*  $\hat{Q}_b^{\text{MC}}$  of  $Q$  is the simplest possible estimator of  $\mathbb{E}[Q]$ . The estimator is given by

$$\hat{Q}_b^{\text{MC}} := \frac{1}{N_{\text{MC}}} \sum_{j=1}^{N_{\text{MC}}} Q_b(\omega^{(j)}),$$

where the  $\omega^{(j)}$  are independent and identically distributed samples from the probability space  $\Omega$ .

One would expect that reducing  $b$  and increasing  $N_{\text{MC}}$  would give a more accurate approximation of  $\mathbb{E}[Q]$ . Therefore our analysis of  $\hat{Q}_b^{\text{MC}}$  seeks to answer the question ‘How should we choose  $b$  and  $N_{\text{MC}}$  to ensure the error is less than a prescribed tolerance  $\varepsilon > 0$  (with minimal computational work)?’ The standard relationship between  $\varepsilon$  and  $N_{\text{MC}}$  is that one should take  $N_{\text{MC}} \sim \varepsilon^{-2}$ , see, e.g., [49, Text after equation (3)]. We prove a generalised version of this relationship in Theorem 5.10 below.

#### Multi-Level Monte-Carlo Estimators

In contrast to the Monte-Carlo estimator, where all of the approximations  $Q_b(\omega^{(j)})$  are performed for a single specified mesh size<sup>1</sup>  $b$ , the Multi-Level Monte-Carlo estimator computes approximations for a hierarchy of mesh sizes  $b_0 \geq b_1 \geq \dots \geq b_L$ . The rationale for this computation is the observation that the telescoping sum identity

$$\mathbb{E}[Q_{b_L}] = \mathbb{E}[Q_{b_0}] + \sum_{l=1}^L \mathbb{E}[Q_{b_l} - Q_{b_{l-1}}] \quad (5.1)$$

<sup>1</sup>For technical reasons due to the randomness of the coefficients, some of these meshes may be refined on a sample-by-sample basis, see Section 5.3 below. We ignore this technicality in the current discussion, but it will be fully addressed in Section 5.3.

holds and therefore, if one computes estimators  $\hat{Y}_0$  for  $\mathbb{E}[Q_{b_0}]$  and  $\hat{Y}_l$  for  $\mathbb{E}[Q_{b_l} - Q_{b_{l-1}}]$ , then one can construct an estimator for  $\mathbb{E}[Q_{b_L}]$ ,

$$\hat{Q}_{b_L}^{\text{ML}} := \hat{Y}_0 + \sum_{l=1}^L \hat{Y}_l.$$

In this chapter, the estimators  $\hat{Y}_l$  will be Monte-Carlo estimators using  $N_0$  samples of  $Q_{b_0}$  (for  $\hat{Y}_0$ ) and  $N_l$  samples of  $Q_{b_l} - Q_{b_{l-1}}$  (for  $\hat{Y}_l$ ,  $l \geq 1$ ).

Our analysis of  $\hat{Q}_{b_L}^{\text{ML}}$  then seeks to answer the question ‘How should we choose  $b_L$  and  $N_0, N_1, \dots, N_L$  to ensure the error is less than a prescribed tolerance  $\varepsilon > 0$  (with minimal computational work)?’ The answer is long, and so we answer this question fully in our main, new result, Theorem 5.13 below.

The reason one expects the Multi-Level Monte-Carlo estimator to require less computational effort than the Monte-Carlo estimator is that one expects the variance  $\mathbb{V}[Q_{b_l} - Q_{b_{l-1}}]$  to decrease as  $l$  increases. One expects this decrease because the quantities of interest  $Q_{b_l}$  and  $Q_{b_{l-1}}$  are obtained from finite-element approximations  $u_{b_l}$  and  $u_{b_{l-1}}$ , and one expects these approximations to get closer together as  $l$  increases. A basic calculation confirms this; indeed, provided the solution  $u$  is sufficiently smooth, and  $b_l \sim b_{l-1}$  uniformly in  $l$ , (e.g., we obtain mesh  $l$  by uniform refinement of mesh  $l - 1$ .) then

$$\|u_{b_l} - u_{b_{l-1}}\|_{H^1} \leq \|u_{b_l} - u\|_{H^1} + \|u - u_{b_{l-1}}\|_{H^1} \lesssim b_l + b_{l-1} \sim b_{l-1} \rightarrow 0 \text{ as } l \rightarrow L.$$

Therefore  $u_{b_l}$  and  $u_{b_{l-1}}$  get closer together as  $l$  increases, and one expects analogous behaviour for  $Q_{b_l}$  and  $Q_{b_{l-1}}$  (if  $Q$  is a continuous function of  $u$ , this behaviour is immediate). Since one takes the number of samples in a Monte-Carlo estimator to be proportional to the variance of the sampled quantity (i.e.,  $Q_{b_0}$  or  $Q_{b_l} - Q_{b_{l-1}}$  in this case), the fact that  $\mathbb{V}[Q_{b_l} - Q_{b_{l-1}}]$  gets smaller as  $l$  increases should mean the number of samples of  $Q_{b_l} - Q_{b_{l-1}}$  can decrease as  $l$  increases. As the computational cost of performing numerical solves is higher for finer meshes (i.e., the cost of computing  $Q_{b_l} - Q_{b_{l-1}}$  increases as  $l$  increases), we expect that the Multi-Level Monte-Carlo estimator allows us to perform a large number of (cheap) solves on the coarser meshes, and a small number of (expensive) solves on the fine meshes, i.e.  $N_0 \geq N_1 \geq \dots \geq N_L$ . Replacing solves on finer meshes with solves on coarser meshes in this way should result in computational savings, as is seen for the stationary diffusion equation in [49].

### 5.2.2 Challenges in Monte-Carlo and Multi-Level Monte-Carlo methods for the Helmholtz equation

Analysing Monte-Carlo and Multi-Level Monte-Carlo methods for the Helmholtz equation has two main challenges that are not present in the analysis of these methods for, e.g., the stationary diffusion equation.

Firstly, the behaviour of Monte-Carlo and Multi-Level Monte-Carlo methods for the Helmholtz equation will be  $k$ -dependent, because the behaviour of the finite-element method for the

Helmholtz equation is  $k$ -dependent, see Section 2.3. Because of this  $k$ -dependent behaviour, we would like our analysis of these methods to be completely  $k$ -explicit. In particular, since we have access to  $k$ -explicit finite-element-error estimates for the Helmholtz equation in Section 2.4 above, we are able to make our analysis of Monte-Carlo and Multi-Level Monte Carlo methods  $k$ -explicit; we can re-prove the standard results on computational complexity for Monte-Carlo and Multi-Level Monte-Carlo methods with the  $k$ -dependence incorporated explicitly.

Secondly, the finite-element approximation  $u_b$  of the solution  $u$  of the stochastic Helmholtz equation may not exist for all  $h > 0$ , and the criteria to prove its existence and uniqueness may be dependent on the coefficients  $A$  and  $n$ . I.e., for fixed  $h$ ,  $u_b(\omega^{(1)})$  may exist and be unique, but  $u_b(\omega^{(2)})$  may not, for some  $\omega^{(1)} \neq \omega^{(2)} \in \Omega$ . To see why this is the case, recall from the definitions of  $(hk^a, hk^b)$ -accuracy and -data-accuracy for the finite-element solution of the Helmholtz equation (Definitions 2.24 and 2.26) that the finite-element approximation  $u_b$  only exists for  $h$  sufficiently small (with the definitions of  $(hk^a, hk^b)$ -accuracy and -data-accuracy defining ‘sufficiently small’ in terms of  $k$  and other quantities). Moreover, the criteria for ‘sufficiently small’ also depend on the coefficients  $A$  and  $n$  (see Remark 2.30). Therefore, when  $A$  and  $n$  are stochastic, the existence and uniqueness of  $u_b(\omega)$  is not only  $h$ -dependent but also  $\omega$ -dependent. Putting the above challenge into the language of the random variables  $Q_b$ , the random variable  $Q_b$  may not exist or be unique for all  $h > 0$ , and moreover, its existence and uniqueness may be sample-dependent. I.e.,  $Q_b(\omega^{(1)})$  may exist and be unique, but  $Q_b(\omega^{(2)})$  may not, for  $\omega^{(1)} \neq \omega^{(2)}$ .

This sample-dependence poses an issue for Monte-Carlo and Multi-Level Monte-Carlo methods. The method may require us to compute  $Q_b(\omega^{(j)})$ , but there is no guarantee that  $Q_b(\omega^{(j)})$  exists. Therefore, we need to modify our methods to deal with this sample-dependence. Such a modification to Monte-Carlo and Multi-Level Monte-Carlo methods for sample-dependent existence and uniqueness criteria was given by Graham, Parkinson, and Scheichl in [104] (and in Parkinson’s PhD thesis [165]), in the context of the Radiative Transport Equation (RTE). The RTE is an integro-differential equation whose numerical approximations have similar sample-dependent existence and uniqueness criteria to the Helmholtz equation. We adopt their approach for dealing with the sample-dependence, this approach is discussed in Section 5.3 below.

### 5.2.3 Literature Review of Multi-Level Monte-Carlo methods

We focus our literature review on (i) foundational works in Multi-Level Monte-Carlo methods, to provide a little context for our work on the Helmholtz equation, and (ii) applications of Multi-Level Monte-Carlo methods to problems sharing the challenges outlined in Section 5.2.2 above. As far as we are aware, there is no prior work on Multi-Level Monte-Carlo methods explicitly incorporating the dependence on an additional parameter, and so we just mention works dealing with sample-dependent criteria for the numerical approximation. For a wider-ranging overview of the literature, we refer the reader to the review article [95] and the webpage [93], the latter of which is kept up-to-date with a range of recent work on Multi-Level Monte-Carlo methods.

Multi-level Monte Carlo methods for stochastic differential equations were first introduced by Giles [94] for time-dependent SDEs, with applications mostly arising in finance, although the ideas

were present in earlier work by Heinrich [111, 112] on multilevel methods for parametric integration. Multi-Level Monte-Carlo methods were first applied to elliptic (i.e., non-time-dependent) PDEs by Barth, Schwab, and Zollinger in [15] and Cliffe, Giles, Scheichl, and Teckentrup in [49] for the stationary diffusion equation with an application in porous media flow. In particular, the statement of the Multi-Level Monte-Carlo complexity theorem in [49, Theorem 1] is the basis for our statement of a Multi-Level Monte-Carlo complexity theorem for the Helmholtz equation in Theorem 5.13 below. We highlight that a key result of [49, Theorem 1] is that Multi-Level Monte-Carlo methods *always* outperform Monte-Carlo methods, at least in the setting given in [49].

We mention the work of Scarabosio [191], who applied the Multi-Level Monte-Carlo method to a Helmholtz transmission problems with an uncertain boundary (i.e. a Helmholtz problem where  $A$  and  $n$  are piecewise constant, both jumping across a random interface). Her particular emphasis was on quantities of interest given by point evaluations of the solution; other UQ algorithms do not behave well for such QoIs, see [191, Section 3.3]. She works under the assumption that  $k$  is small, i.e., a ‘large wavelength assumption’ [191, Assumption 3.1]. In this setting she shows that the Multi-Level Monte-Carlo method for the transmission problem fits into the standard framework of Multi-Level Monte-Carlo methods [191, Proposition 4.2] (as in, e.g., [49, 95]) and that the numerical behaviour of the method is as predicted by the theory [191, Section 6].

We now highlight two bodies of work on Multi-Level Monte-Carlo methods with sample-dependent criteria; the work of Mishra, Schwab, and Šukys on Monte-Carlo and Multi-Level Monte-Carlo methods for time-domain wave propagation and the work of Graham, Parkinson, and Scheichl on Monte-Carlo and Multi-Level Monte-Carlo methods for the Radiative Transport Equation.

The work of Mishra, Schwab, and Šukys covers Monte-Carlo and Multi-Level Monte-Carlo methods for a range of linear and nonlinear hyperbolic problems, see, e.g., [201]. However, we focus just on their results for linear problems, as then the PDE involved is the time-domain wave equation with random coefficients and random initial data, whose Fourier transform in time is the Helmholtz equation (recall the discussion in Section 1.1.1). This work on linear wave propagation is contained in the papers [202, 151] and in Šukys’ PhD thesis [201]. They discretised the individual realisations of the wave problems using a finite-volume method in space and specialised time-stepping algorithms in time (see, e.g., [151, Section 3.1]). Because the PDEs in these works have random coefficients, the CFL condition for the numerical method (this condition depends on the coefficients) is also random, meaning the number of time steps used in the time-stepping algorithm is random. (The spatial discretisation is fixed across all realisations.) In [202] the authors analyse the error against the expected work (analogous to our analysis in Sections 5.4 and 5.5 below). In [151] the authors present more realistic test cases, and a load-balancing algorithm for applying the Multi-Level Monte-Carlo method on high-performance computers. The load-balancing algorithm is needed because the different individual solves have different computational requirements, because of the random number of timesteps mentioned above. They see that the Multi-Level Monte-Carlo method consistently outperforms

the Monte-Carlo method.

Another collection of relevant work is that of Graham, Scheichl, and Parkinson [103, 165, 104] on UQ methods (including Multi-Level Monte-Carlo methods) for the Radiative Transport Equation (RTE), as mentioned above. The main relevance of this work for our study of the Helmholtz equation is that, as mentioned above, proving the numerical approximation of the solution of the RTE exists and is unique requires a coefficient-dependent discretisation condition (see [104, Theorem 4.12]). (This condition is analogous to a mesh constraint, except the RTE is not discretised with a traditional mesh, as it is defined on both spatial and angular variables.) When this discretisation constraint is carried over into a UQ setting, the RTE has a sample-dependent discretisation condition. Therefore, for some samples a given discretisation may be too coarse to guarantee existence and uniqueness. This sample-dependence is very similar to the situation we encounter for the Helmholtz equation, where the condition to ensure data-accuracy is  $A$ - and  $n$ -dependent (see Corollary 2.41 above), and therefore there will be a sample-dependent condition in the UQ setting.

The remedy proposed for this sample-dependence by Graham, Parkinson, and Scheichl is to *selectively* refine the discretisation *only* for those samples that require a finer discretisation. We adopt this strategy for the Helmholtz equation, as outlined in Section 5.3 below. Moreover, Graham, Parkinson, and Scheichl show that (under suitable assumptions on the randomness, that are satisfied for a range of realistic random field models) this sample-wise refinement does not affect the asymptotics of the expected cost of the algorithm, because it only needs to happen on a set of samples of small measure, see [104, Lemma 5.8]. We obtain similar results for the Helmholtz equation in Lemma 5.6 below.

### 5.3 AN ABSTRACT SETTING FOR BOTH THE MULTI-LEVEL MONTE-CARLO AND MONTE-CARLO METHODS, MOTIVATED BY THE HELMHOLTZ EQUATION

We now define the concepts and quantities needed to define and discuss Monte-Carlo and Multi-Level Monte-Carlo methods for the Helmholtz equation. However, at this stage we work at an abstract level, i.e., we consider random variables  $Q$  and  $Q_b$  rather than the solution  $u$  of a stochastic Helmholtz equation and its approximations  $u_b$ . This abstraction will help simplify the presentation of the additional challenges one has for the Helmholtz equation. However, when constructing this abstract setting, our definitions will be motivated by properties of the finite-element solution of the Helmholtz equation. Therefore in Section 5.6 below, we will show that the Helmholtz equation fits into our abstract setting, and therefore our abstract results are applicable to the Helmholtz equation itself.

We assume that we have a parameter  $k > 0$  (corresponding to the wavenumber  $k$  in the Helmholtz equation), that there exists a random variable  $Q : \Omega \rightarrow \mathbb{R}$  depending on  $k$ , and that our goal is to approximate  $\mathbb{E}[Q]$ . Our first aim would be to define a family of random variables  $(Q_b)_{b>0}$  (corresponding to the finite-element approximations  $u_b$ ). However, as has been discussed in Section 5.2.2, the existence and uniqueness of finite-element approximations of the

Helmholtz equation is sample-dependent, and therefore we want our abstract setting to reflect this dependence.

If we recall our finite-element-error bound in Theorem 2.39 above (from which we concluded the  $h$ -finite-element method for the heterogeneous Helmholtz equation is  $hk^{(2p+1)/2p}$ -data-accurate in Corollary 2.41), then we see that the existence and uniqueness criteria and the error bound are  $h$ -,  $k$ -,  $A$ -, and  $n$ -dependent. (For simplicity, in this chapter we assume  $C_{\text{stab}} \sim 1$ , i.e., the Helmholtz problem is nontrapping almost surely, although one could easily generalise the results of this chapter to the trapping case, albeit with a worse  $k$ -dependence.) Therefore, when we move to the UQ case, where  $A$  and  $n$  are random fields, the existence and uniqueness criterion will be  $h$ -,  $k$ -, and sample-dependent. Motivated by this dependence, we make the following assumption on the existence and uniqueness of the random variable  $Q_h$ .

**Assumption 5.1** (Probabilistic version of Theorem 2.39). *There exist random variables  $C_1$  and  $\tilde{c}_1$ , with  $\mathbb{E}[\tilde{c}_1] < \infty$ , and constants  $a, \alpha, \sigma > 0$  all independent of  $h$  and  $k$  such that, for  $h > 0$  if*

$$h < C_1(\omega)k^{-a}, \quad (5.2)$$

then  $Q_h(\omega)$  exists, is unique, and satisfies

$$|Q(\omega) - Q_h(\omega)| \leq \tilde{c}_1(\omega)h^\alpha k^\sigma. \quad (5.3)$$

**Remark 5.2** (Comments on Assumption 5.1).

- *As an example, Theorem 2.39 shows Assumption 5.1 holds for the stochastic Helmholtz equation with  $Q(\cdot) = \|\cdot\|_{H_k^1(D)}$ ,  $a = (2p+1)/2p$ ,  $\alpha = 2p$ , and  $\sigma = 2p+1$ , where we have used the fact that as discussed in Remark 2.40, the final terms in the bounds (2.63) and (2.64) are the dominant terms.*
- *In principle, one can obtain explicit formulae for  $C_1$  and  $\tilde{c}_1$  from (2.62)–(2.64). However, as noted in Remarks 2.43 and 2.45, (2.62)–(2.64) may not depend optimally on  $n$ , and are not completely explicit in their  $A$ -dependence. Therefore, using (2.62)–(2.64) to define  $C_1$  and  $\tilde{c}_1$  would mean  $C_1$  and  $\tilde{c}_1$  may not depend optimally on  $\omega$ , nor be completely explicit in their  $\omega$ -dependence. Therefore we do not specify (here, or in Section 5.6 below) the form of  $C_1$  or  $\tilde{c}_1$ .*

A crucial consequence of Assumption 5.1 is that, as stated above, for a given  $h > 0$  the value  $Q_h(\omega)$  may not be defined for all  $\omega \in \Omega$ . To cope with this issue, we follow the approach of Graham, Parkinson, and Scheichl in [104]. For a fixed  $h > 0$ , we define the set  $\Omega_{\text{bad}} = \{\omega \in \Omega : (5.2) \text{ is not satisfied}\}$ . On  $\Omega_{\text{bad}}$  we refine the mesh on a sample-by-sample basis so that (5.2) is satisfied on the refined mesh. We then show that this additional refinement does not change the  $h$ -dependence of the expected cost of a single sample. (The proof of this fact requires the assumption that  $\Omega_{\text{bad}}$  has small probability; this assumption is stated more formally in Assumption 5.5 below, and is proved by Graham, Scheichl, and Parkinson in the neutron-transport context in [104, Lemma 5.3].)

We now give the above scheme more precisely. For fixed  $h > 0$ , and given  $\omega \in \Omega$ , we define

$$h_\omega^{\max} = C_1(\omega)k^{-a}, \quad (5.4)$$

that is,  $h_\omega^{\max}$  is the largest mesh size that satisfies (5.2). We then define

$$h_\omega = \min\{h, h_\omega^{\max}\}, \quad (5.5)$$

that is, the behaviour of  $h_\omega$  as  $h \downarrow 0$  is governed by  $h$ , but  $h_\omega$  is always small enough so that it satisfies (5.2). We can now define the quantity

$$\tilde{Q}_b(\omega) = Q_{h_\omega}(\omega). \quad (5.6)$$

Observe that, by construction,  $\tilde{Q}_b(\omega)$  exists for all  $\omega \in \Omega$ , because if  $\omega \in \Omega_{\text{bad}}$ , then  $h_\omega = h_\omega^{\max}$  and by definition of  $h_\omega^{\max}$ , the value  $Q_{h_\omega^{\max}}(\omega)$  exists.

**Remark 5.3** (Is  $\tilde{Q}_b$  a random variable?). *Throughout this chapter, we assume  $\tilde{Q}_b$  is a random variable. One could, in principle, prove this fact, but the proof would likely be very involved. One would need to show the map  $(\omega, h) \mapsto Q_b(\omega)$  is measurable (for all pairs  $(\omega, h)$  such that this map is defined) with respect to a suitable  $\sigma$ -algebra, and then combine this fact with the fact that  $h_\omega$  is a random variable (and thus measurable) to conclude that the map  $\omega \mapsto \tilde{Q}_b$  is measurable. Proving that the map  $(\omega, h) \mapsto Q_b(\omega)$  is measurable in the context of finite-element discretisations of the Helmholtz equation would be very technical, and would contribute little to the discussion of Monte-Carlo and Multi-Level Monte-Carlo methods for the Helmholtz equation. Therefore, we instead assume  $\tilde{Q}_b$  is a random variable.*

Because  $\tilde{Q}_b$  is associated with a random mesh size  $h_\omega$ , the cost of computing one realisation of  $\tilde{Q}_b$  will also be a random variable. Therefore, we make the following assumption on the cost of computing one realisation of  $\tilde{Q}_b$ . In particular, we assume that the cost is driven by the *actual* mesh size that is used in the computations,  $h_\omega$ . We let  $\mathcal{C}(\cdot)$  denote the cost of computing one realisation of a random variable.

**Assumption 5.4** (Cost of one realisation of  $\tilde{Q}_b$ ). *There exists  $\gamma > 0$  and a positive random variable  $\tilde{c}_3$ , where  $\tilde{c}_3$  does not depend on  $h$  and  $k$ , such that*

$$\mathcal{C}(\tilde{Q}_b(\omega)) \leq \tilde{c}_3(\omega)h_\omega^{-\gamma},$$

We can now show that, provided the set  $\Omega_{\text{bad}}$  has small probability (in a sense made precise in Assumption 5.5 below), the expected cost of computing one realisation of  $\tilde{Q}_b$  is driven only by  $h$ . I.e., the expectation does not ‘see’ the additional refinement needed for  $\omega \in \Omega_{\text{bad}}$ , because these samples occur with low probability.

**Assumption 5.5.** *The quantity*

$$c_3 := \mathbb{E}\left[\tilde{c}_3\left(1 + C_1^{-\gamma}\right)\right] \quad (5.7)$$

is finite.

One can conclude from Assumption 5.5 that the set  $\Omega_{\text{bad}}$  has low probability, as in [104, Text at the bottom of p. 21]. Observe that  $C_1$  governs where the mesh needs to be refined (since if  $C_1(\omega)$  is small, then a smaller mesh size is needed). Therefore if terms involving  $C_1^{-1}$  have finite expectation (as in (5.7)), then  $C_1$  is small with low probability, i.e.,  $\Omega_{\text{bad}}$  has low probability.

**Lemma 5.6** (Expected cost of one sample of  $\tilde{Q}_b$ ). *If Assumptions 5.4 and 5.5 hold, then*

$$\mathbb{E}[\mathcal{C}(\tilde{Q}_b)] \leq c_3(b^{-\gamma} + k^{a\gamma}). \quad (5.8)$$

*Proof of Lemma 5.6.* The proof follows closely that in [104, Lemma 5.8]. We have

$$\mathcal{C}(\tilde{Q}_b(\omega)) \leq \tilde{c}_3(\omega)h_\omega^{-\gamma} \leq \tilde{c}_3(\omega)(b^{-\gamma} + (h_\omega^{\max})^{-\gamma}) \quad (5.9)$$

by Assumption 5.4 and (5.5). Then using (5.4) and (5.9) we obtain the bound

$$\mathcal{C}(\tilde{Q}_b(\omega)) \leq \tilde{c}_3(\omega)h^{-\gamma} + (\tilde{c}_3 C_1^{-\gamma})(\omega)k^{a\gamma}, \quad (5.10)$$

and therefore since Assumption 5.5 holds, we obtain (5.8).  $\square$

To prove results on the expected computational cost and convergence of Monte-Carlo and Multi-Level Monte-Carlo methods, we need not only the previous lemma on the expected computational cost of a single sample of  $\tilde{Q}_b$ , but also the following lemma on the convergence of  $\tilde{Q}_b$  to  $Q$ , analogous to [104, Theorem 5.14].

**Lemma 5.7** (Convergence of  $\tilde{Q}_b$  to  $Q$ ). *Under Assumption 5.1*

$$\mathbb{E}[|\tilde{Q}_b - Q|] \leq c_1 k^\sigma h^\alpha, \quad (5.11)$$

where  $c_1 = \mathbb{E}[\tilde{c}_1]$ .

*Proof of Lemma 5.7.* The proof is immediate from (5.6), Assumption 5.1 and the fact that  $h_\omega \leq h$  (by (5.5)).  $\square$

**Remark 5.8** (Assumption 5.1 is sufficient, but not necessary). *The proofs of Theorems 5.10 and 5.13 below (our main technical results) only require a bound on*

$$|\mathbb{E}[\tilde{Q}_b - Q]| \leq c_1 k^\sigma h^\alpha; \quad (5.12)$$

*a weaker condition than the bound (5.11) which we use in these proofs. Therefore, in principle one could replace Assumption 5.1, a pathwise assumption which leads to (5.11), with the weaker assumption on the difference in mean (5.12).*

Before we move on to study Monte-Carlo and Multi-Level Monte-Carlo methods, we define the notion of error that we use when studying these methods.

**Definition 5.9** (Root-mean-squared error). *Given a random variable  $Q$  and an estimator  $\hat{Q}$  of  $Q$ , the root-mean-squared error of  $\hat{Q}$  is*

$$\text{Err}(\hat{Q}) := \left( \mathbb{E} \left[ \left| \hat{Q} - Q \right|^2 \right] \right)^{\frac{1}{2}}.$$

## 5.4 MONTE-CARLO METHODS

We now prove a  $k$ -explicit bound on the expected computational complexity of the Monte-Carlo method in the above abstract setting (which is, of course, motivated by the stochastic Helmholtz equation). Recall that the Monte-Carlo estimator of  $Q$  is defined by

$$\hat{Q}_b^{\text{MC}} = \frac{1}{N_{\text{MC}}} \sum_{j=1}^{N_{\text{MC}}} \tilde{Q}_b^{(j)},$$

where the  $\tilde{Q}_b^{(j)}$  are independently and identically distributed samples of  $\tilde{Q}_b$ .

We have the following theorem on the computational complexity of the Monte-Carlo estimator  $\hat{Q}_b^{\text{MC}}$ , which is a generalisation of the standard proof of the complexity of the Monte-Carlo method (see, e.g., [49, Section 2.1]) to the above  $k$ -dependent setting. In this theorem, the notation  $\sim$  denotes a hidden constant that is independent of  $h$ ,  $k$ , and  $\varepsilon$ .

**Theorem 5.10** (Computational complexity of Monte-Carlo). *Let Assumptions 5.1, 5.4, and 5.5 hold. Given  $\varepsilon \in (0, 1)$ , if*

$$h \sim \left( \sqrt{2c_1} \right)^{-\frac{1}{\alpha}} k^{-\frac{\sigma}{\alpha}} \varepsilon^{\frac{1}{\alpha}}, \quad (5.13)$$

and

$$N_{\text{MC}} \sim 2\mathbb{V}[\tilde{Q}_b] \varepsilon^{-2} \quad (5.14)$$

then

$$\text{Err}(\hat{Q}_b^{\text{MC}}) \sim \varepsilon \quad (5.15)$$

and the computational complexity of  $\hat{Q}_b^{\text{MC}}$  satisfies

$$\mathbb{E}[\mathcal{C}(\hat{Q}_b^{\text{MC}})] \sim \mathbb{V}[\tilde{Q}_b] \left( \varepsilon^{-2-\frac{\gamma}{\alpha}} k^{\frac{\gamma\sigma}{\alpha}} + \varepsilon^{-2} k^{a\gamma} \right). \quad (5.16)$$

The first term in (5.16) is analogous to the standard cost term one obtains in the analysis of Monte-Carlo methods (see, e.g., [49, Section 2.1]). The second term in (5.16) arises from the  $k$ -dependence of (5.4). The reason the second term in (5.16) has a better  $\varepsilon$ -dependence than the first term is that (5.2) is an  $\varepsilon$ -independent criterion, whereas ensuring the error is small (via the mesh constraint (5.13)) is an  $\varepsilon$ -dependent criterion.

*Proof of Theorem 5.10.* The proof is nearly identical to the standard proof for Monte-Carlo methods, see, e.g., [49, Section 2.1]. We can first perform a so-called bias-variance decomposition of

the error

$$\begin{aligned}
\text{Err}(\hat{Q}_b^{\text{MC}})^2 &= \mathbb{E}\left[\left|\mathbb{E}[Q] - \mathbb{E}[\hat{Q}_b^{\text{MC}}] + \mathbb{E}[\hat{Q}_b^{\text{MC}}] - \hat{Q}_b^{\text{MC}}\right|^2\right] \\
&= \left|\mathbb{E}[Q] - \mathbb{E}[\hat{Q}_b^{\text{MC}}]\right|^2 + \mathbb{E}\left[\left|\mathbb{E}[\hat{Q}_b^{\text{MC}}] - \hat{Q}_b^{\text{MC}}\right|^2\right] \\
&= \left|\mathbb{E}[Q] - \mathbb{E}[\hat{Q}_b^{\text{MC}}]\right|^2 + \mathbb{V}[\hat{Q}_b^{\text{MC}}],
\end{aligned} \tag{5.17}$$

where the second line follows from first due to the fact that  $\mathbb{E}[\hat{Q}_b^{\text{MC}} - \mathbb{E}[\hat{Q}_b^{\text{MC}}]] = 0$ , and the third line follows from the second by the definition of the variance. The first term in (5.17) is the ‘bias’ (i.e., the error introduced by the discretisation), and the second term in (5.17) is the variance of the estimator  $\hat{Q}_b^{\text{MC}}$ .

By definition of  $\hat{Q}_b^{\text{MC}}$ , and the fact that the samples  $\tilde{Q}_b^{(j)}$  are independent, we have

$$\mathbb{V}[\hat{Q}_b^{\text{MC}}] = \frac{1}{N_{\text{MC}}^2} \sum_{j=1}^{N_{\text{MC}}} \mathbb{V}[\tilde{Q}_b^{(j)}] = \frac{1}{N_{\text{MC}}} \mathbb{V}[\tilde{Q}_b]. \tag{5.18}$$

Therefore we can conclude from (5.17) and (5.18) that the root-mean-squared-error satisfies

$$\text{Err}(\hat{Q}_b^{\text{MC}})^2 = \left|\mathbb{E}[\tilde{Q}_b - Q]\right|^2 + \frac{1}{N_{\text{MC}}} \mathbb{V}[\tilde{Q}_b]. \tag{5.19}$$

By (5.13) and Lemma 5.7 the first term in (5.19) is proportional to  $\varepsilon^2/2$ , and by (5.14) the second term in (5.19) is proportional to  $\varepsilon^2/2$ , and therefore (5.15) holds. All that remains is to estimate the (expected) computational complexity. We have

$$\begin{aligned}
\mathbb{E}[\mathcal{C}(\hat{Q}_b^{\text{MC}})] &= N_{\text{MC}} \mathbb{E}[\mathcal{C}(\tilde{Q}_b)] \\
&\leq N_{\text{MC}} c_3 (b^{-\gamma} + k^{a\gamma}) \text{ by Lemma 5.6,} \\
&\sim 2\mathbb{V}[\tilde{Q}_b] \varepsilon^{-2} \left( c_3 (\sqrt{2}c_1)^{\frac{\gamma}{\alpha}} k^{\frac{\gamma\sigma}{\alpha}} \varepsilon^{-\frac{\gamma}{\alpha}} + k^{a\gamma} \right) \text{ by (5.13) and (5.14)}
\end{aligned}$$

as required. □

## 5.5 MULTI-LEVEL MONTE-CARLO METHODS

We now analyse the Multi-Level Monte-Carlo method in the  $k$ -dependent abstract setting given above. Aside from the  $k$ -dependence and the sample-dependent existence and uniqueness criterion (the latter of which has been discussed and dealt with through introducing the random variables  $\tilde{Q}_b$  above), our approach and final result is analogous to the standard Multi-Level Monte-Carlo complexity result given in, e.g., [49, Theorem 1]. Recall that the goal is to choose the number of levels  $L$  and the numbers of samples on each level  $N_l$  to achieve a root-mean-squared error of at most  $\varepsilon$  with minimal cost. Our main result, showing how to achieve this goal, is Theorem 5.13 below. We now give precise details of the setup for Multi-Level Monte-Carlo.

We define a set of levels  $\{h_l\}_{l=0}^L$  (with  $L$  to be chosen) such that

$$h_l = \frac{h_{l-1}}{s} \quad (5.20)$$

for  $s > 1$  and  $l \geq 1$ . In particular

$$h_L = s^{-L} h_0. \quad (5.21)$$

(Observe that when  $h_l$  corresponds to the mesh width of a finite-element mesh, then (5.20) is achieved if we obtain successive meshes by uniform refinement.) We then define the correction operators between the levels by

$$Y_l := \tilde{Q}_{h_l} - \tilde{Q}_{h_{l-1}}, l \geq 1, \quad Y_0 = \tilde{Q}_{h_0}. \quad (5.22)$$

Observe that by construction

$$\mathbb{E}[Y_0] + \sum_{l=1}^L \mathbb{E}[Y_l] = \mathbb{E} \left[ \tilde{Q}_{h_0} + \sum_{l=1}^L \tilde{Q}_{h_l} - \tilde{Q}_{h_{l-1}} \right] = \mathbb{E}[\tilde{Q}_{h_L}]. \quad (5.23)$$

We let  $\hat{Y}_l$  be the Monte-Carlo estimator of  $Y_l$ , i.e.,

$$\hat{Y}_l := \frac{1}{N_l} \sum_{j=1}^{N_l} Y_l^{(j)}, \quad (5.24)$$

with  $N_l$  to be chosen, where the  $Y_l^{(j)}$  are independent samples of  $Y_l$ . Note that it follows that the estimators  $\hat{Y}_l$  are independent of each other. (To simplify the notation, we do not include  $N_l$  in the notation for  $\hat{Y}_l$ .) Finally we define the *multi-level Monte Carlo estimator* of  $Q$

$$\hat{Q}_{h_L}^{\text{ML}} := \sum_{l=1}^L \hat{Y}_l.$$

As we discussed in Section 5.2.1 above, the reason the Multi-Level Monte-Carlo method delivers a lower computational cost than the Monte-Carlo method is that the variance of the estimators  $\hat{Y}_l$  decreases as  $l$  increases. Therefore the more expensive simulations (for higher  $l$ ) need fewer samples. To quantify the behaviour of these variances, we assume  $\mathbb{V}[Y_l]$  has the following property, c.f. the behaviour of the error in (5.3). (The similarity in the form of (5.25) below and (5.3) is no coincidence, one usually proves bounds of the form (5.25) via bounds of the form (5.3); see the proof of Lemma 5.19 below for an example of this proof technique.)

**Assumption 5.11** (Variance of correction operators). *There exist  $c_2, \beta, \tau > 0$ , such that  $c_2$  is independent of  $h$  and  $k$ , and*

$$V_l := \mathbb{V}[Y_l] \leq c_2 h_l^\beta k^\tau. \quad (5.25)$$

As we will see in Theorem 5.13 below, the interplay between  $\beta$  and  $\gamma$  (i.e., the interplay between the variances and the cost of computing a single sample) governs the behaviour of the

cost of the Multi-Level Monte-Carlo method.

We make the following simplifying assumption, that the coarse mesh  $h_0$  has the same  $k$ -dependence as the criterion for existence and uniqueness (5.2).

**Assumption 5.12** (Dependence of coarse space on  $k$ ). *Let  $C_{\text{coarse}} > 0$  be independent of  $k$  and*

$$h_0 = C_{\text{coarse}} k^{-a}. \quad (5.26)$$

We can now state our main theorem on the complexity of the Multi-Level Monte-Carlo method in the  $k$ -dependent abstract setting above. In particular, we show

- how the number of levels  $L$  should be chosen, and
- how the number of samples  $N_l$  on each level should be chosen

so that the root-mean-squared error of the Multi-Level Monte-Carlo estimator is of the order  $\varepsilon$  with minimal work. Observe that Theorem 5.13 is analogous to the standard Multi-Level Monte-Carlo complexity theorem, see, e.g., [49, Theorem 1], but adapted for our  $k$ -dependent setting.

We let

$$\mathcal{C}_l := c_3 (h_l^{-\gamma} + k^{a\gamma}), \quad (5.27)$$

i.e.,  $\mathcal{C}_l$  is the bound on the expected cost of computing one sample of  $\tilde{Q}_{h_l}$  (see Lemma 5.6).

**Theorem 5.13** (Computational Complexity of Multi-Level Monte-Carlo). *Under Assumptions 5.4, 5.5, 5.1, 5.12, and 5.11, if  $L$  is given by*

$$L = \max \left\{ \left\lceil \frac{1}{\alpha} \log_s \left( \sqrt{2} c_1 C_{\text{coarse}}^\alpha k^{\sigma - a\alpha} \varepsilon^{-1} \right) \right\rceil, 0 \right\}, \quad (5.28)$$

the number of samples on each computational level is given by

$$N_l = \left\lceil 2\varepsilon^{-2} \left( \frac{V_l}{\mathcal{C}_l} \right)^{\frac{1}{2}} \sum_{j=0}^L (V_j \mathcal{C}_j)^{\frac{1}{2}} \right\rceil, \quad (5.29)$$

$\varepsilon < 1$ , and

$$\alpha \geq \frac{1}{2} \min\{\beta, \gamma\},$$

then  $\text{Err}(\hat{Q}_{h_L}^{\text{ML}}) \leq \varepsilon$  and, if  $L \geq 1$ , the computational cost of  $\hat{Q}_{h_L}^{\text{ML}}$  satisfies

$$\mathbb{E} \left[ \mathcal{C}(\hat{Q}_{h_L}^{\text{ML}}) \right] \lesssim \begin{cases} (k^{\tau - a(\beta - \gamma)} + k^{\frac{\gamma\sigma}{\alpha}}) \varepsilon^{-2} & \text{if } \beta > \gamma, \\ k^\tau \varepsilon^{-2} \left( (\log_s(\varepsilon^{-1} k^{\sigma - a\alpha}))^2 + 1 \right) + k^{\frac{\gamma\sigma}{\alpha}} \varepsilon^{-2} & \text{if } \beta = \gamma, \\ k^{\tau + (\gamma - \beta)\frac{\sigma}{\alpha}} \varepsilon^{-2 - \frac{\gamma - \beta}{\alpha}} + k^{\frac{\gamma\sigma}{\alpha}} \varepsilon^{-2} & \text{if } \gamma > \beta. \end{cases} \quad (5.30)$$

However, if  $L = 0$ , then  $\mathcal{C}(\hat{Q}_{h_L}^{\text{ML}})$  is given by Theorem 5.10.

The proof of Theorem 5.13 is given on page 215 below. It is surprising that in (5.28) we must take the maximum to ensure  $L$  is non-negative; this requirement is due to a subtle point about the values of  $a$ ,  $\alpha$ , and  $\sigma$ , see Section 5.6.1 below. The assumption that  $\alpha \geq \min\{\beta, \gamma\}/2$  is standard in studies of Multi-Level Monte-Carlo methods, in order to simplify the expressions involving  $\varepsilon$  in (the equivalent results to) (5.30), see, e.g. [49, Theorem 1].

**Remark 5.14** (The finest mesh size in Theorem 5.13). *Observe that if the number of additional levels  $L$  is given by (5.28), then one can simplify the dependence on  $C_{\text{coarse}}$  and  $a$  using (5.21) and (5.26) to obtain*

$$h_L = \min \left\{ \left( \frac{\varepsilon}{\sqrt{2}c_1 k^\sigma} \right)^{\frac{1}{\alpha}}, h_0 \right\}. \quad (5.31)$$

In the proof of Theorem 5.13, we will need to bound sums of the form  $\sum_{l=0}^L s^{\delta l}$ , where  $L$  is given by (5.28) and  $\delta$  is some constant. Therefore, we first prove these bounds in the following lemma, that contains an abstract version of (5.28), before proceeding to the proof of Theorem 5.13.

**Lemma 5.15** (Bounds on sums occurring in the proof of Theorem 5.13). *If  $L$  is given by*

$$L = \lceil C_L \log_s(\eta \varepsilon^{-1}) \rceil, \quad (5.32)$$

*for some  $C_L, \eta > 0$ , then, for  $s > 1$  and  $\delta \in \mathbb{R}$ , we have the bounds*

$$\sum_{l=0}^L s^{\delta l} \leq \begin{cases} L+1 & \text{if } \delta = 0, \\ \frac{s^\delta}{1-s^{-\delta}} \eta^{\delta C_L} \varepsilon^{-\delta C_L} & \text{if } \delta > 0, \\ \frac{s^{-\delta}}{s^{-\delta}-1} & \text{if } \delta < 0. \end{cases} \quad (5.33)$$

*Proof of Lemma 5.15.* The case  $\delta = 0$  is immediate. For  $\delta \neq 0$  the proof follows that in [49, Appendix A]. We first observe that, for  $\delta > 0$ ,

$$\begin{aligned} \sum_{l=0}^L s^{\delta l} &= \frac{s^{\delta(L+1)} - 1}{s^\delta - 1} \\ &= \frac{s^{\delta L} - s^{-\delta}}{1 - s^{-\delta}} \\ &\leq \frac{s^{\delta L}}{1 - s^{-\delta}}, \end{aligned} \quad (5.34)$$

since  $s^{\delta L} \geq s^{-\delta}$ .

Then, since  $L$  is given by (5.32), it follows that the bound

$$L < C_L \log_s(\eta \varepsilon^{-1}) + 1 \quad (5.35)$$

holds. Rearranging (5.35), we obtain the bound

$$s^L < (\eta\varepsilon^{-1})^{C_L} s. \quad (5.36)$$

From (5.36) we can then obtain

$$s^{\delta L} < \eta^{\delta C_L} \varepsilon^{-\delta C_L} s^{\delta}. \quad (5.37)$$

Combining (5.34) and (5.37), we obtain (5.33) in the case  $\delta > 0$ .

For the case  $\delta < 0$ , we observe

$$\begin{aligned} \sum_{l=0}^L s^{\delta l} &= \frac{s^{\delta(L+1)} - 1}{s^{\delta} - 1} \\ &= \frac{s^{-\delta} - s^{\delta L}}{s^{-\delta} - 1} \\ &\leq \frac{s^{-\delta}}{s^{-\delta} - 1}, \end{aligned}$$

since  $s^{-\delta} \geq s^{\delta L}$ , that is, (5.33) in the case  $\delta < 0$ .  $\square$

We are now in a position to prove Theorem 5.13.

*Proof of Theorem 5.13.* Throughout the proof, we assume  $L > 0$ . In the case  $L = 0$ , the Multi-Level Monte-Carlo estimator becomes the Monte-Carlo estimator, whose behaviour is given by Theorem 5.10.

We recall the bias–variance decomposition of the (squared) mean-squared error analogous to (5.17)

$$\text{Err}(\hat{Q}_{h_L}^{\text{ML}})^2 = \left| \mathbb{E}[\hat{Q}_{h_L}^{\text{ML}} - Q] \right|^2 + \mathbb{V}[\hat{Q}_{h_L}^{\text{ML}}], \quad (5.38)$$

where the first term in (5.38) is the bias and the second term is the variance. We now proceed to choose the parameters  $L$  and  $N_l$ ,  $l = 0, \dots, L$ , such that we can bound both the bias and the variance by  $\varepsilon^2/2$ , thereby making  $\text{Err}(\hat{Q}_{h_L}^{\text{ML}}) \leq \varepsilon$ .

We first bound the bias. To do this, we only need to choose  $L$  large enough, i.e., choose  $h_L$  small enough. By the construction of the Multi-Level Monte-Carlo estimator  $\hat{Q}_{h_L}^{\text{ML}}$ , it follows that  $\mathbb{E}[\hat{Q}_{h_L}^{\text{ML}}] = \mathbb{E}[\tilde{Q}_{h_L}]$ , see (5.23). Therefore the bias term in (5.38) is equal to  $\left| \mathbb{E}[\tilde{Q}_{h_L} - Q] \right|^2$ . By Lemma 5.7 with  $h = h_L$ , a sufficient condition for the bias term to be at most  $\varepsilon^2/2$  is

$$c_1 k^\sigma h_L^\alpha \leq \frac{\varepsilon}{\sqrt{2}}, \quad (5.39)$$

which, when rearranged, gives the first term in (5.31). As  $h_L = h_0 s^{-L}$ , it follows from rearranging (5.39) that a sufficient condition for the bias term to be  $\leq \varepsilon^2/2$  is

$$L = \left\lceil \frac{1}{\alpha} \log_s \left( \sqrt{2} c_1 k^\sigma h_0^\alpha \varepsilon^{-1} \right) \right\rceil. \quad (5.40)$$

Under Assumption 5.12, since  $h_0 = C_{\text{coarse}} k^{-a}$ , we can simplify (5.40) to obtain the first term in (5.28), as required.

We now seek to bound the variance term in (5.38) with minimal cost. I.e., we choose the numbers of samples  $N_l$  such that the variance term is at most  $\varepsilon^2/2$  and the computational cost is minimised. Similar to the expression (5.18) for the variance of the Monte-Carlo estimator, one can show that the variance of the Multi-Level Monte-Carlo estimator is given by

$$\mathbb{V}[\hat{Q}_{b_L}^{\text{ML}}] = \sum_{l=0}^L \frac{V_l}{N_l}, \quad (5.41)$$

and the expected cost of  $\hat{Q}_{b_L}^{\text{ML}}$  is: (following [104])

$$\begin{aligned} \mathbb{E}[\mathcal{C}(\hat{Q}_{b_L}^{\text{ML}})] &\leq \sum_{l=0}^L \mathbb{E}[\mathcal{C}(\hat{Y}_l)] \\ &= \sum_{l=0}^L \sum_{j=1}^{N_l} \mathbb{E}[\mathcal{C}(Y_l^{(j)})] \text{ by the definition of } \hat{Y}_l \text{ (5.24),} \\ &\leq \sum_{l=0}^L \sum_{j=0}^{N_l} (\mathbb{E}[\mathcal{C}(\tilde{Q}_{b_l})] + \mathbb{E}[\mathcal{C}(\tilde{Q}_{b_{l-1}})]) \text{ by the definition of } Y_l \text{ (5.22),} \\ &= \sum_{l=0}^L N_l (1 + s^{-\gamma}) c_3 (b_l^{-\gamma} + k^{a\gamma}) \text{ by Lemma 5.6,} \\ &= (1 + s^{-\gamma}) \sum_{l=0}^L N_l \mathcal{C}_l, \end{aligned} \quad (5.42)$$

by the definition of  $\mathcal{C}_l$ , (5.27).

We now find an optimal number of samples for each level. To find this optimal number of samples we formulate this task as an optimisation problem: Find  $N_0, N_1, \dots, N_L > 0$  to minimise (5.42) subject to

$$\sum_{l=0}^L \frac{V_l}{N_l} = \frac{\varepsilon^2}{2}.$$

This is exactly the formulation used in [95, Section 1.3], and therefore as in [95, Section 1.3] we can use a Lagrange multiplier to solve this minimisation problem, resulting in the values of  $N_l$  as defined in (5.29). (The ceiling function in (5.29) is introduced because the values of  $N_l$  solving the optimisation problem may not be integers, however, the number of samples in the Multi-Level Monte-Carlo method must be integers. Increasing the optimal values of  $N_l$  slightly (by using the ceiling) will decrease the variance (as the variance is given by (5.41)), and so we will still have  $\mathbb{V}[\hat{Q}_{b_L}^{\text{ML}}] \leq \varepsilon^2/2$ .)

We now infer the expected computational complexity of the Multi-Level Monte-Carlo method with  $L$  given by (5.28) and the  $N_l$  given by (5.29). To simplify the calculation, we first bound  $\mathcal{C}_l$  purely in terms of  $b_l$ , rather than  $b_l$  and  $k$ , as in Lemma 5.6. From Lemma 5.6 we have

$\mathcal{C}_l \leq c_3(h_l^{-\gamma} k^{a\gamma})$ , and therefore

$$\begin{aligned} \mathcal{C}_l &\leq c_3(h_l^{-\gamma} + C_{\text{coarse}}^\gamma h_0^{-\gamma}) \quad \text{by Assumption 5.12} \\ &= c_3(h_l^{-\gamma} + C_{\text{coarse}}^\gamma (h_l s^l)^{-\gamma}) \quad \text{by definition of } h_l \text{ (5.20)} \\ &= c_3(1 + C_{\text{coarse}}^\gamma s^{-\gamma l}) h_l^{-\gamma} \\ &\leq c_3(1 + C_{\text{coarse}}^\gamma) h_l^{-\gamma} \end{aligned}$$

because  $\gamma > 0$  and  $s > 1$ , so  $s^{-\gamma l} < 1$ .

We can now bound the expected computational complexity. From the expression (5.42), we have

$$\begin{aligned} \mathbb{E}[\mathcal{C}(\hat{Q}_{h_L}^{\text{ML}})] &\leq (1 + s^{-\gamma}) \sum_{l=0}^L \mathcal{C}_l N_l \\ &\leq (1 + s^{-\gamma}) \sum_{l=0}^L \mathcal{C}_l \left( \frac{2}{\varepsilon^2} \left( \frac{V_l}{\mathcal{C}_l} \right)^{\frac{1}{2}} \sum_{j=0}^L (V_j \mathcal{C}_j)^{\frac{1}{2}} + 1 \right) \quad \text{(by (5.29))}, \\ &= 2\varepsilon^{-2} (1 + s^{-\gamma}) \left( \sum_{l=0}^L (V_l \mathcal{C}_l)^{\frac{1}{2}} \right)^2 + (1 + s^{-\gamma}) \sum_{l=0}^L \mathcal{C}_l \\ &= 2c_2 c_3 (1 + C_{\text{coarse}}^\gamma) (1 + s^{-\gamma}) k^\tau \varepsilon^{-2} \left( \sum_{l=0}^L h_l^{\frac{\beta-\gamma}{2}} \right)^2 \\ &\quad + c_3 (1 + C_{\text{coarse}}^\gamma) (1 + s^{-\gamma}) \sum_{l=0}^L h_l^{-\gamma} \\ &\quad \text{(by Assumptions 5.11 and 5.12 and Lemma 5.6),} \\ &= 2c_2 c_3 (1 + C_{\text{coarse}}^\gamma) (1 + s^{-\gamma}) k^\tau \varepsilon^{-2} h_0^{\beta-\gamma} \left( \sum_{l=0}^L s^{l(\frac{\gamma-\beta}{2})} \right)^2 \\ &\quad + c_3 (1 + C_{\text{coarse}}^\gamma) (1 + s^{-\gamma}) h_0^{-\gamma} \sum_{l=0}^L s^{\gamma l}, \end{aligned} \tag{5.43}$$

by definition of  $h_l$ .

We now bound the two sums in (5.43) using Lemma 5.15. Using Lemma 5.15 with  $C_L = 1/\alpha$ ,  $\eta = \sqrt{2}c_1 C_{\text{coarse}}^\alpha k^{\sigma-a\alpha}$ , and  $\delta = \gamma > 0$ , the second term in (5.43) can be bounded by

$$\begin{aligned} &c_3 (1 + C_{\text{coarse}}^\gamma) \frac{(1 + s^{-\gamma}) h_0^{-\gamma} s^\gamma (\sqrt{2}c_1)^{\frac{\gamma}{\alpha}} C_{\text{coarse}}^\gamma}{1 - s^{-\gamma}} k^{\frac{\gamma\sigma}{\alpha} - a\gamma} \varepsilon^{-\frac{\gamma}{\alpha}} \\ &= \frac{(1 + s^{-\gamma}) c_3 (\sqrt{2}c_1)^{\frac{\gamma}{\alpha}} s^\gamma (1 + C_{\text{coarse}}^\gamma)}{1 - s^{-\gamma}} k^{\frac{\gamma\sigma}{\alpha}} \varepsilon^{-\frac{\gamma}{\alpha}} \\ &\leq \frac{(1 + s^{-\gamma}) c_3 (\sqrt{2}c_1)^{\frac{\gamma}{\alpha}} s^\gamma (1 + C_{\text{coarse}}^\gamma)}{1 - s^{-\gamma}} k^{\frac{\gamma\sigma}{\alpha}} \varepsilon^{-2}, \end{aligned} \tag{5.44}$$

since  $\alpha \geq \gamma/2$ .

To bound the first sum in (5.43), we must distinguish three cases,  $\gamma = \beta$ ,  $\gamma > \beta$ , and  $\gamma < \beta$ .

If  $\gamma = \beta$ , then the first part of (5.43) becomes (using Lemma 5.15 with  $C_L$  and  $\eta$  as above, and  $\delta = 0$  and (5.28))

$$\begin{aligned}
& 2c_2c_3(1 + C_{\text{coarse}}^\gamma)(1 + s^{-\gamma})k^\tau \varepsilon^{-2}(L + 1)^2 \\
& \leq 2c_2c_3(1 + C_{\text{coarse}}^\gamma)(1 + s^{-\gamma})k^\tau \varepsilon^{-2} \left( \frac{1}{\alpha} \log_s(\varepsilon^{-1} \sqrt{2}c_1 C_{\text{coarse}}^\alpha k^{\sigma - a\alpha}) + 2 \right)^2 \\
& = 2c_2c_3(1 + C_{\text{coarse}}^\gamma)(1 + s^{-\gamma})k^\tau \varepsilon^{-2} \left( \frac{1}{\alpha} (\log_s(\varepsilon^{-1} k^{\sigma - a\alpha}) + \log_s(\sqrt{2}c_1 C_{\text{coarse}}^\alpha) + 2) \right)^2 \\
& \lesssim k^\tau \varepsilon^{-2} \left( (\log_s(\varepsilon^{-1} k^{\sigma - a\alpha}))^2 + 1 \right) \tag{5.45}
\end{aligned}$$

In the case  $\gamma > \beta$ , to simplify the notation, we let

$$C_{\text{sum}, \delta} := \frac{(\sqrt{2}c_1)^{\frac{\delta}{\alpha}} C_{\text{coarse}}^\delta}{1 - s^{-\delta}}.$$

Then using Lemma 5.15 with  $C_L$  and  $\eta$  as above, but  $\delta = (\gamma - \beta)/2 > 0$ , the first term in (5.43) becomes

$$\begin{aligned}
& \varepsilon^{-2} 2c_2c_3(1 + C_{\text{coarse}}^\gamma)(1 + s^{-\gamma})k^\tau \varepsilon^{-2} h_0^{\beta - \gamma} \left( C_{\text{sum}, \frac{\gamma - \beta}{2}} s^{\frac{\gamma - \beta}{2}} k^{\frac{\gamma - \beta}{2} \frac{\sigma}{\alpha}} k^{-a \frac{\gamma - \beta}{2}} \varepsilon^{-\frac{\gamma - \beta}{2\alpha}} \right)^2 \\
& = C_{\gamma > \beta} k^{\tau + (\gamma - \beta) \frac{\sigma}{\alpha}} \varepsilon^{-2 - \frac{\gamma - \beta}{\alpha}}, \tag{5.46}
\end{aligned}$$

where

$$C_{\gamma > \beta} := 2c_2c_3(1 + C_{\text{coarse}}^\gamma)(1 + s^{-\gamma}) C_{\text{sum}, \frac{\gamma - \beta}{2}}^2 s^{\gamma - \beta} C_{\text{coarse}}^{\beta - \gamma};$$

and the second equality in (5.46) follows from the definition of  $h_0$  in Assumption 5.12.

If  $\gamma < \beta$ , then using Lemma 5.15 with  $C_L$  and  $\eta$  as above, but with  $\delta = (\gamma - \beta)/2 < 0$ , the first term in (5.43) is

$$C_{\gamma < \beta} k^{\tau - a(\beta - \gamma)} \varepsilon^{-2}, \tag{5.47}$$

where

$$C_{\gamma < \beta} := 2c_2c_3(1 + C_{\text{coarse}}^\gamma)(1 + s^{-\gamma}) C_{\text{coarse}}^{\beta - \gamma} \frac{s^{\beta - \gamma}}{\left( s^{\frac{\beta - \gamma}{2}} - 1 \right)^2}.$$

We now combine (5.43)–(5.47) and suppress all the constants to obtain (5.30).  $\square$

## 5.6 PLACING THE STOCHASTIC HELMHOLTZ EQUATION IN THE ABSTRACT $k$ -DEPENDENT SETTING

We now show that the stochastic Helmholtz equation fits into the abstract  $k$ -dependent setting given above. We use the abstract results on the computational complexity of Monte-Carlo and Multi-Level Monte-Carlo methods in Theorems 5.10 and 5.13 to derive fully  $k$ -explicit complexity

bounds for Monte-Carlo and Multi-Level Monte-Carlo methods for the stochastic Helmholtz equation, given in Theorem 5.22.

### 5.6.1 Model problem and quantities of interest

We let  $u : \Omega \rightarrow H_k^1(D)$  solve the TEDP-analogue of Problem 3.1 (see Remark 3.13), and let  $\tilde{u}_b : \Omega \rightarrow V_{b,p}$  solve the stochastic analogue of Problem 2.20. (I.e.,  $\tilde{u}_b$  solves Problem 2.20 sample-wise with coefficients  $A(\omega)$  and  $n(\omega)$ ,  $T = ik$ , and meshsize  $h_\omega$ .) We assume  $\tilde{u}_b$  is measurable, see Remark 5.3. Further, we assume that the stochastic Helmholtz equation is nontrapping almost surely, i.e., the TEDP-analogues of Conditions 3.6 and 3.8 and Theorem 3.10 hold. We consider two quantities of interest (QoIs) of the solution  $u$ ; the two norms  $\|u\|_{L^2(D)}$  and  $\|u\|_{H_k^1(D)}$ , where  $D$  is the computational domain.

**Remark 5.16** (Why consider these QoIs?). *We consider the norms  $\|u\|_{L^2(D)}$  and  $\|u\|_{H_k^1(D)}$  as QoIs because the Helmholtz equation is an elliptic PDE, and therefore it is natural to consider terms depending on  $u$  and  $\nabla u$  (and these are, arguably, the simplest such terms). Moreover, we expect different  $k$ -dependence of the computational complexity for QoIs involving  $u$  compared to QoIs involving  $\nabla u$  (c.f., Theorem 2.39 and Assumption 5.1). Considering both  $\|u\|_{L^2(D)}$  and  $\|u\|_{H_k^1(D)}$  as QoIs will allow us to see if this is the case.*

#### The values of $\alpha$ , $\sigma$ , $\beta$ , $\tau$ , $\gamma$ , and $a$

For the two QoIs  $\|u\|_{L^2(D)}$  and  $\|u\|_{H_k^1(D)}$ , the provable values of  $\alpha$ ,  $\sigma$ ,  $\beta$ , and  $\tau$  are given in Lemmas 5.19 and 5.20 below. Their values are obtained straightforwardly from Theorem 2.39 above. However, determining the value of  $\gamma$ , and especially the value of  $a$ , is more involved. In addition, we note that in practice the value of, in particular,  $\alpha$  may be larger than predicted by the theory, see, e.g., [49, Section 4]. In this section, however, we will work with the provable values.

*The value of  $\gamma$*  The value of  $\gamma$  represents the efficiency of the solver one uses to solve the linear systems arising from the finite-element discretisations of the individual Helmholtz problems. Recall that the number of degrees of freedom in the linear systems is of the order  $h^{-d}$  (if the mesh size for the finite-element mesh is  $h$ ). In the following analysis we take  $\gamma = d$ , i.e. we assume that we have access to an optimal Helmholtz solver, that can solve linear systems with  $N$  unknowns arising from finite-element discretisations of Helmholtz problems in  $\mathcal{O}(N)$  time. Obtaining such a solver is the subject of much current research, and we refer to, e.g., the recent works [102, 221, 203] for a selection of modern solvers achieving close to this optimal scaling.

*The values of  $a$*  In our analysis below, we consider two different values of  $a$ ,  $a = (2p + 1)/2p$  (where  $p$  is the polynomial degree of the finite-elements) and  $a = 1$ . We now explain why  $a = (2p + 1)/2p$  is the natural choice, but has some limitations in the Multi-Level Monte-Carlo method. We then go on to explain how the choice  $a = 1$  removes these limitations, and we discuss whether the choice  $a = 1$  is reasonable.

The first choice of  $a$  is motivated by the finite-element results in Theorem 2.39 above. Recall from (2.62) that if  $hk^{(2p+1)/2p}$  is sufficiently small (if  $C_{\text{stab}} \sim 1$ , with hidden constant dependent on  $A$  and  $n$ ), then the finite-element solution  $u_b$  exists, is unique, and satisfies the error bounds in both the  $L^2$ - and  $H_k^1$ -norms. Therefore, since Assumption 5.1 is concerned with existence, uniqueness, and error bounds for  $Q_b$ , the choice  $a = (2p + 1)/2p$  is natural.

However, certain choices of  $a$  and  $Q$  mean that the number of levels  $L$  will not grow with  $k$ . In (5.28) above,  $L$  depends on  $k^{\sigma-a\alpha}$ ; i.e., the  $k$ -dependence of  $L$  is governed by the relationship between  $a$  (i.e., the  $k$ -dependence of the coarsest level) and  $\sigma/\alpha$  (the  $k$ -dependence of the finest level—see (5.31)). Taking  $a = (2p + 1)/2p$  and  $Q = \|u\|_{H_k^1(D)}$ , so that  $\alpha = 2p$  and  $\sigma = 2p + 1$  (see Lemma 5.19 below for details of why these values for  $\alpha$  and  $\sigma$  are correct) then  $k^{\sigma-a\alpha} = 1$ , and therefore  $L$  is  $k$ -independent.

It may, however, be interesting to study the case where the number of levels  $L$  increases with  $k$ . If we instead choose  $a = 1$  (i.e., the condition for existence and uniqueness (5.2) simply requires a fixed number of points per wavelength), then we would have  $k^{\sigma-a\alpha} = k$ , and so the number of levels  $L$  would increase with  $k$ .

Therefore, the question arises, ‘Is the choice  $a = 1$  (with  $\alpha$  and  $\sigma$  still given by Theorem 2.39) reasonable?’

In 1-d, the answer is ‘yes’. In [121, Corollary 3.2] and [118, Theorem 4.27 and equation (4.7.41)] Ihlenburg and Babuška prove that the  $h$ -finite-element method for the homogeneous Helmholtz equation in 1-d is  $(hk^1, hk^{(2p+1)/2p})$ -accurate; i.e., finite-element error bounds of a form similar to those in Theorem 2.39 hold if  $hk$  is sufficiently small. Translated into the multi-level context, this result implies that for  $d = 1$ , Assumption 5.1 holds with  $a = 1$ . We note that this result has *not* been proved in higher dimensions (see the discussion in Section 2.3.3). However, we will assume that Assumption 5.1 holds in higher dimensions with  $a = 1$ ; i.e., we make the following assumptions, and we will prove results on the computational complexity of Monte-Carlo and Multi-Level Monte-Carlo under these assumptions (as well as when  $a = (2p + 1)/2p$ ).

**Assumption 5.17** (Assumptions for  $Q(u) = \|u\|_{H_k^1(D)}$  with  $a = 1$ ). *In the setting given at the beginning of Section 5.6.1, if  $Q(u) = \|u\|_{H_k^1(D)}$ , then Assumptions 5.1 and 5.11 hold with  $a = 1$ ,  $\alpha = 2p$ ,  $\sigma = 2p + 1$ ,  $\beta = 4p$ , and  $\tau = 4p + 2$ , for some random variables  $C_1$ ,  $\tilde{c}_1$ , and  $c_2$ .*

**Assumption 5.18** (Assumptions for  $Q(u) = \|u\|_{L^2(D)}$  with  $a = 1$ ). *In the setting given at the beginning of Section 5.6.1, if  $Q(u) = \|u\|_{L^2(D)}$ , then Assumptions 5.1 and 5.11 hold with  $a = 1$ ,  $\alpha = 2p$ ,  $\sigma = 2p$ ,  $\beta = 4p$ , and  $\tau = 4p$ , for some random variables  $C_1$ ,  $\tilde{c}_1$ , and  $c_2$ .*

## 5.6.2 Main result on Monte-Carlo and Multi-Level Monte-Carlo methods for the Helmholtz equation

We are now in a position to state our main result on the computational complexity of Monte-Carlo and Multi-Level Monte-Carlo methods applied to the Helmholtz equation, Theorem 5.22 below. We first verify Assumption 5.1 for each of our QoIs in the following two lemmas.

**Lemma 5.19** (Verifying assumptions for  $Q(u) = \|u\|_{H_k^1(D)}$ ). *In the setting given at the beginning of Section 5.6.1, if  $Q(u) = \|u\|_{H_k^1(D)}$ , then Assumptions 5.1 and 5.11 hold with  $a = (2p + 1)/2p$ ,  $\alpha = 2p$ ,  $\sigma = 2p + 1$ ,  $\beta = 4p$ ,  $\tau = 4p + 2$ , and  $C_1$  and  $\tilde{c}_1$  given by the constants in (2.62) and (2.64) respectively, and  $c_2 = \mathbb{E}[\tilde{c}_1^2](1 + s^\alpha)^2$ .*

*Proof of Lemma 5.19.* By the assumptions of this lemma, it is immediate from (2.64) that Assumption 5.1 holds with  $\alpha = 2p$  and  $\sigma = 2p + 1$ . (See Remark 2.40 for why we can neglect the lower-order terms in (2.64).) To show Assumption 5.11, we follow [42, Proof of Proposition 4.2] and use the triangle inequality and Assumption 5.1 to show

$$\left| \hat{Y}_l(\omega) \right| \leq \left| (\tilde{Q}_{b_l} - Q)(\omega) \right| + \left| (Q - \tilde{Q}_{b_{l-1}})(\omega) \right| \leq \tilde{c}_1(\omega)(h_l^\alpha + h_{l-1}^\alpha)k^\sigma C_{f,gl}. \quad (5.48)$$

We then use (5.48) and the fact that  $\mathbb{V}[\hat{Y}_l] = \mathbb{E}\left[|\hat{Y}_l|^2\right] - \left|\mathbb{E}[\hat{Y}_l]\right|^2 \leq \mathbb{E}\left[|\hat{Y}_l|^2\right]$  to show (5.25), with  $c_2 = \mathbb{E}[\tilde{c}_1^2](1 + s^\alpha)^2$ .  $\square$

**Lemma 5.20** (Verifying assumptions for  $Q(u) = \|u\|_{L^2(D)}$ ). *In the setting given at the beginning of Section 5.6.1, if  $Q(u) = \|u\|_{L^2(D)}$ , then Assumptions 5.1 and 5.11 hold with  $a = (2p + 1)/2p$ ,  $\alpha = 2p$ ,  $\sigma = 2p$ ,  $\beta = 4p$ ,  $\tau = 4p$ , and  $C_1$  and  $\tilde{c}_1$  given by the constants in (2.62) and (2.63) respectively, and  $c_2 = \mathbb{E}[\tilde{c}_1^2](1 + s^\alpha)^2$ .*

*Proof of Lemma 5.20.* The proof is exactly analagous to the proof of Lemma 5.19, except we use (2.63) instead of (2.64).  $\square$

We require the following assumption on the variance of the approximations  $\tilde{Q}_b$ . Such an assumption is standard, see, e.g., [49, Text below equation (3)].

**Assumption 5.21.** *The variance  $\mathbb{V}[\tilde{Q}_b]$  is constant with respect to  $h$ .*

**Theorem 5.22** (Computational complexity of Monte-Carlo and Multi-Level Monte-Carlo methods for the Helmholtz equation). *Suppose Assumptions 5.4, 5.5, and 5.12 (on the cost of one realisation of  $\tilde{Q}_b$ , on the integrability of a combination of constants related to the size of  $\Omega_{\text{bad}}$ , and on the coarse space) and Assumption 5.21 hold.*

1. *If the assumptions of Lemmas 5.19 and 5.20 hold, and  $c_2$  as defined in Lemmas 5.19 and 5.20 is finite, then the Monte-Carlo and Multi-Level Monte-Carlo methods achieve a root-mean-squared error of at most  $\varepsilon$ , and their computational complexity (up to factors independent of  $h$  and  $k$ ) is given by the first two lines of Table 5.1, where ‘ $k\varepsilon$  small’ means*

$$k\varepsilon < \sqrt{2}c_1 C_{\text{coarse}}^{2p}. \quad (5.49)$$

2. *If Assumptions 5.17 and 5.18 hold instead of the assumptions of Lemmas 5.19 and 5.20, then the Monte-Carlo and Multi-Level Monte-Carlo methods achieve a root-mean-squared error of at most  $\varepsilon$ , and their computational complexity (up to factors independent of  $h$  and  $k$ ) is given by the last two lines of Table 5.1.*

| $Q(u)$             | $a$               | Monte-Carlo                                          | Multi-Level Monte-Carlo                                                                                           |
|--------------------|-------------------|------------------------------------------------------|-------------------------------------------------------------------------------------------------------------------|
| $\ u\ _{H_k^1(D)}$ | $\frac{2p+1}{2p}$ | $k^{d\frac{2p+1}{2p}} \varepsilon^{-2-\frac{d}{2p}}$ | $k^{d\frac{2p+1}{2p}} \varepsilon^{-2}$                                                                           |
| $\ u\ _{L^2(D)}$   | $\frac{2p+1}{2p}$ | $k^{d\frac{2p+1}{2p}} \varepsilon^{-2-\frac{d}{2p}}$ | $k^d \varepsilon^{-2}$ if $k\varepsilon$ small,<br>otherwise $k^{d\frac{2p+1}{2p}} \varepsilon^{-2-\frac{d}{2p}}$ |
| $\ u\ _{H_k^1(D)}$ | 1                 | $k^{d\frac{2p+1}{2p}} \varepsilon^{-2-\frac{d}{2p}}$ | $k^{d+2} \varepsilon^{-2}$                                                                                        |
| $\ u\ _{L^2(D)}$   | 1                 | $k^d \varepsilon^{-2-\frac{d}{2p}}$                  | $k^d \varepsilon^{-2}$                                                                                            |

Table 5.1: Computational complexity of Monte-Carlo and Multi-Level Monte-Carlo algorithms

The proof of Theorem 5.22 is given on page 224 below.

We now discuss the results in Theorem 5.22. The results for Multi-Level Monte-Carlo methods are consistently better than those for Monte-Carlo methods in terms of  $\varepsilon$ -dependence, unless the condition for existence and uniqueness of  $u_h$  is more restrictive than the condition to keep the error bounded<sup>2</sup>. In such a case, for  $\varepsilon$  small and/or  $k$  small, Multi-Level Monte-Carlo outperforms Monte-Carlo, but if  $\varepsilon$  and/or  $k$  are large, then Multi-Level Monte-Carlo is identical to Monte-Carlo (because there are no additional levels, i.e., in (5.28) the first term in the maximum is negative, and so  $L = 0$ ).

However, the  $k$ -dependence of the Multi-Level Monte-Carlo and Monte-Carlo methods (and which method has the more favourable  $k$ -dependence) is more complicated, and so we discuss it in more detail.

*Identical  $k$ -dependence* Observe that in the cases (i)  $a = (2p+1)/2p$  and  $Q(u) = \|u\|_{H_k^1(D)}$  and (ii)  $a = 1$  and  $Q(u) = \|u\|_{L^2(D)}$ , the  $k$ -dependence of the Monte-Carlo and Multi-Level Monte-Carlo methods is the same. This is unsurprising; in each case the criterion on the coarse space (5.26) has the same  $k$ -dependence as the definition of  $h_L$  (5.31), since  $a = \sigma/\alpha$ . Consequently, the number of levels  $L$  is independent of  $k$  (see also that the factor  $k^{\sigma-a\alpha}$  in (5.28) is  $k$ -independent in each of these cases). Since the number of levels is independent of  $k$ , and the  $k$ -dependence of the coarse and fine levels is the same, it is unsurprising that the computational complexity of the two estimators has the same  $k$ -dependence.

However, in the other two cases, the  $k$ -dependence of the complexity of the Multi-Level Monte-Carlo method is different to that of the Monte-Carlo method. We consider each of these cases in turn.

*When Multi-Level Monte-Carlo has better  $k$ -dependence* In the case  $a = (2p+1)/2p$  and  $Q(u) = \|u\|_{L^2(D)}$ , we see that the  $k$ -dependence of the Multi-Level Monte-Carlo method is *better* than that

<sup>2</sup>In the cases we consider, this scenario only occurs when we take  $a = (2p+1)/2p$  and  $Q(u) = \|u\|_{L^2(D)}$ , so  $\alpha = \sigma = 2p$ .

of the Monte-Carlo method. To understand this improvement, observe that the  $k$ -dependence of the criterion (5.26) on the coarse space ( $h_0 \lesssim k^{-(2p+1)/2p}$ ) is *more* restrictive than the  $k$ -dependence one would otherwise impose to ensure the bias error is small ( $h_L^{2p} k^{2p}$  is sufficiently small). Therefore, if we take  $h_L \sim k^{-(2p+1)/2p}$ , (to satisfy the coarse space requirement) the bias error (of the order  $h_L^{2p} k^{2p} = k^{-1}$ ) will *decrease* as  $k$  increases. Moreover, the variance of the Multi-Level Monte-Carlo estimator (given by (5.41)) will also decrease as  $k$  increases. Even on the coarsest level, the variance will be of the order  $h_0^{4p} k^{4p} = k^{-2}$  (see Assumption 5.11). Therefore, because the variance on each level decreases as  $k$  increases, the number of samples on each level will also decrease as  $k$  increases, reducing the overall computational cost in a  $k$ -dependent way.

*When Multi-Level Monte-Carlo has worse  $k$ -dependence* Conversely, in the case  $a = 1$  and  $Q(u) = \|u\|_{H_k^1(D)}$ , we see that the  $k$ -dependence of the cost of the Multi-Level Monte-Carlo estimator is *worse* than that of the Monte-Carlo estimator. The reason for this worse dependence is, in essence, the converse of the reason for the improved dependence in the discussion above. The difference between the coarse space (of the order  $k^{-1}$ ) and the fine space (with  $h_L$  of the order  $k^{-(2p+1)/2p}$ , see (5.31)) increases as  $k$  increases, and therefore the number of levels  $L$  will increase as  $k$  increases (see (5.28), and observe that in this case  $k^{\sigma-a\alpha} = k$ ). Moreover, on any level  $l$  where  $h_l \gtrsim k^{-(2p+1)/2p}$  (i.e., not the finest level), the variance  $V_l$  will increase as  $k$  increases, since  $V_l \sim h_l^{4p} k^{4p+1}$ . Therefore, on each level the variance (and thus the number of samples) will increase as  $k$  increases, resulting in an overall  $k$ -dependent increase in the computational cost.

It remains to be seen how these theoretical predictions are borne out in numerical computations; such computations should be the subject of future research.

**Remark 5.23** (Proving probabilistic bounds on the cost). *In [104], the authors extend their bounds on the expectation of the computational cost for Monte-Carlo and Multi-Level Monte-Carlo methods for the radiative transport equation to bounds on the exceedance probabilities of the computational cost. I.e., they prove bounds of the form*

$$\mathbb{P}(\mathcal{C}(\hat{Q}) < M(\varepsilon, \delta, \hat{Q})) > 1 - \delta^2, \quad (5.50)$$

for some function  $M$ , where  $\hat{Q}$  is the Monte-Carlo or Multi-Level Monte-Carlo estimator (see [104, Theorems 5.12 and 5.13]). They make only mild additional assumptions on the randomness to prove bounds of the form (5.50); these assumptions mean they can bound  $\mathbb{V}[\tilde{Q}_b]$  and hence  $\mathbb{V}[\mathcal{C}(\hat{Q})]$ . The probabilistic bounds (5.50) then follow from bounds on  $\mathbb{V}[\mathcal{C}(\hat{Q})]$  using Chebyshev's inequality.

We could apply these proof techniques to prove a probabilistic bound of the form (5.50) for Monte-Carlo and Multi-Level Monte-Carlo methods for the Helmholtz equation. However, the calculations for the Helmholtz equation would be conceptually similar to those in [104], albeit more involved, as we would need to keep track of the  $k$ -dependence. Given we expect the results we obtain would be similar to those in [104], we elect not to pursue them.

### 5.6.3 Proof of Theorem 5.22

We first prove that the assumptions in the abstract setting of Sections 5.3–5.5 hold for the stochastic Helmholtz equation, before applying the theory developed in Sections 5.3–5.5 to prove Theorem 5.22. Recall that we have assumed the stochastic Helmholtz problem is almost-surely nontrapping. In particular, when we apply Theorem 2.39, the constant  $C_{\text{stab}}$  will be independent of  $k$ .

*Proof of Theorem 5.22.* The proof follows immediately from the case  $\beta > \gamma$  in Theorem 5.13, because we have  $\beta = 4p$  or  $4p + 1$  (depending on the QoI), for  $p \geq 1$  and  $\gamma = d$ . We then substitute the appropriate values of  $a$ ,  $\alpha$ ,  $\beta$ , etc. into (5.30), and identify which of the terms  $k^{\tau-a(\beta-\gamma)}$  or  $k^{\gamma\sigma/\alpha}$  dominates for large  $k$ , and which of the terms  $\varepsilon^{-2}$  or  $\varepsilon^{-\gamma/\alpha}$  dominates for small  $\varepsilon$ .

Two cases require explaining a little further. Firstly, the case  $a = (2p + 1)/2p$  and  $Q(u) = \|u\|_{L^2(D)}$  (so  $\alpha = \sigma = 2p$ .) In this case, the expression for  $L$  in (5.28) evaluates as

$$L = \max \left\{ \left\lceil \frac{1}{2p} \log_s \left( \sqrt{2} c_1 C_{\text{coarse}}^{2p} k^{-1} \varepsilon^{-1} \right) \right\rceil, 0 \right\} \quad (5.51)$$

Observe that for  $k$  (or  $\varepsilon$ ) sufficiently large, the first term in the maximum in (5.51) may be negative (i.e. if  $(\varepsilon k)^{-1}$  is sufficiently close to 0, then the logarithm will be negative). In such a case, the maximum of the two quantities on the right-hand side of (5.51) will be 0, and in such a case the Multi-Level Monte-Carlo method reverts to the Monte-Carlo algorithm since  $L = 0$ , i.e., there are no additional levels of refinement. The criterion for the first term to be positive (and so for the Multi-Level Monte-Carlo method to be distinct from the Monte-Carlo method) is

$$\sqrt{2} c_1 C_{\text{coarse}}^{2p} k^{-1} \varepsilon^{-1} > 1,$$

which is equivalent to the condition (5.49).

In the case that the condition (5.49) holds we can apply (5.30), and by substituting in the appropriate values of  $a$ ,  $\alpha$ , etc., the right-hand side of (5.30) becomes

$$k^{\left(\frac{2p+1}{2p}d\right)-2} \varepsilon^{-2} + k^d \varepsilon^{-\frac{d}{2p}}. \quad (5.52)$$

To see which of the  $k$ -dependent terms in (5.52) dominates for large  $k$ , observe that

$$d \geq \frac{2p+1}{2p}d - 2$$

if, and only if,  $p \geq d/4$ . As  $p \geq 1$  and  $d \leq 3$ , we always have  $p \geq d/4$ , and hence the  $k^d$  term dominates.

Secondly, when  $a = 1$  and  $Q(u) = \|u\|_{H_k^1(D)}$ , on substituting the appropriate values of  $a$ ,  $\alpha$ ,

etc. into (5.30), the right-hand side of (5.30) becomes

$$k^{d+2}\varepsilon^{-2} + k^{\frac{2p+1}{2p}d}\varepsilon^{-\frac{d}{2p}}. \quad (5.53)$$

Observe that, analogously to above,  $d+2 \geq (2p+1)d/(2p)$  if, and only if,  $p \geq d/4$ , and therefore the  $k^{d+2}$  term in (5.53) dominates.  $\square$

## 5.7 SUMMARY AND FUTURE WORK

### 5.7.1 Summary

In this chapter we analysed the computational cost of Monte-Carlo (MC) and Multi-Level Monte-Carlo (MLMC) methods for the Helmholtz equation. In particular:

- In Sections 5.3–5.5 we adapted the standard Monte-Carlo and Multi-Level Monte-Carlo complexity theory to the  $k$ -dependent case.
- In Section 5.6 we applied the adapted theory to two Quantities of interest (QoIs), under two different assumptions on the behaviour of the underlying finite-element method, and saw that MLMC is consistently cheaper than MC, with respect to the required tolerance  $\varepsilon$ .

### 5.7.2 Future work

There are several immediate possibilities for building on the work in this chapter:

- Applying the adapted theory to other, more physically realistic QoIs, e.g., the far-field pattern of  $u$  (see, e.g., [50, Section 2.5]).
- Performing numerical experiments to investigate if the predicted speedup of MLMC methods over MC methods is obtained in practice.
- Investigating extensions of MLMC methods, as has already been done for the stationary diffusion equation, e.g., Multi-Level Quasi-Monte-Carlo methods, see, e.g., [130] and Multi-Level Markov-Chain Monte-Carlo methods [56].



# Failure of Fredholm theory for a stochastic variational formulation of Helmholtz problems

The standard approach to proving existence and uniqueness of a (deterministic) Helmholtz BVP is to show that the associated sesquilinear form satisfies a Gårding inequality, and then apply Fredholm theory to deduce that existence and uniqueness are equivalent; see, e.g., [146, Theorem 4.10]. This procedure relies on the fact that the inclusion  $H_{0,D}^1(D_R) \hookrightarrow L^2(D_R)$  is compact; see, e.g., [146, Theorem 3.27].

As noted in Section 3.1.4, the analysis in [80] of Problem 3.3 for the Helmholtz Interior Impedance Problem mimics this approach and assumes that  $L^2(\Omega; H^1(D))$  is compactly contained in  $L^2(\Omega; L^2(D))$ , where  $D$  is the spatial domain. Here we briefly show  $L^2(\Omega; H^1(D))$  is *not* compactly contained in  $L^2(\Omega; L^2(D))$  by giving an explicit example of a bounded sequence in  $L^2(\Omega; H^1(D))$  that has no convergent subsequence in  $L^2(\Omega; L^2(D))$ . Necessary and sufficient conditions for a subset of  $L^p([0, T]; B)$ , for  $B$  a Banach space, to be compact, can be found in [197]. In particular, [197] shows that a space  $C$  being compactly contained in a space  $B$  does not by itself imply  $L^2([0, T]; C)$  is compactly contained in  $L^2([0, T]; B)$ .

**Example A.1.** Let  $(\Omega, \mathcal{F}, \mathbb{P}) = ([0, 1], \mathcal{B}([0, 1]), \lambda)$ . Let  $D$  be a compact subset of  $\mathbb{R}^d$ . Since  $L^2(\Omega)$  is separable, it has an orthonormal basis, which we denote by  $(f_m)_{m \in \mathbb{N}}$ . Let  $u_m \in L^2(\Omega; H^1(D))$  be defined by  $u_m(\omega)(x) := f_m(\omega)$ , for all  $x \in D$ , i.e., for each value of  $\omega$ ,  $u_m(\omega)$  is a constant function on  $D$  and so  $\|u_m(\omega)\|_{H^1(D)} = \|u_m(\omega)\|_{L^2(D)}$ . Then

$$\|u_m\|_{L^2(\Omega; H^1(D))}^2 = \int_{\Omega} \|u_m(\omega)\|_{H^1(D)}^2 d\mathbb{P}(\omega) = \lambda(D)^2 \int_{\Omega} |f_m(\omega)|^2 d\mathbb{P}(\omega) = \|f_m\|_{L^2(\Omega)}^2 \lambda(D)^2,$$

and so  $u_m$  is a bounded sequence in  $L^2(\Omega; H^1(D))$ . However, for  $n \neq m$ , we have

$$\begin{aligned} \|u_m - u_n\|_{L^2(\Omega; L^2(D))}^2 &= \int_{\Omega} \|u_m(\omega) - u_n(\omega)\|_{L^2(D)}^2 d\mathbb{P}(\omega) \\ &= \lambda(D)^2 \int_{\Omega} |f_m(\omega) - f_n(\omega)|^2 d\mathbb{P}(\omega) = \lambda(D)^2 \|f_m - f_n\|_{L^2(\Omega)}^2 = 2\lambda(D)^2 \end{aligned}$$

if  $n \neq m$ , since the  $f_m$  form an orthonormal basis for  $L^2(\Omega)$ . Therefore  $(u_m)_{m \in \mathbb{N}}$  is bounded in  $L^2(\Omega; H^1(D))$  but does not have a convergent subsequence in  $L^2(\Omega; L^2(D))$ , and thus the inclusion of  $L^2(\Omega; H^1(D))$  into  $L^2(\Omega; L^2(D))$  cannot be compact.



# Recap of basic material on measure theory and Bochner spaces

Recall that here, and in the rest of this thesis,  $(\Omega, \mathcal{F}, \mathbb{P})$  is a complete probability space.

## B.1 RECAP OF MEASURE THEORY RESULTS

We first recall some results from measure theory, with our main reference [26]. Even though [26] mainly considers maps with image  $\mathbb{R}$ , the results we quote for more general images are straightforward generalisations of the results in [26].

**Definition B.1** (Measurable map). *If  $(M, \mathcal{M})$  and  $(N, \mathcal{N})$  are measurable spaces, we say that  $f : M \rightarrow N$  is measurable (with respect to  $(\mathcal{M}, \mathcal{N})$ ) if  $f^{-1}(E) \in \mathcal{M}$  for all  $E \in \mathcal{N}$ .*

**Definition B.2** (Borel  $\sigma$ -algebra). *If  $(S, \mathcal{T}_S)$  is a topological space, the Borel  $\sigma$ -algebra  $\mathcal{B}(S)$  on  $S$  is the  $\sigma$ -algebra generated by  $\mathcal{T}_S$ .*

If  $V$  is any topological space (including a Hilbert, Banach, metric, or normed vector space) then we will always take the Borel  $\sigma$ -algebra on  $V$  unless stated otherwise.

**Lemma B.3** (Continuous maps are measurable [26, Theorem 2.1.2]). *Any continuous function between two topological spaces is measurable.*

**Lemma B.4** (The composition of a measurable map with a continuous map is measurable [26, Text at the top of p. 146]). *Let  $(M, \mathcal{M})$  be a measurable space and let  $(S, \mathcal{T}_S)$  and  $(T, \mathcal{T}_T)$  be topological spaces. Let  $f : M \rightarrow S$  be measurable and let  $h : S \rightarrow T$  be continuous. Then  $h \circ f$  is measurable.*

**Definition B.5** (Product  $\sigma$ -algebra [57, Section IV.11]). *Let  $(M_1, \mathcal{M}_1), \dots, (M_m, \mathcal{M}_m)$  be measurable spaces. The product  $\sigma$ -algebra  $\mathcal{M}_1 \otimes \dots \otimes \mathcal{M}_m$  is defined as the  $\sigma$ -algebra generated by the set of measurable rectangles  $\{R_1 \times \dots \times R_m : R_1 \in \mathcal{M}_1, \dots, R_m \in \mathcal{M}_m\}$ .*

**Lemma B.6** (Measurability of the Cartesian product of measurable functions).

*Let  $(M_1, \mathcal{M}_1), \dots, (M_m, \mathcal{M}_m)$  be measurable spaces and  $h_j : \Omega \rightarrow M_j$ ,  $j = 1, \dots, m$  be measurable functions. Then the product map  $P : \Omega \rightarrow M_1 \times \dots \times M_m$  given by  $P(\omega) := (h_1(\omega), \dots, h_m(\omega))$  is measurable with respect to  $(\mathcal{F}, \mathcal{M}_1 \otimes \dots \otimes \mathcal{M}_m)$ .*

*Sketch proof of Lemma B.6.* Let  $\text{Rect}(\mathcal{M}_1, \dots, \mathcal{M}_m)$  denote the set of measurable rectangles, as in Definition B.5. Let  $\mathcal{P} := \{C \subseteq M_1 \times \dots \times M_m : P^{-1}(C) \in \mathcal{F}\}$ . The proof of the lemma consists of the following straightforward steps, whose proofs are omitted: (i) Show  $\text{Rect}(\mathcal{M}_1, \dots, \mathcal{M}_m) \subseteq \mathcal{P}$ . (ii) Show  $\mathcal{P}$  is a  $\sigma$ -algebra. (iii) Deduce  $\mathcal{M}_1 \otimes \dots \otimes \mathcal{M}_m \subseteq \mathcal{P}$  (since  $\mathcal{M}_1 \otimes \dots \otimes \mathcal{M}_m$  is generated

by measurable rectangles). (iv) Conclude  $P$  is measurable with respect to  $(\mathcal{F}, \mathcal{M}_1 \otimes \cdots \otimes \mathcal{M}_m)$ .  $\square$

**Lemma B.7** (The product of Borel  $\sigma$ -algebras is the Borel  $\sigma$ -algebra of the product [26, Lemma 6.2.1 (i)]). *Let  $H_1, H_2$  be Hausdorff spaces and let  $H_2$  have a countable base (e.g.  $H_2$  could be a separable metric space). Then  $\mathcal{B}(H_1 \times H_2) = \mathcal{B}(H_1) \otimes \mathcal{B}(H_2)$ , where  $\mathcal{B}(H_1 \times H_2)$  is the Borel  $\sigma$ -algebra of the product topology on  $H_1 \times H_2$ .*

## B.2 RECAP OF RESULTS ON BOCHNER SPACES

We now recap the theory of Bochner spaces, using [54] as our main reference. In what follows the space  $V$  is always a Banach space.

**Definition B.8** (Simple function). *A function  $v : \Omega \rightarrow V$  is simple if there exist  $v_1, \dots, v_m \in V$  and  $E_1, \dots, E_m \in \mathcal{F}$  such that  $v = \sum_{i=1}^m v_i \chi_{E_i}$ , where  $\chi_{E_i}$  is the indicator function on  $E_i$ .*

**Definition B.9** (Strongly measurable). *A function  $v : \Omega \rightarrow V$  is strongly measurable<sup>1</sup> if there exists a sequence of simple functions  $(v_n)_{n \in \mathbb{N}}$  such that  $\lim_{n \rightarrow \infty} \|v_n - v\|_V = 0$ ,  $\mathbb{P}$ -almost everywhere.*

**Definition B.10** (Bochner integrable [54, p. 49]). *A strongly measurable function  $v : \Omega \rightarrow V$  is called Bochner integrable if there exists a sequence of simple functions  $(v_n)_{n \in \mathbb{N}}$  such that*

$$\lim_{n \rightarrow \infty} \int_{\Omega} \|v_n(\omega) - v(\omega)\|_V d\mathbb{P}(\omega) = 0.$$

**Theorem B.11** (Condition for Bochner integrability [54, Theorem II.2.2]).

*A strongly measurable function  $v : \Omega \rightarrow V$  is Bochner integrable if and only if  $\int_{\Omega} \|v\|_V d\mathbb{P} < \infty$ .*

**Corollary B.12** (Sufficient condition for Bochner integrability). *Let  $p \geq 1$ . If a strongly measurable function  $v : \Omega \rightarrow V$  has  $\int_{\Omega} \|v\|_V^p d\mathbb{P} < \infty$ , then  $v$  is Bochner integrable.*

**Definition B.13** (Bochner norm). *For a Bochner integrable function  $v : \Omega \rightarrow V$ , let*

$$\|v\|_{L^p(\Omega; V)} := \left( \int_{\Omega} \|v(\omega)\|_V^p d\mathbb{P}(\omega) \right)^{1/p}, \quad 1 \leq p < \infty, \quad \text{and} \quad \|v\|_{L^\infty(\Omega; V)} := \operatorname{ess\,sup}_{\omega \in \Omega} \|v(\omega)\|_V.$$

**Definition B.14** (Bochner space). *Let  $1 \leq p \leq \infty$ . Then*

$$L^p(\Omega; V) := \left\{ v : \Omega \rightarrow V : v \text{ is Bochner integrable, } \|v\|_{L^p(\Omega; V)} < \infty \right\}.$$

**Definition B.15** (Complete probability space). *A probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  is complete if for every  $E_1 \in \mathcal{F}$  with  $\mathbb{P}(E_1) = 0$ , the inclusion  $E_2 \subseteq E_1$  implies that  $E_2 \in \mathcal{F}$ .*

**Definition B.16** (Separable space). *A topological space is separable if it contains a countable, dense subset.*

<sup>1</sup>In [54] the authors use the term  $\mu$ -measurable instead of strongly measurable (where  $\mu$  is the measure on the domain of the functions under consideration).

**Definition B.17** ( $\sigma$ -finite). A probability space  $(\Omega, \mathcal{F}, \mathbb{P})$  is  $\sigma$ -finite if there exist  $E_1, E_2, \dots \in \mathcal{F}$  such that  $\Omega = \bigcup_{m=1}^{\infty} E_m$ .

**Theorem B.18** (Pettis measurability theorem [186, Proposition 2.15]). Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be complete and  $\sigma$ -finite. The following are equivalent for a function  $v : \Omega \rightarrow V$ : (i)  $v$  is strongly measurable, (ii)  $v$  is measurable and  $\mathbb{P}$ -essentially separably valued.

**Corollary B.19** (Equivalence of measurable and strongly measurable when the image is separable). Let  $(\Omega, \mathcal{F}, \mathbb{P})$  be  $\sigma$ -finite. If  $V$  is a separable Banach space, then a function  $v : \Omega \rightarrow V$  is strongly measurable if, and only if, it is measurable.

**Lemma B.20** (The composition of a continuous map and a  $\mathbb{P}$ -essentially separably valued map). Let  $(S, \mathcal{T}_S)$  and  $(T, \mathcal{T}_T)$  be topological spaces. If  $f_1 : \Omega \rightarrow S$  and  $f_2 : S \rightarrow T$  are such that  $f_1$  is  $\mathbb{P}$ -essentially separably valued and  $f_2$  is continuous, then  $f_2 \circ f_1$  is  $\mathbb{P}$ -essentially separably valued.

*Proof of Lemma B.20.* As  $f_1$  is  $\mathbb{P}$ -essentially separably valued, there exists  $E \in \mathcal{F}$  such that  $\mathbb{P}(E) = 1$  and  $f_1(E) \subseteq G \subseteq S$ , where  $G$  is separable. As  $f_2$  is continuous,  $f_2(G)$  is separable [215, Theorem 16.4(a)]. Therefore, since  $(f_2 \circ f_1)(E) \subseteq f_2(G)$ , it follows that  $f_2 \circ f_1$  is  $\mathbb{P}$ -essentially separably valued.  $\square$

**Lemma B.21** (The composition of a continuous map and a strongly measurable map). If  $B_1$  and  $B_2$  are Banach spaces and there exist  $f_1 : \Omega \rightarrow B_1$  and  $f_2 : B_1 \rightarrow B_2$  such that  $f_1$  is strongly measurable and  $f_2$  is continuous, then  $f_2 \circ f_1$  is strongly measurable.

*Proof of Lemma B.21.* By Theorem B.18,  $f_1$  is both measurable and  $\mathbb{P}$ -essentially separably valued. Therefore we can apply Lemmas B.4 and B.20 to conclude  $f_2 \circ f_1$  is both measurable and  $\mathbb{P}$ -essentially separably valued. Hence by Theorem B.18  $f_2 \circ f_1$  is strongly measurable.  $\square$

**Lemma B.22** (Zero in all integrals implies zero almost everywhere [54, Corollary II.2.5]). If  $\alpha$  is Bochner integrable and  $\int_E \alpha(\omega) d\mathbb{P}(\omega) = 0$  for each  $E \in \mathcal{F}$  then  $\alpha = 0$   $\mathbb{P}$ -almost everywhere.

**Lemma B.23** (Cartesian product of  $\mathbb{P}$ -essentially separably valued maps). Let  $(\mathcal{C}_1, \mathcal{T}_{\mathcal{C}_1}), \dots, (\mathcal{C}_m, \mathcal{T}_{\mathcal{C}_m})$  be topological spaces, and let  $s_j : \Omega \rightarrow \mathcal{C}_j$ ,  $j = 1, \dots, m$  be  $\mathbb{P}$ -essentially separably valued. Define  $\mathcal{C} := \mathcal{C}_1 \times \dots \times \mathcal{C}_m$  and equip  $\mathcal{C}$  with the product topology. Then the map  $f : \Omega \rightarrow \mathcal{C}$  given by  $s(\omega) := (s_1(\omega), \dots, s_m(\omega))$  is  $\mathbb{P}$ -essentially separably valued.

The proof of Lemma B.23 is straightforward and omitted.



## Measurability of series expansions (used in Section 3.1.2)

Here we collect together results from measure theory that allow us to conclude in Lemma C.12 that the series expansions for  $A$  and  $n$  in Section 3.1.2 are measurable. As mentioned in Section 3.1.2, the proof that the sum of measurable functions is measurable is standard, but we have not been able to find this result stated in the literature for this particular setting of mappings into a separable subspace of a general normed vector space.

**Lemma C.1.** *If  $U$  is a separable normed vector space,  $m \in \mathbb{N}$ , and  $\phi_j : \Omega \rightarrow U$ ,  $j = 1, \dots, m$  are measurable functions, then  $\phi_1 + \dots + \phi_m : \Omega \rightarrow U$  is measurable.*

*Sketch proof of Lemma C.1.* By induction, it is sufficient to show the result for  $m = 2$ . We let  $B_r^U(v)$  denote the ball of radius  $r > 0$  about  $v \in U$ . To show  $\phi_1 + \phi_2$  is measurable, we let  $v \in U$ ,  $r > 0$  and we show  $(\phi_1 + \phi_2)^{-1}(B_r^U(v)) \in \mathcal{F}$ . Let  $\mathbb{Q}_U$  denote a countable dense subset of  $U$ , which exists as  $U$  is separable. Let  $\mathbb{Q}_{\mathbb{F}}$  denote a countable dense subset of the field  $\mathbb{F}$ , which exists as  $\mathbb{F} = \mathbb{R}$  or  $\mathbb{C}$ .

For  $s \in \mathbb{Q}_{\mathbb{F}}$ ,  $q \in \mathbb{Q}_U$  let

$$S_{s,q} = \left\{ \omega \in \Omega : \left\| \phi_1(\omega) - \frac{1}{2}v - q \right\|_U < s \right\} \cap \left\{ \omega \in \Omega : \left\| \phi_2(\omega) - \frac{1}{2}v + q \right\|_U < r - s \right\}.$$

We claim

$$(\phi_1 + \phi_2)^{-1}(B_r^U(v)) = \bigcup_{s \in \mathbb{Q}_{\mathbb{F}}} \bigcup_{q \in \mathbb{Q}_U} S_{s,q}, \quad (\text{C.1})$$

and the result then follows as the right-hand side is an element of the  $\sigma$ -algebra  $\mathcal{F}$ . To show (C.1), let  $\omega \in \bigcup_{s \in \mathbb{Q}_{\mathbb{F}}} \bigcup_{q \in \mathbb{Q}_U} S_{s,q}$ , and let  $s \in \mathbb{Q}_{\mathbb{F}}$ ,  $q \in \mathbb{Q}_U$  be such that  $\omega \in S_{s,q}$ . Then it follows from the triangle inequality that  $\omega \in (\phi_1 + \phi_2)^{-1}(B_r^U(v))$ . Now let  $\omega \in (\phi_1 + \phi_2)^{-1}(B_r^U(v))$ , define  $r_\omega := r - \|\phi_1(\omega) + \phi_2(\omega) - v\|_U > 0$ , fix  $s \in \mathbb{Q}_{\mathbb{F}} \cap (0, r_\omega/2)$ , and choose  $q \in \mathbb{Q}_U$  such that  $\|\phi_1(\omega) - v/2 - q\|_U < s$ . Then again it follows from the triangle inequality that  $\omega \in S_{s,q}$ , and thus (C.1) holds, as required.  $\square$

**Corollary C.2.** *If  $V$  is a normed vector space,  $U \subseteq V$  is a separable subspace, and  $\phi_j : \Omega \rightarrow U$ ,  $j = 1, \dots, m$  are measurable functions, then  $\phi_1 + \dots + \phi_m : \Omega \rightarrow U$  is measurable.*

**Lemma C.3.** *Let  $V$  be a normed vector space. If  $v \in V$  and  $Y : \Omega \rightarrow \mathbb{F}$  is a measurable function, then  $Yv : \Omega \rightarrow V$  is a measurable function.*

*Proof of Lemma C.3.* The map  $M_v : \mathbb{F} \rightarrow V$  given by  $M_v(x) = xv$  is continuous. As  $Yv = M_v \circ Y$ , it follows from Lemma B.4 that  $Yv$  is measurable.  $\square$

**Lemma C.4.** *If  $V$  is a normed vector space and  $U \subseteq V$ , then the inclusion map  $\iota : U \rightarrow V$  is measurable.*

*Proof of Lemma C.4.* As  $\iota$  is continuous, it immediately follows that it is measurable.  $\square$

**Corollary C.5.** *If  $V$  is a normed vector space,  $U \subseteq V$  and  $\phi : \Omega \rightarrow U$  is measurable, then  $\phi : \Omega \rightarrow V$  is measurable.*

*Proof of Corollary C.5.* This is immediate from Lemma C.4 and Lemma B.4.  $\square$

**Lemma C.6.** *If  $V$  is a normed vector space,  $m \in \mathbb{N}$ , and  $\phi_1, \dots, \phi_m \in V$  for  $j = 1, \dots, m$  then  $\text{span}\{\phi_1, \dots, \phi_m\}$  is a separable subspace of  $V$ .*

*Sketch Proof of Lemma C.6.* As  $\mathbb{F} = \mathbb{R}$  or  $\mathbb{C}$ , it has a separable subset  $\mathbb{Q}_{\mathbb{F}}$ . Since a finite product of countable sets is countable, the set

$$\left\{ B_{1/n}^V(q_1\phi_1 + \dots + q_m\phi_m) : n \in \mathbb{N}, q_1, \dots, q_m \in \mathbb{Q}_{\mathbb{F}} \right\}$$

is a countable base for the topology on  $\text{span}\{\phi_1, \dots, \phi_m\}$  induced by the norm  $\|\cdot\|_V$ .  $\square$

**Lemma C.7.** *If  $V$  is a normed vector space,  $m \in \mathbb{N}$ , and for  $j = 1, \dots, m$ ,  $\phi_j \in V$  and  $Y_j : \Omega \rightarrow \mathbb{F}$  are measurable, then the function  $\phi : \Omega \rightarrow V$  given by*

$$\phi(\omega) = \phi_0 + \sum_{j=1}^m Y_j(\omega)\phi_j$$

*is measurable.*

*Proof of Lemma C.7.* The subspace  $U = \text{span}\{\phi_0, \phi_1, \dots, \phi_m\}$  is separable by Lemma C.6, and it is clear that the image of  $\phi$  lies in  $U$ . By Lemma C.3 and Corollary C.2,  $\phi : \Omega \rightarrow U$  is measurable, and therefore  $\phi : \Omega \rightarrow V$  is measurable by Corollary C.5.  $\square$

We now prove that almost-surely convergent sequences of measurable functions are measurable, and we then apply this result to the partial sums in the definitions of  $A$  and  $n$  in (3.13).

We will use the following theorem to establish that the almost-sure limit of a sequence of measurable functions is measurable.

**Theorem C.8** ([60, Theorem 4.2.2]). *Let  $(W, d)$  be a metric space. Suppose the functions  $\zeta_j : \Omega \rightarrow W$  are measurable, for all  $j \in \mathbb{N}$ . If the limit*

$$\zeta(\omega) = \lim_{j \rightarrow \infty} \zeta_j(\omega)$$

*exists for every  $\omega \in \Omega$ , then the function  $\zeta : \Omega \rightarrow W$  is measurable.*

**Corollary C.9.** *Let  $(W, d)$  be a metric space. Suppose the functions  $\zeta_m : \Omega \rightarrow W$  are measurable, for all  $m \in \mathbb{N}$ . If the limit*

$$\lim_{m \rightarrow \infty} \zeta_m(\omega) \tag{C.2}$$

exists almost surely, then there exists a measurable function  $\zeta : \Omega \rightarrow W$  such that

$$\zeta(\omega) = \lim_{m \rightarrow \infty} \zeta_m(\omega)$$

whenever the limit exists.

*Proof of Corollary C.9.* Following [65], we define  $\tilde{\Omega} = \{\omega \in \Omega : \text{(C.2) exists}\}$ . Then, for  $m \in \mathbb{N}$  define  $\tilde{\zeta}_m : \Omega \rightarrow W$  by

$$\tilde{\zeta}_m(\omega) = \begin{cases} \zeta_m(\omega) & \text{if } \omega \in \tilde{\Omega} \\ 0 & \text{if } \omega \notin \tilde{\Omega} \end{cases}$$

Observe that, by construction, the limit  $\tilde{\zeta}(\omega) = \lim_{m \rightarrow \infty} \tilde{\zeta}_m(\omega)$  exists for all  $\omega \in \Omega$  and the functions  $\tilde{\zeta}_m$  are measurable. Therefore, by Theorem C.8,  $\tilde{\zeta}$  is measurable.  $\square$

**Lemma C.10.** *Let  $V$  be a normed vector space. If there exist  $\phi_j \in V$ ,  $j = 0, 1, \dots$  and measurable functions  $Y_j : \Omega \rightarrow \mathbb{F}$ ,  $j \in \mathbb{N}$  such that the series*

$$\phi_0 + \sum_{j=1}^{\infty} Y_j(\omega) \phi_j$$

*exists in  $V$  almost surely, then there exists a measurable function  $\phi : \Omega \rightarrow V$  such that*

$$\phi(\omega) = \phi_0 + \sum_{j=1}^{\infty} Y_j(\omega) \phi_j$$

*almost surely.*

*Proof of Lemma C.10.* By Lemma C.7, the partial sums  $\phi_0 + \sum_{j=1}^m Y_j(\omega) \phi_j$ , for  $m \in \mathbb{N}$  are measurable, and by assumption their limit as  $m \rightarrow \infty$  exists almost surely. Therefore, applying Corollary C.9 to the partial sums, we obtain the result.  $\square$

**Lemma C.11.** *The series expansions for both  $A$  and  $n$  defined by (3.13) exist in  $W^{1,\infty}(D_R; \mathbb{R}^{d \times d})$  and  $W^{1,\infty}(D_R; \mathbb{R})$  almost surely, respectively.*

*Proof of Lemma C.11.* The spaces  $W^{1,\infty}(D_R; \mathbb{R}^{d \times d})$  and  $W^{1,\infty}(D_R; \mathbb{R})$  are Banach spaces, by definition of their norms (see (3.4) and (3.5)). Therefore it suffices to show that the partial sums of the series expansions for  $A$  and  $n$  in (3.13) are Cauchy sequences. As the proofs for  $A$  and  $n$  are completely analogous, we only give the proof for  $A$  here.

First observe that since each of the random variables  $Y_j$  in (3.13) is uniformly distributed on  $[-1/2, 1/2]$ , it follows that for all  $j \in \mathbb{N}$ ,  $\text{ess sup}_{\omega \in \Omega} |Y_j(\omega)| = \frac{1}{2}$ . Therefore, we can conclude that the bound  $\text{ess sup}_{\omega \in \Omega} \sup_{j \in \mathbb{N}} |Y_j(\omega)| \leq \frac{1}{2}$  holds. (For if not, then, there would exist  $\hat{\Omega} \subseteq \Omega$  with  $\mathbb{P}(\hat{\Omega}) > 0$  such that for all  $\omega \in \hat{\Omega}$ ,  $\sup_{j \in \mathbb{N}} |Y_j(\omega)| > \frac{1}{2}$ . Then there would exist  $\hat{j} \in \mathbb{N}$  such that  $|Y_{\hat{j}}(\omega)| > 1/2$  for all  $\omega \in \hat{\Omega}$ , which would give the contradiction  $\text{ess sup}_{\omega \in \Omega} |Y_{\hat{j}}(\omega)| > \frac{1}{2}$ .)

It now suffices to show that for  $\mathbb{P}$ -almost every  $\omega \in \Omega$ , the partial sums of the series expansion in (3.13) form a Cauchy sequence. Recall that for  $\mathbb{P}$ -almost every  $\omega \in \Omega$

$$\sup_{j \in \mathbb{N}} |Y_j(\omega)| \leq \frac{1}{2}.$$

For such an  $\omega$ , and  $m \in \mathbb{N}$ , define the  $m$ th partial sum

$$A_m(\omega) = A_0 + \sum_{j=1}^m Y_j(\omega) \Psi_j.$$

It is straightforward to show that  $(A_m(\omega))_{m \in \mathbb{N}}$  is a Cauchy sequence in  $W^{1,\infty}(D_R; \mathbb{R}^{d \times d})$ , using the assumption (3.14); therefore, the series expansion for  $A(\omega)$  in (3.13) exists almost surely.  $\square$

**Lemma C.12.** *The functions  $A$  and  $n$  defined by (3.13) are measurable.*

*Proof of Lemma C.12.* The result immediately follows from Lemmas C.10 and C.11.  $\square$

# Error estimators for complex-valued random variables

In this appendix, we recall the definition of some elementary properties of *complex-valued* random variables, properties that are slightly different to their analogues for real-valued random variables. We then prove that the standard unbiased estimator of the variance (with Bessel's correction) of a complex-valued random variable is indeed an unbiased estimator of the variance.

**Definition D.1** (Complex-valued random variable [165, Equation (4.2)]). *If  $(\Omega, \mathcal{F}, \mathbb{P})$  is a probability space, then a complex-valued random variable is a map  $Y : \Omega \rightarrow \mathbb{C}$ , such that  $\Re Y$  and  $\Im Y$  are real-valued random variables.*

**Definition D.2** (Mean and variance of a complex-valued random variable [165, Equations (5.8) and (5.27)]). *Let  $Y$  be a complex-valued random variable. The expectation of  $Y$  is*

$$\mathbb{E}[Y] := \mathbb{E}[\Re Y] + i\mathbb{E}[\Im Y],$$

*if it exists. The variance of  $Y$  is*

$$\mathbb{V}[Y] := \mathbb{E}[|Y|^2] - |\mathbb{E}[Y]|^2$$

*if it exists.*

**Definition D.3** ( $\sigma$ -algebra generated by a random variable). *Let  $Y$  be a complex-valued random variable. The  $\sigma$ -algebra generated by  $Y$  is*

$$\sigma(Y) := \{Y^{-1}(E) : E \in \sigma(\mathbb{C})\},$$

*where  $Y^{-1}$  denotes the pullback.*

**Definition D.4** (Independent  $\sigma$ -algebras). *Two  $\sigma$ -algebras  $\mathcal{F}_1$  and  $\mathcal{F}_2$  on  $\Omega$  are independent if all their sets are independent, i.e.*

$$\mathbb{P}(E_1) \cap \mathbb{P}(E_2) = \mathbb{P}(E_1)\mathbb{P}(E_2)$$

*for all  $E_1 \in \mathcal{F}_1$  and  $E_2 \in \mathcal{F}_2$ .*

**Definition D.5** (Independent random variables). *Two complex-valued random variables  $Y_1$  and  $Y_2$  are independent if their respective generated  $\sigma$ -algebras are independent.*

**Lemma D.6** (Independent implies uncorrelated). *If  $Y_1$  and  $Y_2$  are independent complex-valued random variables, then*

$$\mathbb{E}[Y_1 \overline{Y_2}] = \mathbb{E}[Y_1] \mathbb{E}[\overline{Y_2}].$$

The proof of Lemma D.6 is identical to the real case.

**Definition D.7** (Monte-Carlo estimator for  $\mathbb{E}[Y]$ ). *Let  $Y$  be a complex-valued random variable, and  $Y_1, \dots, Y_N$  be independent and identically distributed to  $Y$ . The Monte-Carlo estimator of  $\mathbb{E}[Y]$  is*

$$\hat{Y} := \frac{1}{N} \sum_{l=1}^N Y_l.$$

**Definition D.8** (Unbiased estimator of the variance of the Monte-Carlo estimator). *Let  $Y$  be a complex-valued random variable, and  $\hat{Y}$  the Monte-Carlo estimator of  $\mathbb{E}[Y]$ . The estimator  $s_N(\hat{Y})$  of  $\mathbb{V}[\hat{Y}]$  is*

$$s_N(\hat{Y}) := \frac{1}{N(N-1)} \sum_{j=1}^N |Y_j - \hat{Y}|^2. \quad (\text{D.1})$$

Observe that (D.1) defines an estimator for the variance of *the Monte-Carlo estimator*  $\hat{Y}$ ; this is in contrast to more standard statistical settings, where one constructs an estimator of the variance of  $Y$ . The factor  $1/(N-1)$  in (D.1) is known as *Bessel's correction*, and ensures the estimator is unbiased, as we now prove.

**Lemma D.9** (The unbiased estimator is unbiased). *Let  $Y$  be a complex-valued random variable and  $\hat{Y}$  the Monte-Carlo estimator of  $\mathbb{E}[Y]$ . Then  $s_N(\hat{Y})$  is unbiased, i.e.,*

$$\mathbb{E}[s_N(\hat{Y})] = \mathbb{V}[\hat{Y}].$$

The proof of Lemma D.9 is nearly identical to the proof for an unbiased estimator for  $\mathbb{V}[Y]$  in the real-valued case. Nevertheless, we write the proof out in full, as we have not been able to find this exact result anywhere in the literature.

*Proof of Lemma D.9.* Firstly, note that

$$\mathbb{V}[\hat{Y}] = \mathbb{V}\left[\frac{1}{N} \sum_{l=1}^N Y_l\right] = \frac{1}{N^2} \mathbb{V}\left[\sum_{l=1}^N Y_l\right] = \frac{1}{N} \mathbb{V}[Y].$$

Therefore, it is sufficient to show that  $s_N(\hat{Y})$  is an unbiased estimator for  $\mathbb{V}[Y]/N$ , or equivalently,  $Ns_N(\hat{Y})$  is an unbiased estimator for  $\mathbb{V}[Y]$  (i.e.  $\mathbb{E}[Ns_N(\hat{Y})] = \mathbb{V}[Y]$ ). We show the latter by direct computation. Observe that

$$\mathbb{E}[Ns_N(\hat{Y})] = \frac{1}{N-1} \mathbb{E}\left[\sum_{j=1}^N |Y_j - \hat{Y}|^2\right] = \frac{1}{N-1} \sum_{j=1}^N \mathbb{E}[|Y_j - \hat{Y}|^2],$$

therefore, it is sufficient for us to show

$$\mathbb{E}[|Y_j - \hat{Y}|^2] = \frac{N-1}{N} \mathbb{V}[Y], \quad (\text{D.2})$$

as the  $Y_j$ s are all independently and identically distributed. We show (D.2) by direct computation.

$$\begin{aligned}
\mathbb{E}\left[|Y_j - \hat{Y}|^2\right] &= \mathbb{E}\left[|Y_j|^2 - Y_j \bar{Y} - \bar{Y} Y_j + |\bar{Y}|^2\right] \\
&= \mathbb{E}\left[|Y_j|^2\right] - \frac{1}{N} \sum_{l=1}^N \left(\mathbb{E}\left[Y_j \bar{Y}_l + \bar{Y}_j Y_l\right]\right) + \mathbb{E}\left[|\bar{Y}|^2\right] \\
&= \mathbb{E}\left[|Y_j|^2\right] - \frac{2}{N} \mathbb{E}\left[|Y_j|^2\right] - \frac{1}{N} \sum_{l \neq j} \left(\mathbb{E}\left[Y_j\right] \mathbb{E}\left[\bar{Y}_l\right] + \mathbb{E}\left[\bar{Y}_j\right] \mathbb{E}\left[Y_l\right]\right) + \mathbb{E}\left[|\bar{Y}|^2\right] \\
&= \frac{N-2}{N} \mathbb{E}\left[|Y|^2\right] - \frac{2(N-1)}{N} |\mathbb{E}[Y]|^2 + \mathbb{E}\left[|\bar{Y}|^2\right], \tag{D.3}
\end{aligned}$$

since the  $Y_l$ s have the same distribution as  $Y$ . We now turn our attention to simplifying  $\mathbb{E}\left[|\bar{Y}|^2\right]$ . We have

$$\begin{aligned}
\mathbb{E}\left[|\bar{Y}|^2\right] &= \mathbb{E}\left[\left(\frac{1}{N} \sum_{l=1}^N Y_l\right) \overline{\left(\frac{1}{N} \sum_{m=1}^N Y_m\right)}\right] \\
&= \frac{1}{N^2} \mathbb{E}\left[\sum_{l=1}^N |Y_l|^2 + \sum_{l=1}^N \sum_{l \neq m} Y_l \bar{Y}_m\right] \\
&= \frac{1}{N} \mathbb{E}\left[|Y|^2\right] + \frac{N(N-1)}{N^2} |\mathbb{E}[Y]|^2, \tag{D.4}
\end{aligned}$$

since the  $Y_l$  are i.i.d. Therefore combining (D.3) and (D.4), we obtain

$$\begin{aligned}
\mathbb{E}\left[|Y_j - \hat{Y}|^2\right] &= \frac{N-2}{N} \mathbb{E}\left[|Y|^2\right] - \frac{2(N-1)}{N} |\mathbb{E}[Y]|^2 + \frac{1}{N} \mathbb{E}\left[|Y|^2\right] + \frac{N-1}{N} |\mathbb{E}[Y]|^2 \\
&= \frac{N-1}{N} \mathbb{E}\left[|Y|^2\right] - \frac{N-1}{N} |\mathbb{E}[Y]|^2 \\
&= \frac{N-1}{N} (\mathbb{V}[Y] + |\mathbb{E}[Y]|^2) - \frac{N-1}{N} |\mathbb{E}[Y]|^2,
\end{aligned}$$

which gives us (D.2), as required. □



# Numerical investigation of QMC convergence for the Helmholtz equation

In this appendix we give plots showing the dependence of the QMC error  $\text{Err}_{\text{QMC}}(N_{\text{QMC}}, N_{\text{shifts}})$  on  $N_{\text{QMC}}$  for *fixed* values of  $k$  and increasing  $N_{\text{QMC}}$ . Aside from the fact that we vary  $N_{\text{QMC}}$ , our computational setup is as in Section 4.6.3.

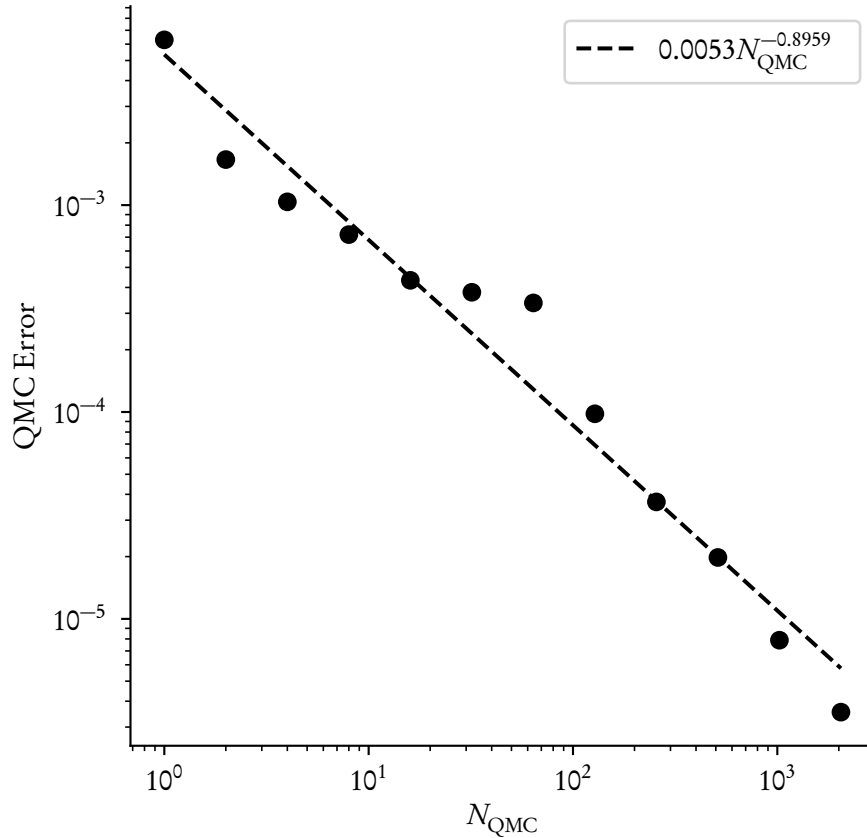


Figure E.1: Quasi-Monte-Carlo error with increasing number of Quasi-Monte-Carlo points, for  $Q = \int_D u$  and  $k = 10$ .

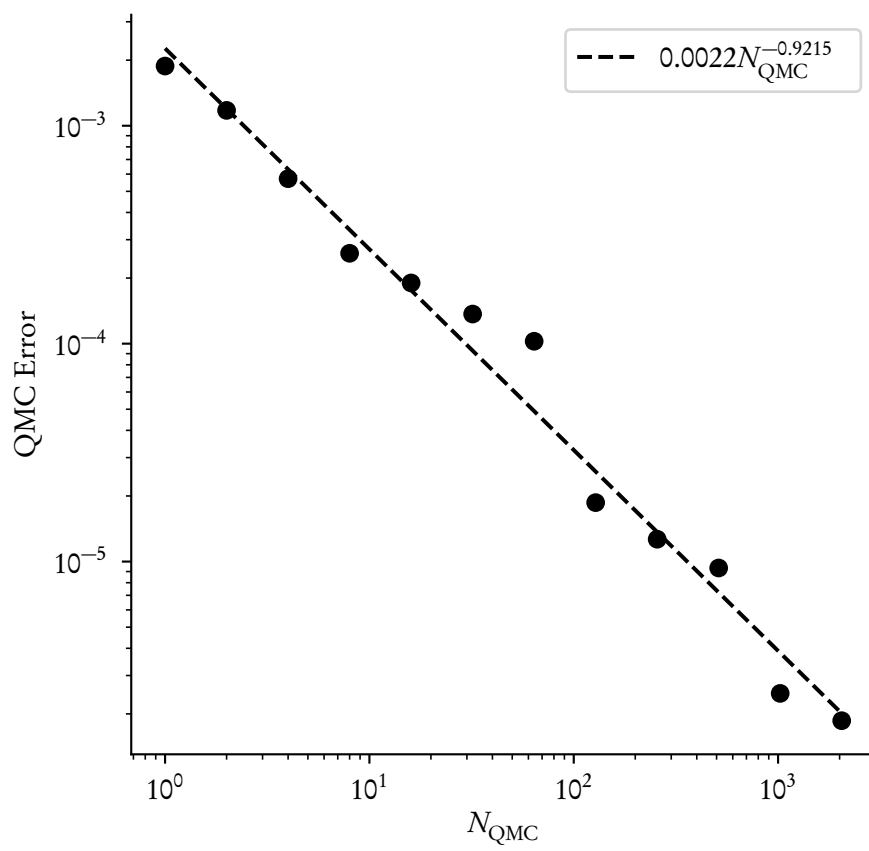


Figure E.2: Quasi-Monte-Carlo error with increasing number of Quasi-Monte-Carlo points, for  $Q = \int_D u$  and  $k = 20$ .

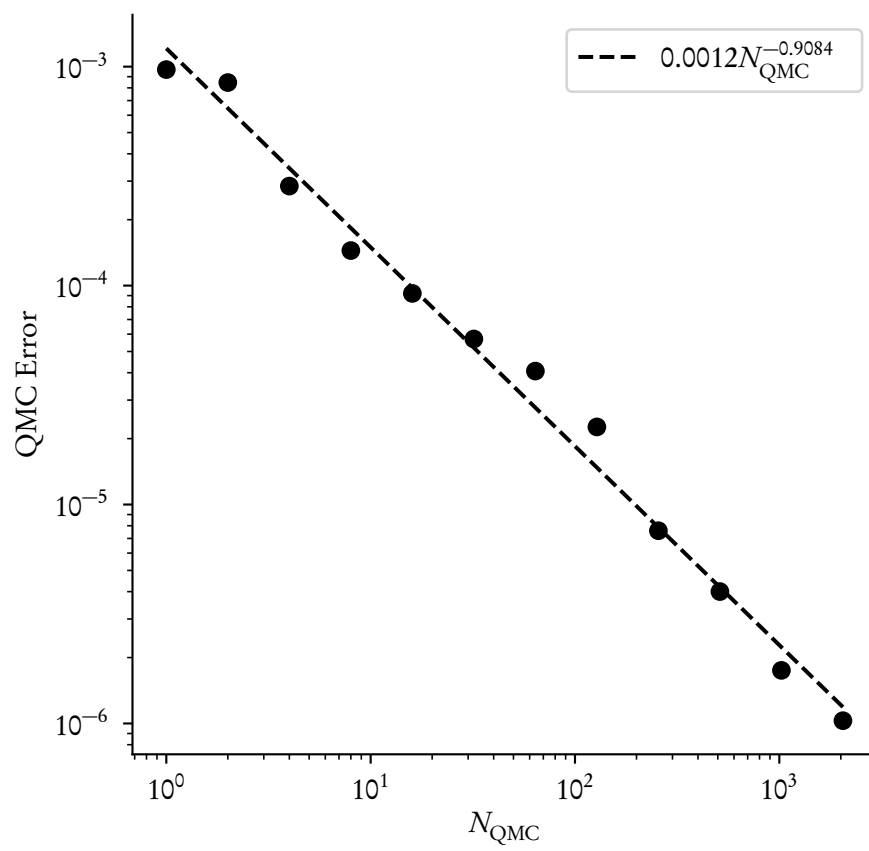


Figure E.3: Quasi-Monte-Carlo error with increasing number of Quasi-Monte-Carlo points, for  $Q = \int_D u$  and  $k = 30$ .

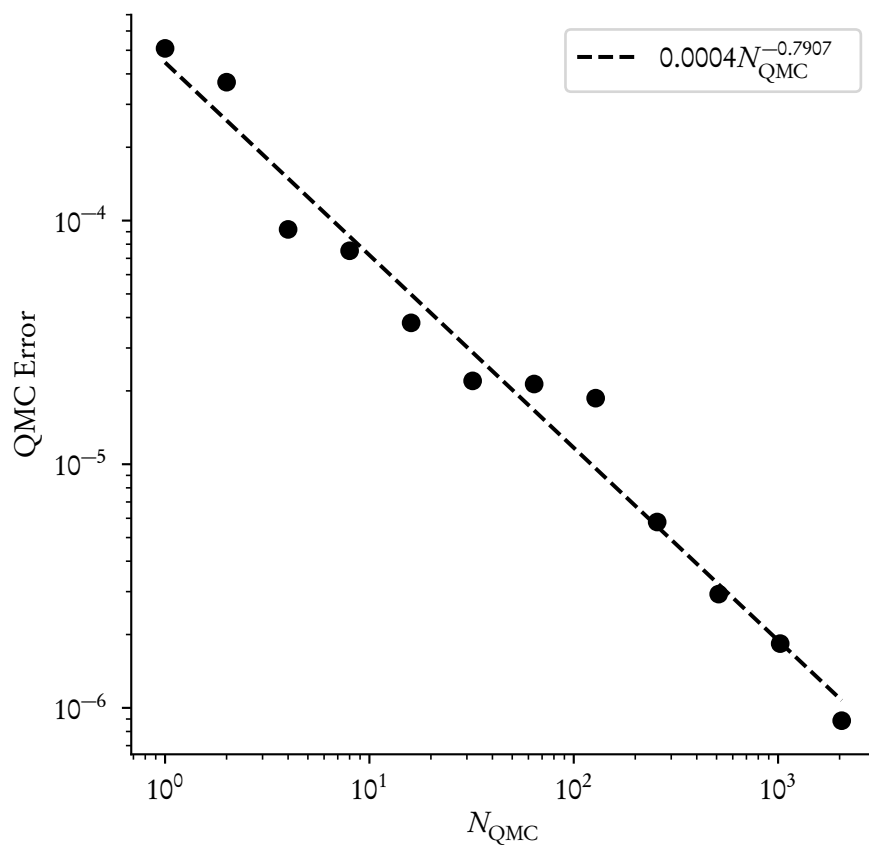


Figure E.4: Quasi-Monte-Carlo error with increasing number of Quasi-Monte-Carlo points, for  $Q = \int_D u$  and  $k = 40$ .

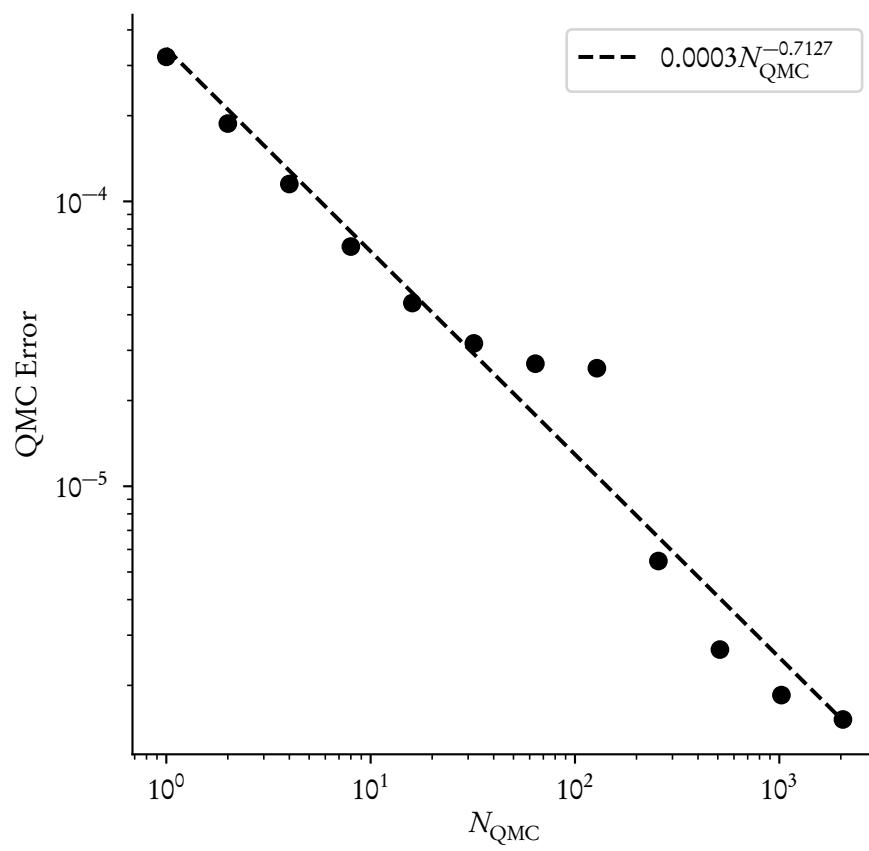


Figure E.5: Quasi-Monte-Carlo error with increasing number of Quasi-Monte-Carlo points, for  $Q = \int_D u$  and  $k = 50$ .

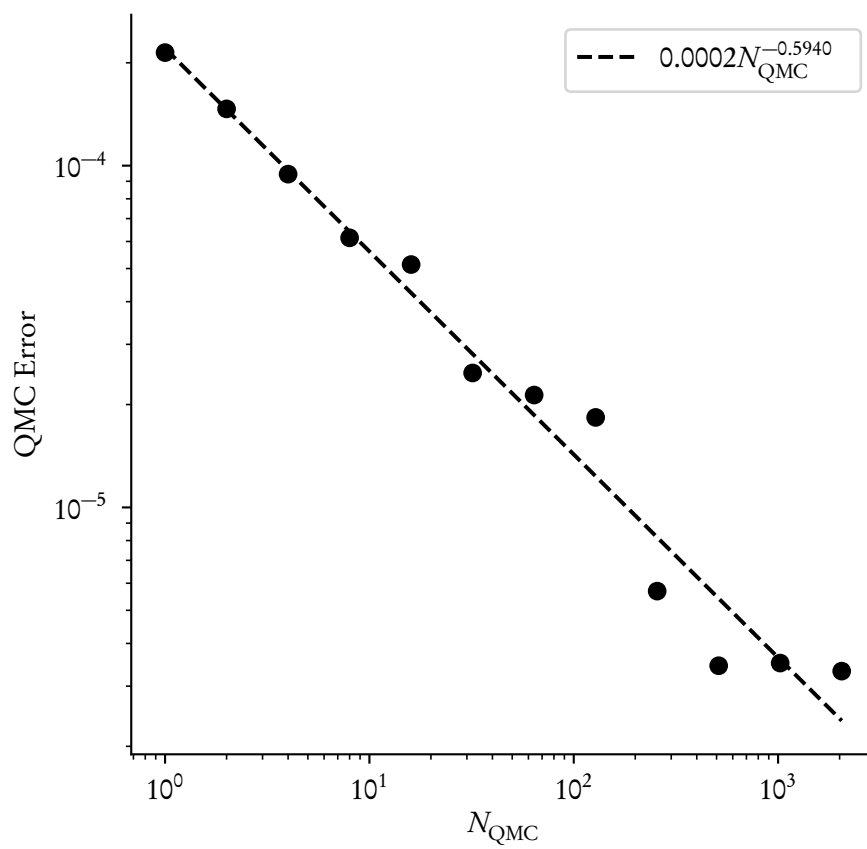


Figure E.6: Quasi-Monte-Carlo error with increasing number of Quasi-Monte-Carlo points, for  $Q = \int_D u$  and  $k = 60$ .

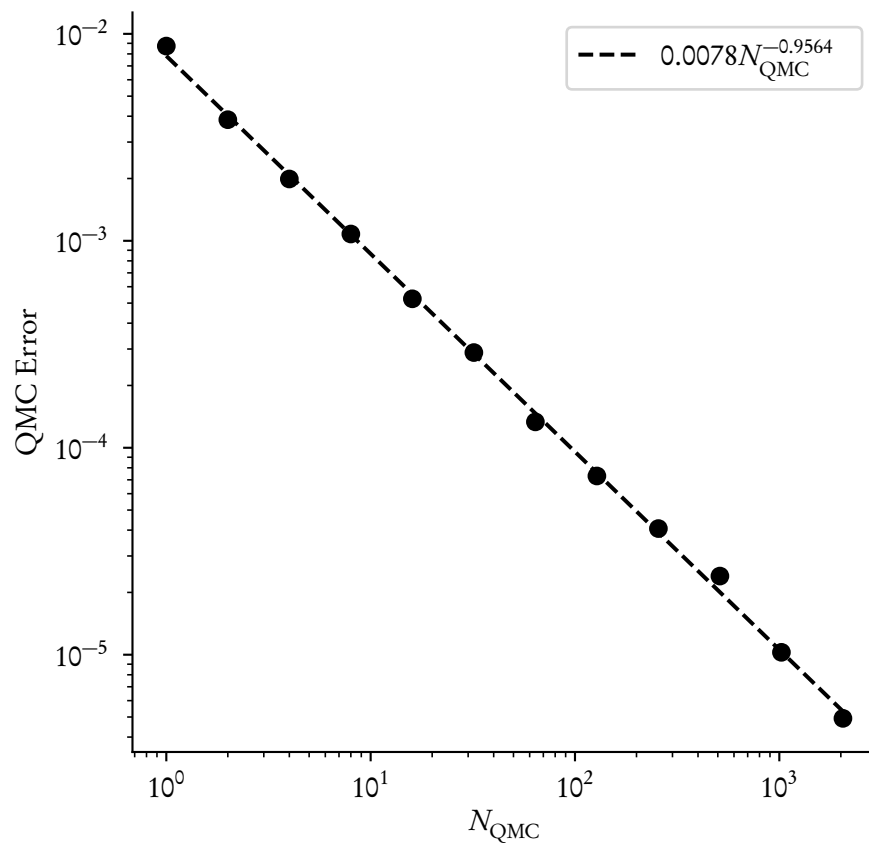


Figure E.7: Quasi-Monte-Carlo error with increasing number of Quasi-Monte-Carlo points, for  $Q = u(0)$  and  $k = 10$ .

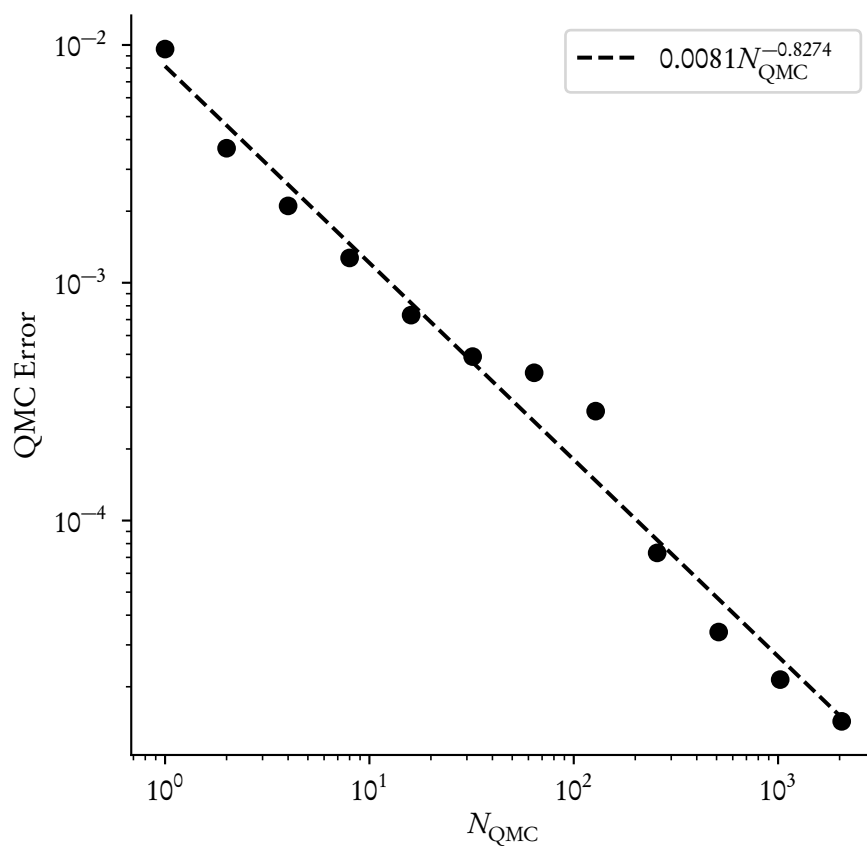


Figure E.8: Quasi-Monte-Carlo error with increasing number of Quasi-Monte-Carlo points, for  $Q = u(0)$  and  $k = 20$ .

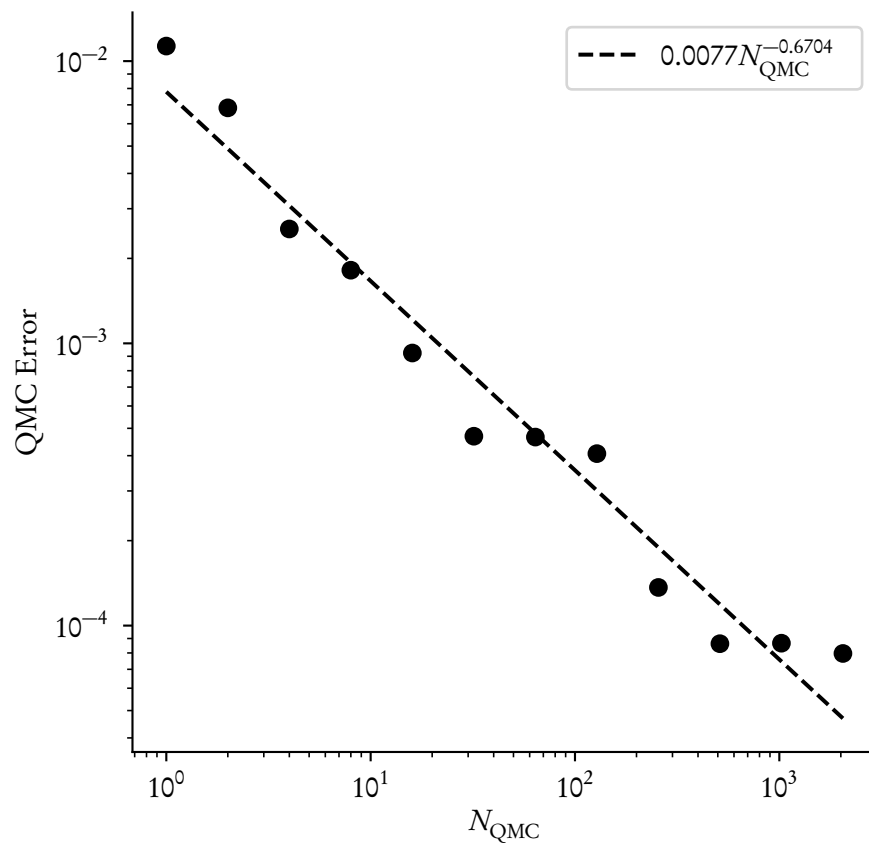


Figure E.9: Quasi-Monte-Carlo error with increasing number of Quasi-Monte-Carlo points, for  $Q = u(0)$  and  $k = 30$ .

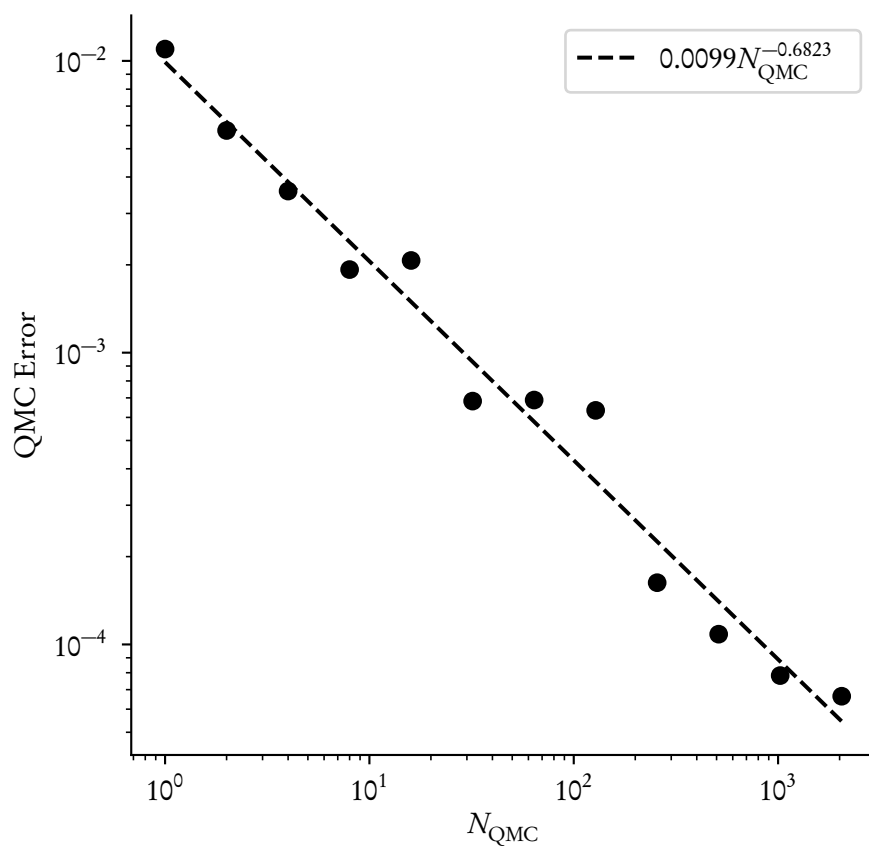


Figure E.10: Quasi-Monte-Carlo error with increasing number of Quasi-Monte-Carlo points, for  $Q = u(0)$  and  $k = 40$ .

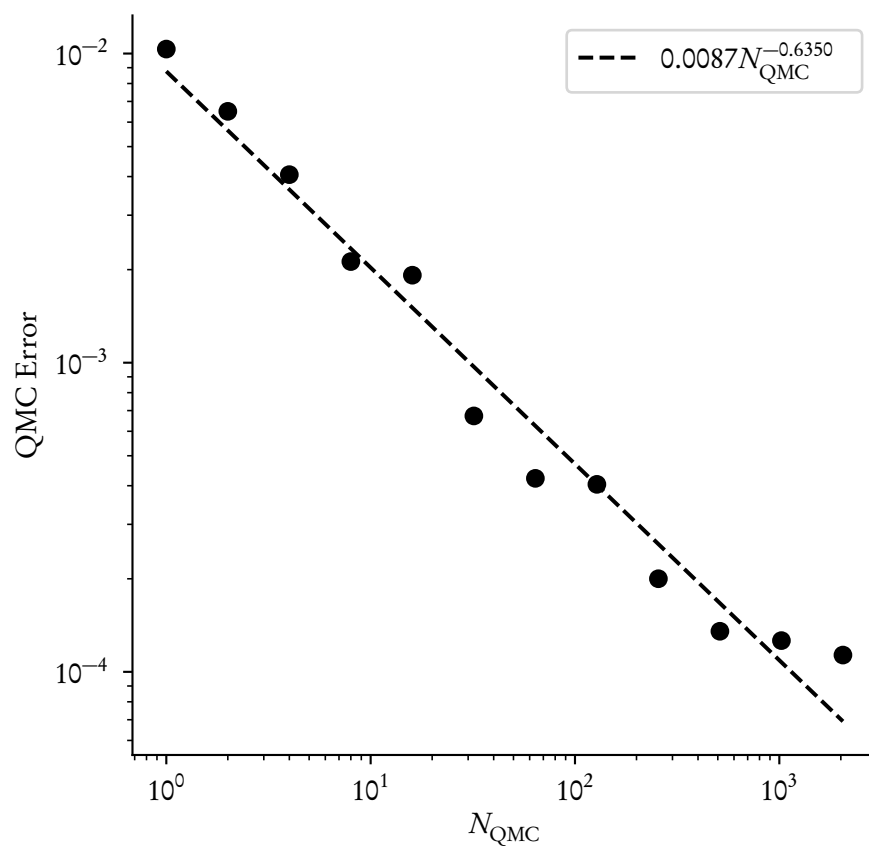


Figure E.11: Quasi-Monte-Carlo error with increasing number of Quasi-Monte-Carlo points, for  $Q = u(0)$  and  $k = 50$ .

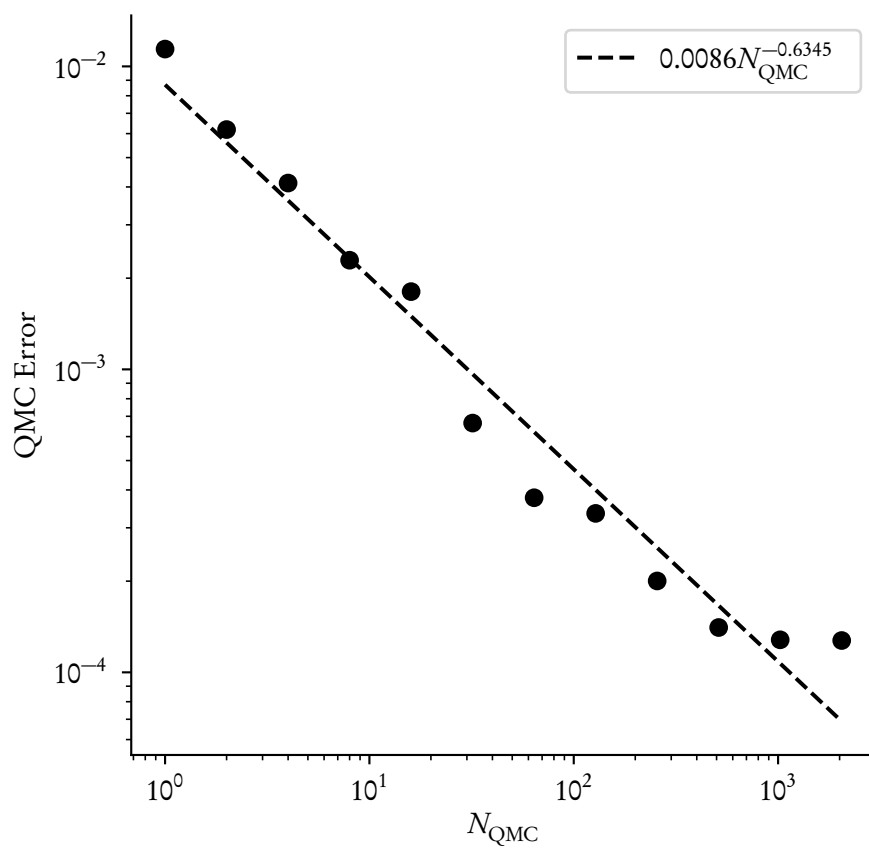


Figure E.12: Quasi-Monte-Carlo error with increasing number of Quasi-Monte-Carlo points, for  $Q = u(0)$  and  $k = 60$ .

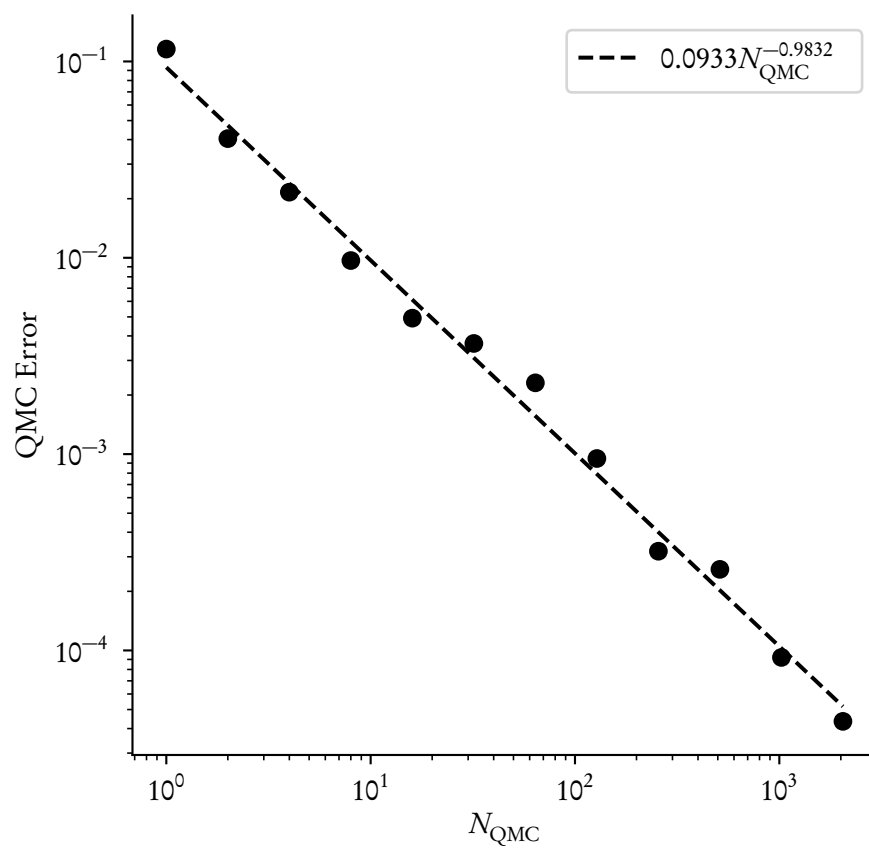


Figure E.13: Quasi-Monte-Carlo error with increasing number of Quasi-Monte-Carlo points, for  $Q = u((1, 1))$  and  $k = 10$ .

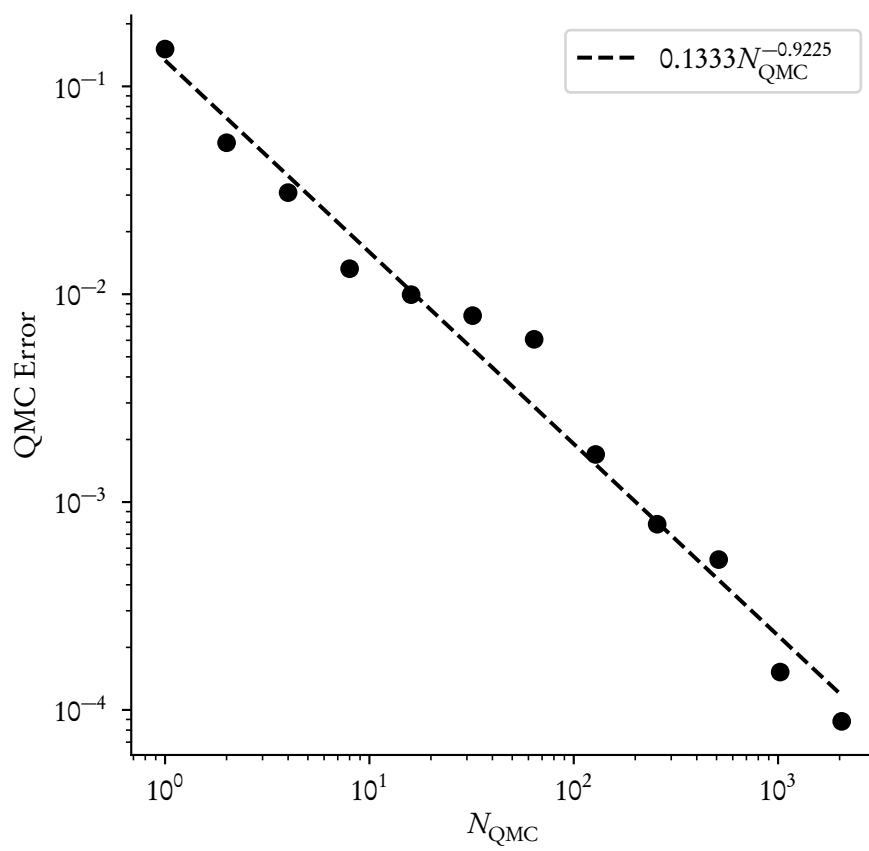


Figure E.14: Quasi-Monte-Carlo error with increasing number of Quasi-Monte-Carlo points, for  $Q = u((1, 1))$  and  $k = 20$ .

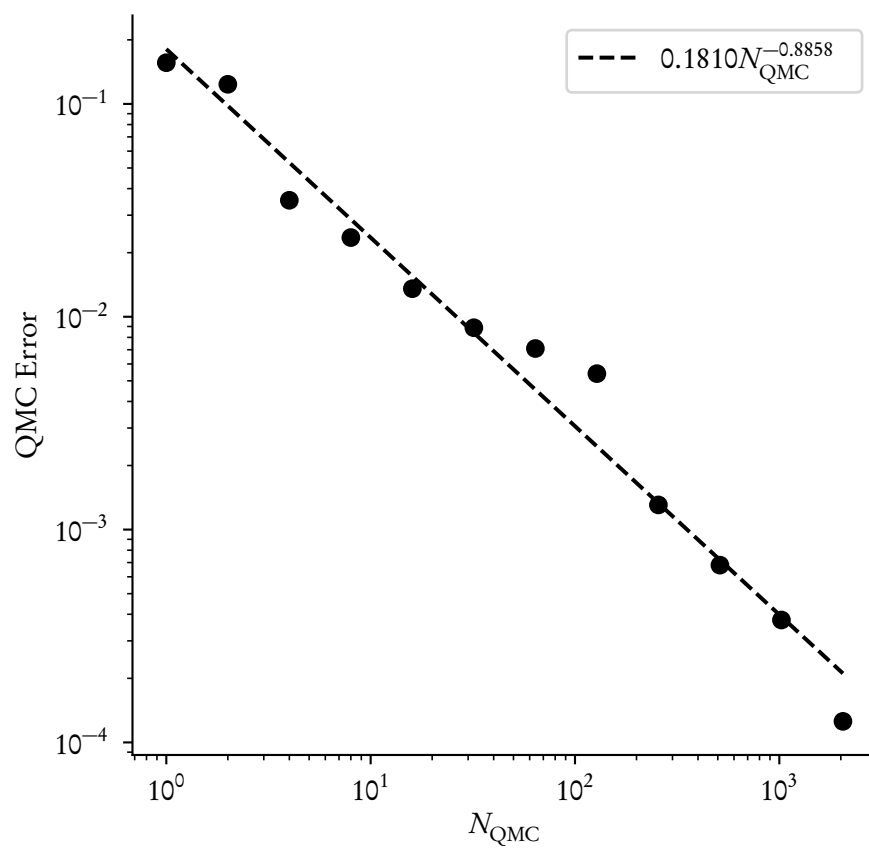


Figure E.15: Quasi-Monte-Carlo error with increasing number of Quasi-Monte-Carlo points, for  $Q = u((1, 1))$  and  $k = 30$ .

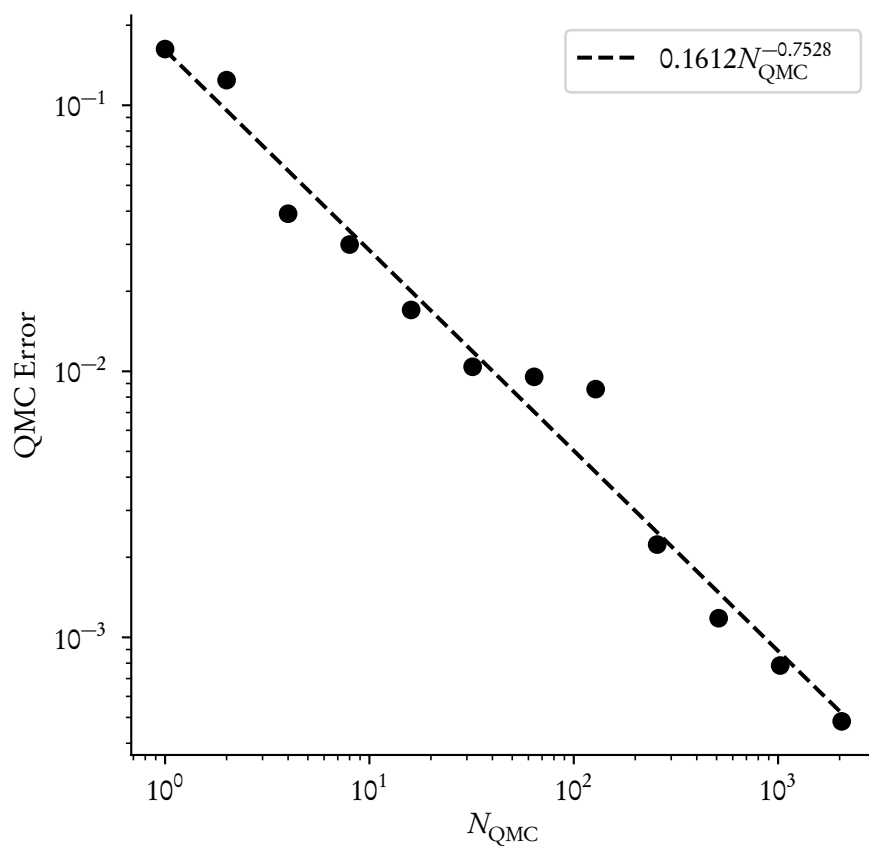


Figure E.16: Quasi-Monte-Carlo error with increasing number of Quasi-Monte-Carlo points, for  $Q = u((1, 1))$  and  $k = 40$ .

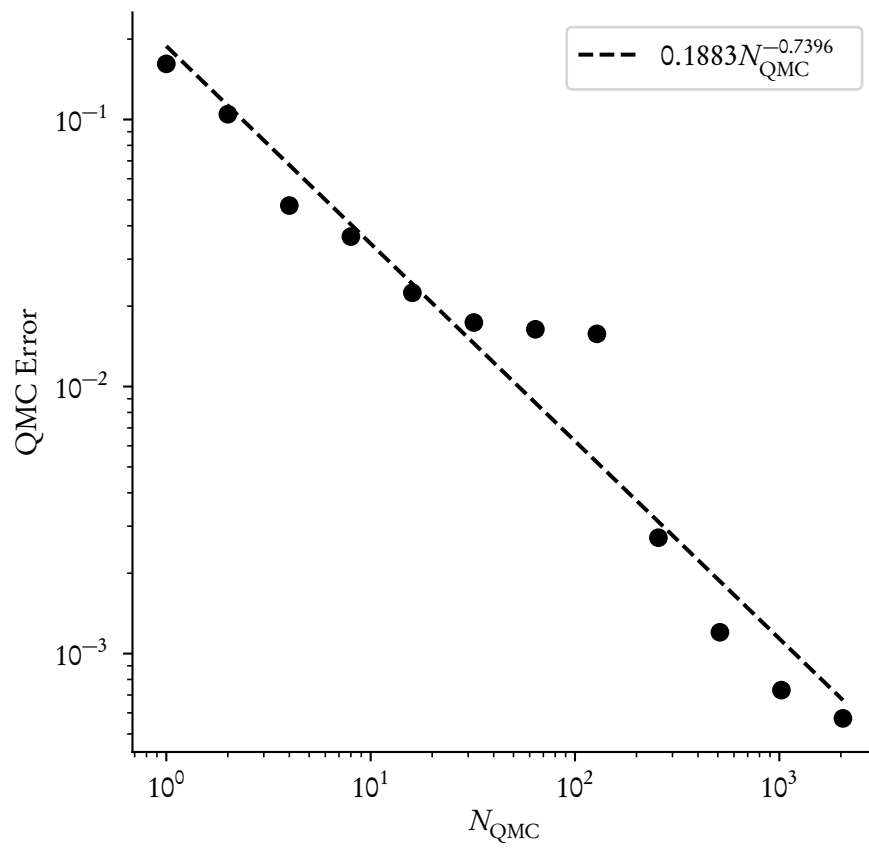


Figure E.17: Quasi-Monte-Carlo error with increasing number of Quasi-Monte-Carlo points, for  $Q = u((1, 1))$  and  $k = 50$ .

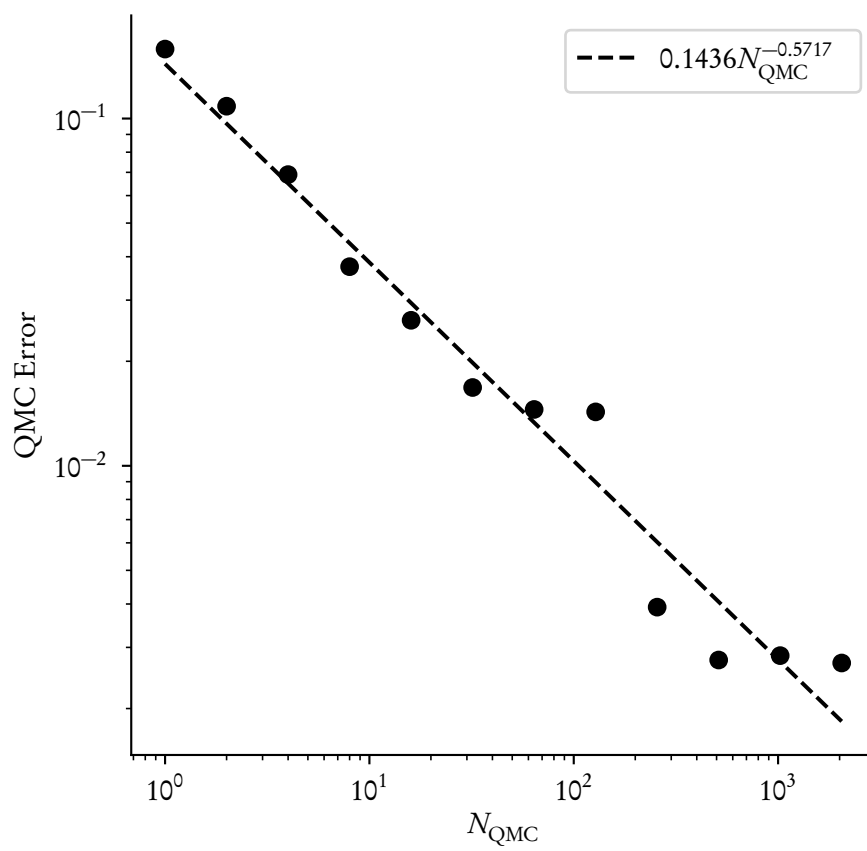


Figure E.18: Quasi-Monte-Carlo error with increasing number of Quasi-Monte-Carlo points, for  $Q = u((1, 1))$  and  $k = 60$ .

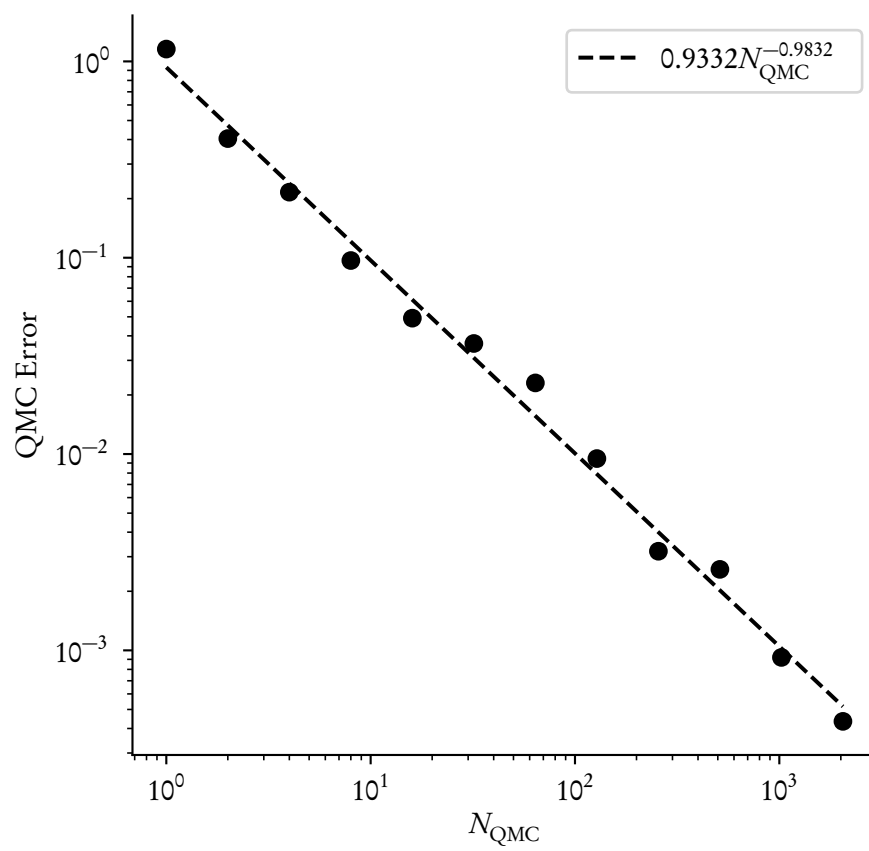


Figure E.19: Quasi-Monte-Carlo error with increasing number of Quasi-Monte-Carlo points, for  $Q = \nabla u((1, 1))$  and  $k = 10$ .

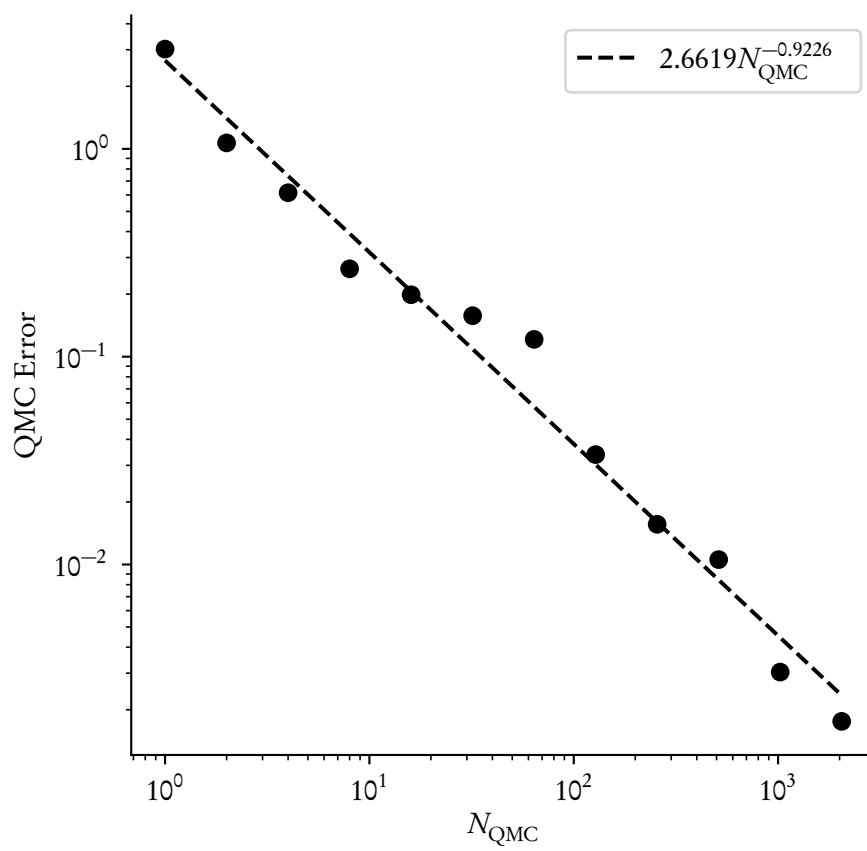


Figure E.20: Quasi-Monte-Carlo error with increasing number of Quasi-Monte-Carlo points, for  $Q = \nabla u((1, 1))$  and  $k = 20$ .

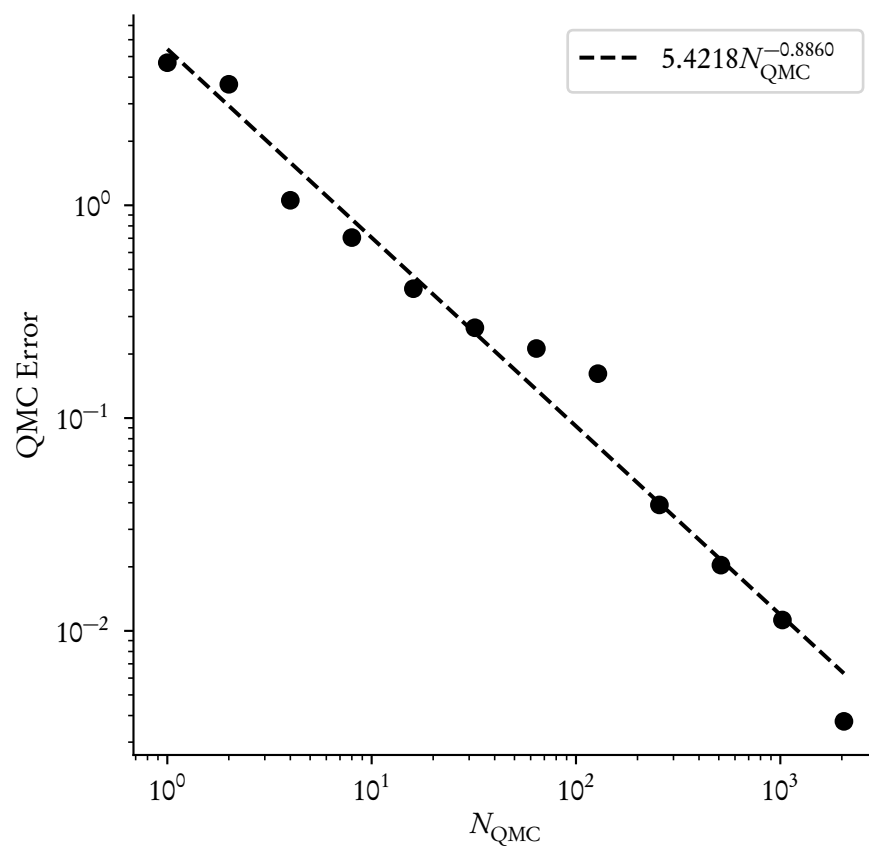


Figure E.21: Quasi-Monte-Carlo error with increasing number of Quasi-Monte-Carlo points, for  $Q = \nabla u((1, 1))$  and  $k = 30$ .

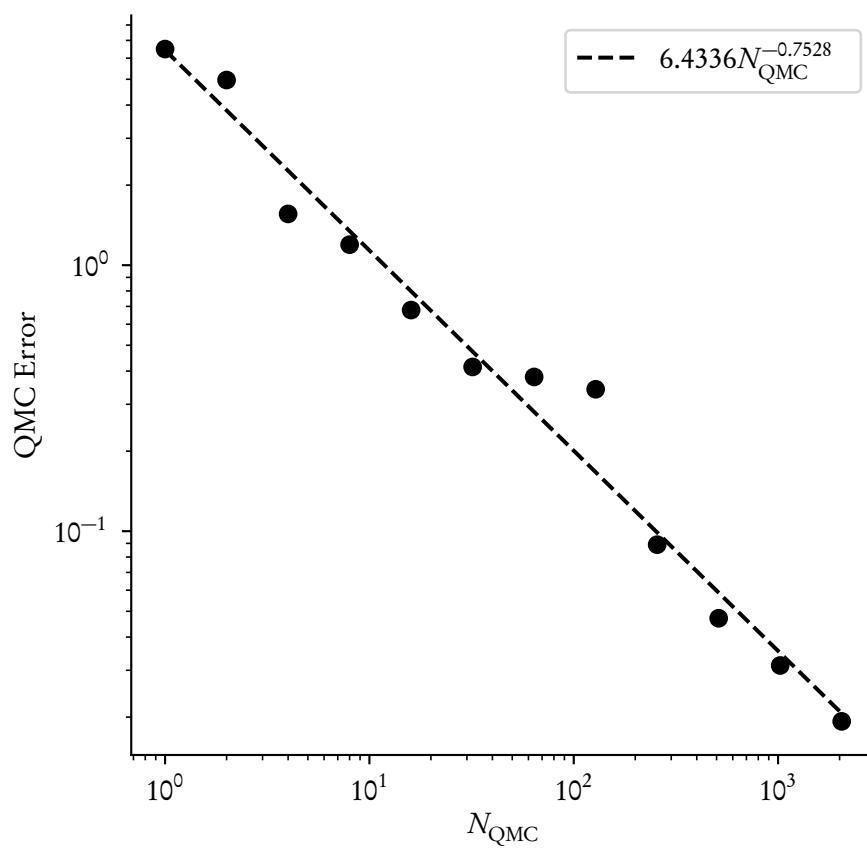


Figure E.22: Quasi-Monte-Carlo error with increasing number of Quasi-Monte-Carlo points, for  $Q = \nabla u((1, 1))$  and  $k = 40$ .

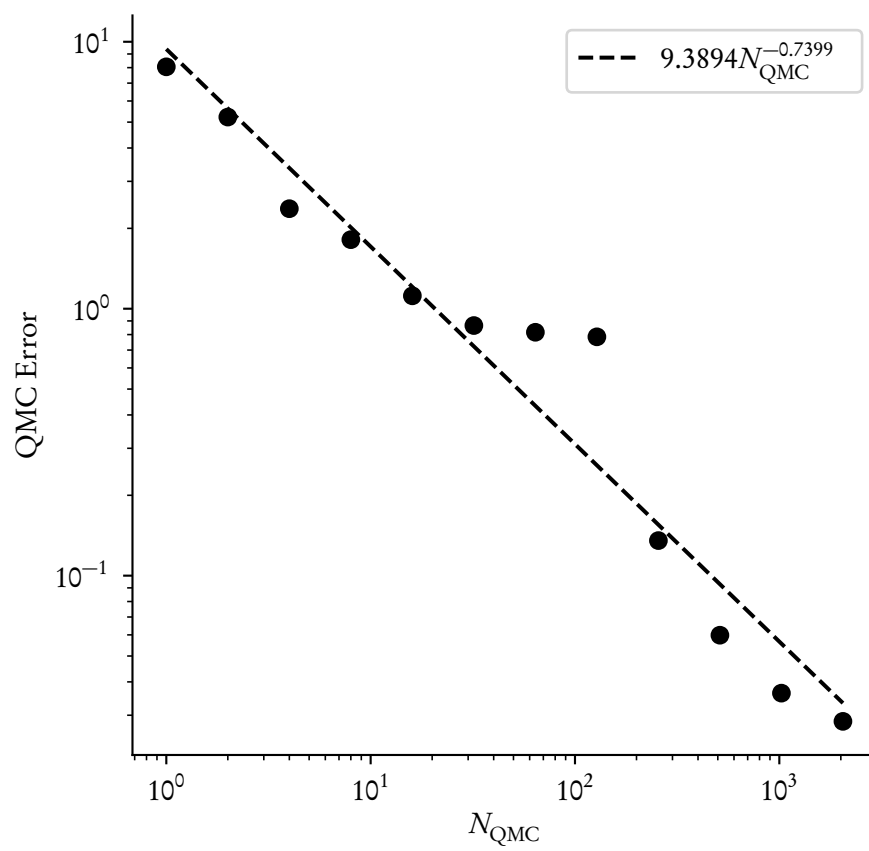


Figure E.23: Quasi-Monte-Carlo error with increasing number of Quasi-Monte-Carlo points, for  $Q = \nabla u((1, 1))$  and  $k = 50$ .

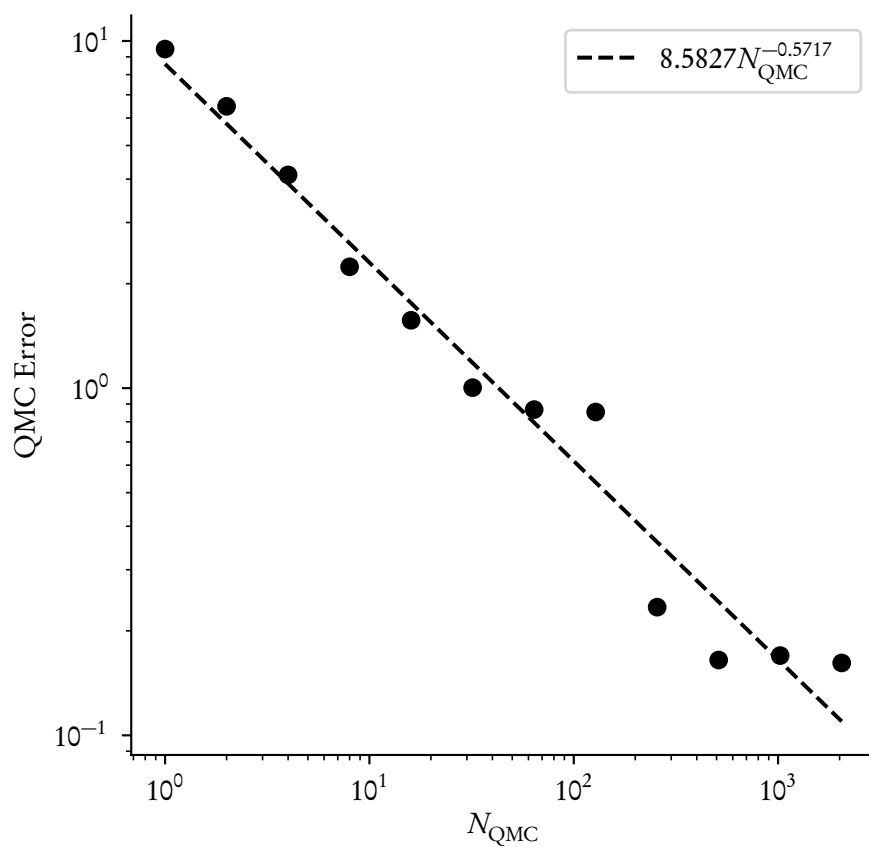


Figure E.24: Quasi-Monte-Carlo error with increasing number of Quasi-Monte-Carlo points, for  $Q = \nabla u((1, 1))$  and  $k = 60$ .

# Additional probabilistic results for nearby preconditioning

**Lemma F.1** (Maximum number of GMRES iterations). *Let  $0 < \varepsilon < 1$ ,  $n^{(2)} : \Omega \rightarrow L^\infty(D; \mathbb{R})$  be a random field, and  $n_1, D_-,$  and  $f$  be as in Problem 4.1, and let  $N$  denote the number of degrees of freedom, i.e. the size of the matrices  $A^{(1)}$  and  $A^{(2)}$ . Then there exists a function  $G_\varepsilon : \mathbb{R}^+ \rightarrow [0, N]$  such that*

$$\text{GMRES}(\varepsilon, n^{(1)}, n^{(2)}) \leq G_\varepsilon(n^{(1)} - n^{(2)}).$$

Moreover,  $G_\varepsilon$  is given by

$$G_\varepsilon\left(\|n^{(1)} - n^{(2)}\|_{L^\infty(D; \mathbb{R})}\right) = \begin{cases} \min\left\{N, \frac{\ln \varepsilon}{\ln\left(\frac{2\alpha^{1/2}}{(1+\alpha)^2}\right)} + 1\right\} & \text{if } \alpha < 1 \\ N & \text{if } \alpha \geq 1, \end{cases} \quad (\text{F.1})$$

where  $\alpha = C_2 k \|n^{(1)} - n^{(2)}\|_{L^\infty(D; \mathbb{R})}$ , where  $C_2$  is given by (4.32).

See Figure F.1 for some examples of the function  $G_\varepsilon$ .

The proof of Lemma F.1 uses the following corollary [187, Corollary 3] of [187, Proposition 2] on the ‘lucky breakdown’ of GMRES.

**Corollary F.2** (Guaranteed convergence of GMRES). *For an  $N \times N$  problem GMRES converges in at most  $N$  iterations.*

*Proof of Lemma F.1.* For  $\alpha \geq 1$ , the result is immediate from Corollary F.2. For  $\alpha < 1$ , if we insert (4.60) (the corollary of the Elman estimate) into the Elman estimate (4.59) (with  $D = 1$ , so  $\|\cdot\|_D = \|\cdot\|_2$ ), we obtain, for  $m \in \mathbb{N}$

$$\frac{\|r_m\|_2}{\|r_0\|_2} \leq \left(\frac{2\sqrt{\alpha}}{(1+\alpha)^2}\right)^m. \quad (\text{F.2})$$

To obtain a bound on the number of iterations needed to obtain the solution to within a tolerance  $\varepsilon$ , we set the right-hand side of (F.2) to be less than  $\varepsilon$  and solve for  $m$  to obtain that the GMRES residual is less than  $\varepsilon$  (recall we assume  $\|r_0\|_2 = 1$ , see Definition 4.39) if

$$m \geq \frac{\ln \varepsilon}{\ln\left(\frac{2\alpha^{1/2}}{(1+\alpha)^2}\right)}. \quad (\text{F.3})$$

Hence, if  $m^*$  is the smallest integer satisfying (F.3), then

$$m^* \leq \frac{\ln \varepsilon}{\ln\left(\frac{2\alpha^{1/2}}{(1+\alpha)^2}\right)} + 1. \quad (\text{F.4})$$

The result for  $\alpha < 1$  therefore follows from (F.4) and Corollary F.2, since GMRES will have converged to within a tolerance  $\varepsilon$  within  $m^*$  iterations.  $\square$

**Remark F.3** (Why not use the ceiling in (F.4)?). *One could replace the bound (F.4) by the equality*

$$m^* = \left\lceil \frac{\ln \varepsilon}{\ln\left(\frac{2\alpha^{1/2}}{(1+\alpha)^2}\right)} \right\rceil.$$

*However, the change in the definition of  $G_\varepsilon$  (F.1) would mean  $G_\varepsilon$  would only be piecewise continuous. As we must use numerical methods to calculate probabilities associated with  $G_\varepsilon$  (see Theorem F.5 and Remark F.6 below), it is convenient if  $G_\varepsilon$  is continuous, and so we use (F.4).*

**Remark F.4** (Why the dependence on  $\alpha$  in Lemma F.1?). *The reason that (F.1) has two cases depending on  $\alpha = C_2 k \left\| n^{(1)} - n^{(2)} \right\|_{L^\infty(D; \mathbb{R})}$  is because Corollary 4.26 only holds if  $\alpha < 1$ . Therefore if  $\alpha \geq 1$  the only result available to us is Corollary F.2.*

Lemma F.1 gives us the relationship between the (bound on the) number of GMRES iterations required for convergence and  $\left\| n^{(1)} - n^{(2)}(\omega) \right\|_{L^\infty(D; \mathbb{R})}$ . We can use this relationship to infer probabilistic properties of the number of GMRES iterations required for convergence from the probability distribution of  $\left\| n^{(1)} - n^{(2)}(\omega) \right\|_{L^\infty(D; \mathbb{R})}$ . (For the probabilistic notation, we refer the reader to Chapter 3.)

**Theorem F.5** (Probabilistic GMRES convergence). *Let  $n^{(1)} \in L^\infty(D; \mathbb{R})$  be fixed, and let  $n^{(2)} : \Omega \rightarrow L^\infty(D; \mathbb{R})$  be a random field. Let  $\varepsilon$  and  $N$  be as in Lemma F.1, and let  $A^{(1)} = A^{(2)} = I$ . Fix  $R \in \mathbb{N}$ . Then*

$$\mathbb{P}\left(G_\varepsilon\left(\left\| n^{(1)} - n^{(2)} \right\|_{L^\infty(D; \mathbb{R})}\right) \leq R\right) \leq \mathbb{P}\left(\text{GMRES}(\varepsilon, n^{(1)}, n^{(2)}) \leq R\right). \quad (\text{F.5})$$

*Proof of Theorem F.5.* By Lemma F.1 we have the implication: if  $G_\varepsilon(n^{(1)} - n^{(2)}(\omega)) \leq R$ , then  $\text{GMRES}(\varepsilon, n^{(1)}, n^{(2)}(\omega)) \leq R$ . Therefore we have the set inclusion

$$\left\{ \omega \in \Omega : G_\varepsilon(n^{(1)} - n^{(2)}(\omega)) \leq R \right\} \subseteq \left\{ \omega \in \Omega : \text{GMRES}(\varepsilon, n^{(1)}, n^{(2)}(\omega)) \leq R \right\}.$$

The result immediately follows.  $\square$

**Remark F.6** (The expression (F.5) is computable). *Because the function  $G_\varepsilon$  is not invertible (as is clear from Figure F.1), one cannot write the left-hand side of (F.5) as*

$$\mathbb{P}\left(\left\| n^{(1)} - n^{(2)} \right\|_{L^\infty(D; \mathbb{R})} \leq G_\varepsilon^{-1}([0, R])\right).$$

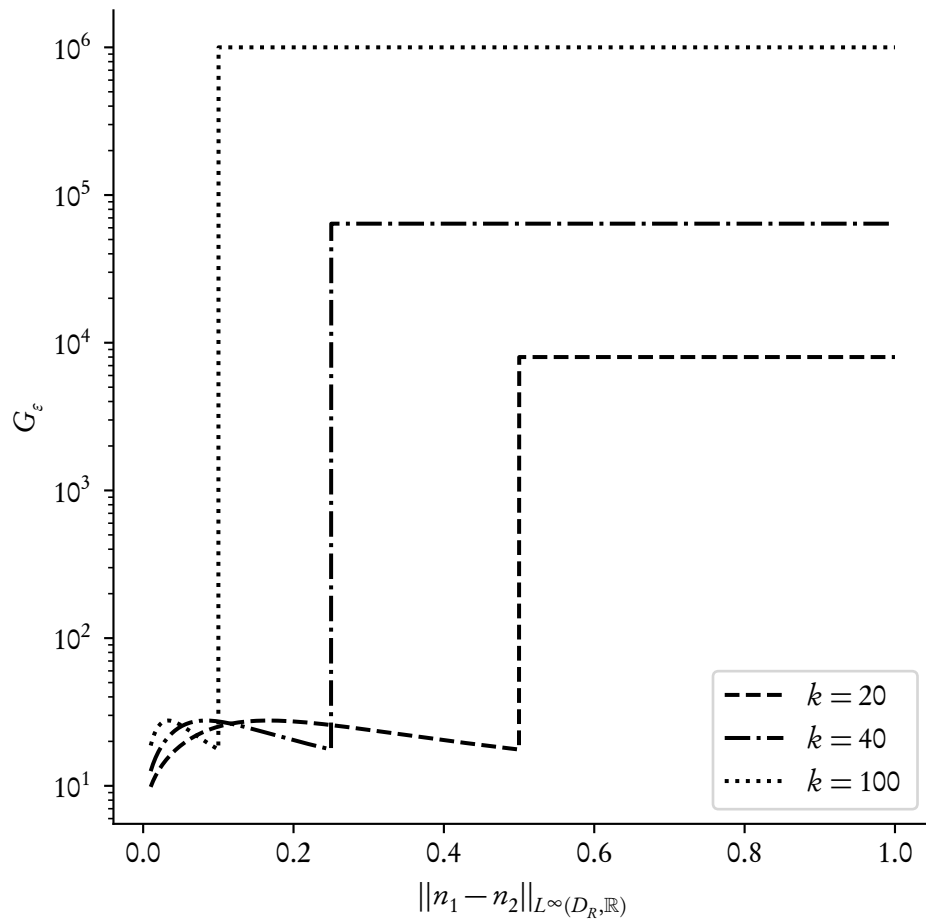


Figure F.1: The function  $G_\varepsilon$  for  $\|n_1 - n_2\|_{L^\infty(D; \mathbb{R})} \in (0.01, 1.0)$ , for  $k = 20, 40, 100$ ,  $C_2 = 0.1$ ,  $N = \lceil k^3 \rceil$ , and  $\varepsilon = 10^{-5}$ .

However, one can still compute the set

$$G_\varepsilon^{-1}([0, R]) = \{\alpha : G_\varepsilon(\alpha) \in [0, R]\} \quad (\text{F.6})$$

(where  $G_\varepsilon^{-1}$  in (F.6) denotes the pullback), and therefore one can compute the probabilities in (F.5). The main effort in computing  $G_\varepsilon^{-1}([0, R])$  is finding if there are any values of  $\alpha < 1$  such that  $G_\varepsilon(\alpha) = R$ , since the existence, or not, of such values determines the range of  $\alpha$  over which we must integrate. However, these values can be computed numerically using standard root-finding algorithms.

## Computational set-up

All of the computations in this thesis were performed with the following computational setup, unless otherwise stated.

*PDE/FEM* The PDE we solve is the Interior Impedance Problem, i.e., Problem 2.12 with  $D_- = \emptyset$ , posed on the 2-d unit square;  $D = [0, 1]^2$ . We use first-order continuous finite elements, with  $h = k^{-3/2}$ . We use regular grids, see Figure G.1 for an example grid. Where we needed to calculate a preconditioner  $(A^{(1)})^{-1}$ , we calculated the exact  $LU$  decomposition of  $A^{(1)}$ .

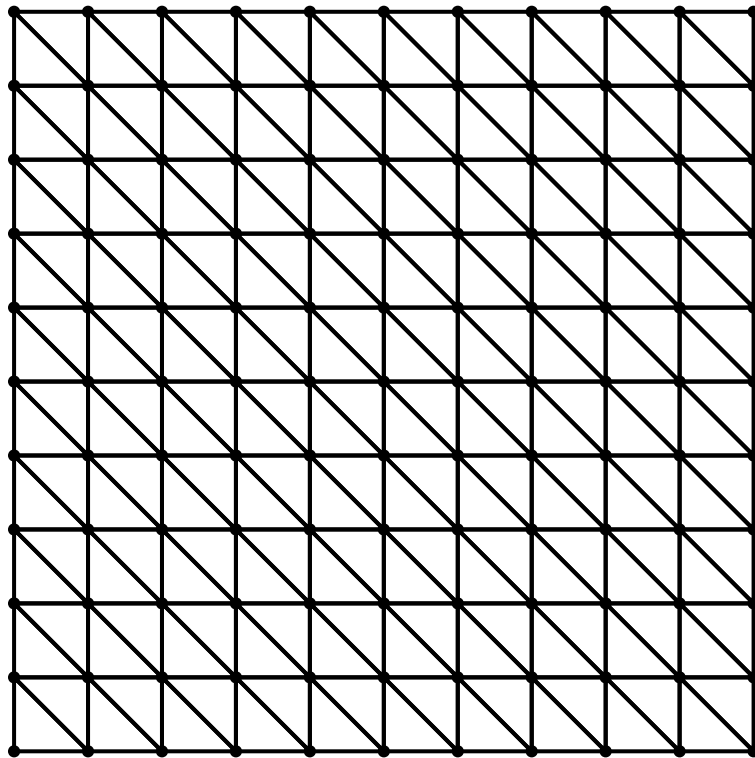


Figure G.1: A sample mesh, similar to those used in all the computations in this thesis.

*Numerical setup* All finite-element calculations were carried out using the Firedrake software library [183, 144], which uses PETSc to perform its linear solves [12, 11, 53, 13] and Chaco [113] to perform graph partitions. PETSc uses the MUMPS [2, 3] solver to perform  $LU$  factorisations and direct solves. When GMRES was used, the stopping criterion was a relative error (relative to the 2-norm of the right-hand side) of  $10^{-5}$  or an absolute error of  $10^{-50}$ . To generate QMC points, we made use of Dirk Nuyen’s ‘Magic Point Shop’ code [161, 131], which uses a base-2 lattice sequence with generating vector from [51]. Many of the computations were carried out on the Balena High Performance Computing (HPC) Service at the University of Bath.

*Code Access* The  $\LaTeX$  files used to produce this thesis, along with all the code required to produce the figures and tables, and to perform the numerical experiments can be found at <https://github.com/orpembery/thesis> and in other repositories accessible from that repository. Snapshots of all the repositories and data used in the production of this thesis can be found at [168, 169, 171, 170, 175, 173, 172, 174].

# Bibliography

- [1] Mark Ainsworth. Discrete Dispersion Relation for  $hp$ -Version Finite Element Approximation at High Wave Number. *SIAM Journal on Numerical Analysis*, 42(2):553–575, 2004.
- [2] Patrick R. Amestoy, Iain S. Duff, Jean-Yves L'Excellent, and Jacko Koster. A Fully Asynchronous Multifrontal Solver Using Distributed Dynamic Scheduling. *SIAM Journal on Matrix Analysis and Applications*, 23(1):15–41, 2001.
- [3] Patrick R. Amestoy, Abdou Guermouche, Jean-Yves L'Excellent, and Stéphane Pralet. Hybrid scheduling for the parallel solution of linear systems. *Parallel Computing*, 32(2):136–156, 2006.
- [4] Fred Aminzadeh, Jean Brac, and Tim Kunz. *3-D Salt and Overthrust Models*, volume 1 of *SEG/EAGE 3-D Modeling Series*. Society of Exploration Geophysicists, 1997.
- [5] Jean-Pierre Aubin. Behaviour of the error of the approximate solutions of boundary value problems for linear elliptic operators by Galerkin's and finite difference methods. *Annali della Scuola Normale Superiore di Pisa, Classe di Scienze Série 3*, 21(4):599–637, 1967.
- [6] Abdul Kadir Aziz, R. Bruce Kellogg, and Arthur Brooke Stephens. A Two Point Boundary Value Problem with a Rapidly Oscillating Solution. *Numerische Mathematik*, 53(1–2):107–121, 1988.
- [7] I. Babuška, Fabio Nobile, and Raul F. Tempone. A Stochastic Collocation Method for Elliptic Partial Differential Equations with Random Input Data. *SIAM Journal on Numerical Analysis*, 45(3):1005–1034, 2007.
- [8] Ivo M. Babuška, Raul F. Tempone, and Georgios E. Zouraris. Galerkin Finite Element Approximations of Stochastic Elliptic Partial Differential Equations. *SIAM Journal on Numerical Analysis*, 42(2):800–825, 2004.
- [9] George Bachmann, Lawrence Narici, and Edward Beckenstein. *Fourier and Wavelet Analysis*. Universitext. Springer Science+Business Media, New York, 2000.
- [10] Trygve Bærland, Miroslav Kuchta, and Kent-Andre Mardal. Multigrid Methods for Discrete Fractional Sobolev Spaces. *SIAM Journal on Scientific Computing*, 41(2):A948–A972, 2019.
- [11] Satish Balay, William D. Gropp, Lois Curfman McInnes, and Barry F. Smith. Efficient Management of Parallelism in Object Oriented Numerical Software Libraries. In E. Arge, A. M. Bruaset, and H. P. Langtangen, editors, *Modern Software Tools in Scientific Computing*, pages 163–202. Birkhäuser Press, 1997.

- [12] Satish Balay, Shrirang Abhyankar, Mark F. Adams, Jed Brown, Peter Brune, Kris Buschelman, Lisandro Dalcin, Victor Eijkhout, William D. Gropp, Dinesh Kaushik, Matthew G. Knepley, Dave A. May, Lois Curfman McInnes, Richard Tran Mills, Todd Munson, Karl Rupp, Patrick Sanan, Barry F. Smith, Stefano Zampini, and Hong Zhang. PETSc Users Manual. Technical Report ANL-95/11 - Revision 3.9, Argonne National Laboratory, 2018. URL <https://www.mcs.anl.gov/petsc/petsc-current/docs/manual.pdf>.
- [13] Satish Balay, Shrirang Abhyankar, Mark F. Adams, Jed Brown, Peter Brune, Kris Buschelman, Lisandro Dalcin, Alp Dener, Victor Eijkhout, William D. Gropp, Dmitry Karpeyev, Dinesh Kaushik, Matthew G. Knepley, Dave A. May, Lois Curfman McInnes, Richard Tran Mills, Todd Munson, Karl Rupp, Patrick Sanan, Barry F. Smith, Stefano Zampini, and Hong Zhang. PETSc Web page, 2019. URL <https://www.mcs.anl.gov/petsc>. Last accessed 04/09/2019.
- [14] Gang Bao, Yanzhao Cao, Yongle Hao, and Kai Zhang. A Robust Numerical Method for the Random Interface Grating Problem via Shape Calculus, Weak Galerkin Method, and Low-Rank Approximation. *Journal of Scientific Computing*, 77(1), 2018.
- [15] Andrea Barth, Christoph Schwab, and Nathaniel Zollinger. Multi-level Monte Carlo Finite Element method for elliptic PDEs with stochastic coefficients. *Numerische Mathematik*, 119(1):123–161, 2011.
- [16] H el ene Barucq, Th eophile Chaumont-Frelet, and Christian Gout. Stability analysis of heterogeneous Helmholtz problems and finite element solution based on propagation media approximation. *Mathematics of Computation*, 86(307):2129–2157, 2017.
- [17] Dean Baskin, Euan A. Spence, and Jared Wunsch. Sharp High-Frequency Estimates for the Helmholtz Equation and Applications to Boundary Integral Equations. *SIAM Journal on Mathematical Analysis*, 48(1):229–267, 2016.
- [18] Alvin Bayliss, Charles I. Goldstein, and Eli Turkel. On Accuracy Conditions for the Numerical Computation of Waves. *Journal of Computational Physics*, 59(3):396–404, 1985.
- [19] Bernhard Beckkermann, Sergei A. Goreinov, and Eugene E. Tyrtshnikov. Some Remarks on the Elman Estimate for GMRES. *SIAM Journal on Matrix Analysis and Applications*, 27(3):772–778, 2006.
- [20] Ali Behzadan and Michael J. Holst. Multiplication in Sobolev Spaces, Revisited. *arXiv preprint arXiv:1512.07379*, 2017.
- [21] Mourad Bellassoued. Carleman estimates and distribution of resonances for the transparent obstacle and application to the stabilization. *Asymptotic Analysis*, 35(3, 4):257–279, 2003.
- [22] Jean-Pierre Berenger. A Perfectly Matched Layer for the Absorption of Electromagnetic Waves. *Journal of Computational Physics*, 114(2):185–200, 1994.

- [23] Timo Betcke, Simon N. Chander-Wilde, Ivan G. Graham, Stephen Langdon, and Marko Lindner. Condition Number Estimates for Combined Potential Integral Operators in Acoustics and Their Boundary Element Discretisation. *Numerical Methods for Partial Differential Equations*, 27(1):31–69, 2011.
- [24] Clifford O. Bloom. Estimates for Solutions of Reduced Hyperbolic Equations of the Second Order with a Large Parameter. *Journal of Mathematical Analysis and Applications*, 44(2):310–332, 1973.
- [25] Clifford O. Bloom and Nicholas D. Kazarinoff. A priori bounds for solutions of the Dirichlet problem for  $[\Delta + \lambda^2 n(x)]u = f(x, \lambda)$  on an exterior domain. *Journal of Differential Equations*, 24(3):437–465, 1977.
- [26] Vladimir I. Bogachev. *Measure Theory*. Springer, Berlin Heidelberg, 2007.
- [27] Marcella Bonazzoli, Victorita Dolean, Ivan G. Graham, Euan A. Spence, and Pierre-Henri Tournier. Domain decomposition preconditioning for the high-frequency time-harmonic Maxwell equations with absorption. *Mathematics of Computation*, 88(320):2559–2604, 2019.
- [28] Dietrich Braess. *Finite Elements: Theory, Fast Solvers, and Applications in Elasticity Theory*. Cambridge University Press, Cambridge, UK, 3rd edition, 2007.
- [29] Susanne C. Brenner and L. Ridgway Scott. *The Mathematical Theory of Finite Element Methods*, volume 15 of *Texts in Applied Mathematics*. Springer Science+Business Media, New York, 3rd edition, 2008.
- [30] Donald L. Brown, Dietmar Gallistl, and Daniel Peterseim. Multiscale Petrov–Galerkin Method for High-Frequency Heterogeneous Helmholtz Equations. In Michael Griebel and Marc Alexander Schweitzer, editors, *Meshfree Methods for Partial Differential Equations VIII*, Lecture Notes in Computational Science and Engineering, pages 85–115, Cham, Switzerland, 2017. Springer.
- [31] BS EN ISO 16810:2014. Non-destructive testing — Ultrasonic testing — General principles (ISO 16810:2012). Standard, The British Standards Institution, 2014.
- [32] Tan Bui-Thanh and Omar Ghattas. An Analysis of Infinite Dimensional Bayesian Inverse Shape Acoustic Scattering and Its Numerical Approximation. *SIAM/ASA Journal on Uncertainty Quantification*, 2(1):203–222, 2014.
- [33] Nicolas Burq. Décroissance de l'énergie locale de l'équation des ondes pour le problème extérieur et absence de résonance au voisinage du réel. *Acta Mathematica*, 180(1):1–29, 1998.
- [34] Nicolas Burq. Semi-Classical Estimates for the Resolvent in Nontrapping Geometries. *International Mathematics Research Notices*, 2002(5):221–241, 2002.

- [35] Xiao-Chuan Cai and Olof B. Widlund. Domain decomposition algorithms for indefinite elliptic problems. *SIAM Journal on Scientific and Statistical Computing*, 13(1):243–258, 1992.
- [36] Yves Capdeboscq. On the scattered field generated by a ball inhomogeneity of constant index. *Asymptotic Analysis*, 77(3–4):197–246, 2012.
- [37] Fernando Cardoso, Georgi Popov, and Georgi Vodev. Distribution of resonances and local energy decay in the transmission problem II. *Mathematical Research Letters*, 6:377–396, 1999.
- [38] Simon N. Chandler-Wilde and Peter Monk. Wave-Number-Explicit Bounds in Time-Harmonic Scattering. *SIAM J. Math. Anal.*, 39(5):1428–1455, 2008.
- [39] Simon N. Chandler-Wilde, Ivan G. Graham, Stephen Langdon, and Euan A. Spence. Numerical-asymptotic boundary integral methods in high-frequency acoustic scattering. *Acta Numerica*, 21(1):89–305, 2012.
- [40] Simon N. Chandler-Wilde, Euan A. Spence, Andrew Gibbs, and Valery P. Smyshlyaev. High-frequency bounds for the Helmholtz equation under parabolic trapping and applications in numerical analysis. *SIAM Journal on Mathematical Analysis*, to appear, 2019.
- [41] Julia Charrier. Strong and Weak Error Estimates for Elliptic Partial Differential Equations with Random Coefficients. *SIAM Journal on Numerical Analysis*, 50(1):216–246, 2012.
- [42] Julia Charrier, Robert Scheichl, and Aretha L. Teckentrup. Finite Element Error Analysis of Elliptic PDEs with Random Coefficients and Its Application to Multilevel Monte Carlo Methods. *SIAM Journal on Numerical Analysis*, 51(1):322–352, 2013.
- [43] Théophile Chaumont-Frelet. *Approximation par éléments finis de problèmes d’Helmholtz pour la propagation d’ondes sismiques*. PhD thesis, INSA Rouen, 2015.
- [44] Théophile Chaumont-Frelet and Serge Nicaise. High-frequency behaviour of corner singularities in Helmholtz problems. *ESAIM: Mathematical Modelling and Numerical Analysis*, 52(5):1803–1845, 2018.
- [45] Théophile Chaumont-Frelet and Serge Nicaise. Wavenumber explicit convergence analysis for finite element discretizations of general wave propagation problem. *IMA Journal of Numerical Analysis*, to appear, 2019.
- [46] Théophile Chaumont-Frelet, Dietmar Gallistl, Serge Nicaise, and Jérôme Tomezyk. Wavenumber explicit convergence analysis for finite element discretizations of time-harmonic wave propagation problems with perfectly matched layers. *HAL preprint hal-01887267*, 2018.

- [47] Théophile Chaumont-Frelet, Serge Nicaise, and Jérôme Tomezyk. Uniform a priori estimates for elliptic problems with impedance boundary conditions. *Communications on Pure and Applied Mathematics*, to appear, 2019.
- [48] Phillippe G. Ciarlet. *Basic Error Estimates for Elliptic Problems*, volume II of *Handbook of Numerical Analysis*, pages 18–351. North–Holland, Amsterdam, 1991.
- [49] K. Andrew Cliffe, Michael B. Giles, Robert Scheichl, and Aretha L. Teckentrup. Multi-level Monte Carlo methods and applications to elliptic PDEs with random coefficients. *Computing and Visualization in Science*, 14(1):3–15, 2011.
- [50] David Colton and Rainer Kress. *Inverse Acoustic and Electromagnetic Scattering Theory*, volume 93 of *Applied Mathematical Sciences*. Springer Science+Business Media, New York, 3rd edition, 2013.
- [51] Ronald Cools, Frances Y. Kuo, and Dirk Nuyens. Constructing Embedded Lattice Rules for Multivariate Integration. *SIAM Journal on Scientific Computing*, 28(6):2162–2188, 2006.
- [52] Peter Cummings and Xiaobing H. Feng. Sharp regularity coefficient estimates for complex-valued acoustic and elastic Helmholtz equations. *Mathematical Models and Methods in Applied Sciences*, 16(1):139–160, 2006.
- [53] Lisandro D. Dalcin, Rodrigo R. Paz, Pablo A. Kler, and Alejandro Cosimo. Parallel distributed computing using Python. *Advances in Water Resources*, 34(9):1124–1139, 2011. New Computational Methods and Software Tools.
- [54] Joseph Diestel and Jerry J. Uhl, Jr. *Vector Measures*, volume 15 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 1977.
- [55] Ganesh C. Diwan, Andrea Moiola, and Euan A. Spence. Can coercive formulations lead to fast and accurate solution of the Helmholtz equation? *Journal of Computational and Applied Mathematics*, 352:110–131, 2019.
- [56] Tim J. Dodwell, Chris Ketelsen, Robert Scheichl, and Aretha L. Teckentrup. A Hierarchical Multilevel Markov Chain Monte Carlo Algorithm with Applications to Uncertainty Quantification in Subsurface Flow. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):1075–1108, 2015.
- [57] Joseph L. Doob. *Measure Theory*, volume 143 of *Graduate Texts in Mathematics*. Springer-Verlag, New York, 1994.
- [58] Jim Douglas Jr., Juan E. Santos, Dongwoo Sheen, and Lynn Schreyer Bennethum. Frequency domain treatment of one-dimensional scalar waves. *Mathematical Models and Methods in Applied Sciences*, 3(2):171–194, 1993.

- [59] Yu Du and Haijun Wu. Preasymptotic Error Analysis of Higher Order FEM and CIP-FEM for Helmholtz Equation with High Wave Number. *SIAM Journal on Numerical Analysis*, 53(2):782–804, 2015.
- [60] Richard M. Dudley. *Real Analysis and Probability*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, UK, 2002.
- [61] Iain S. Duff, Albert M. Erisman, and John K. Reid. On George’s Nested Dissection Method. *SIAM Journal on Numerical Analysis*, 13(5):686–695, 1976.
- [62] Iain S. Duff, Albert M. Erisman, and John K. Reid. *Direct Methods for Sparse Matrices*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, UK, second edition, 2017.
- [63] Michael Eiermann, Oliver G. Ernst, and Elisabeth Ullmann. Computational aspects of the stochastic finite element method. *Computing and Visualization in Science*, 10(1):3–15, 2007.
- [64] Stanley C. Eisenstat, Howard C. Elman, and Martin H. Schultz. Variational Iterative Methods for Nonsymmetric Systems of Linear Equations. *SIAM Journal on Numerical Analysis*, 20:345–357, 1983.
- [65] Nate Eldredge. Measurable functions with values in Banach spaces. Mathematics Stack Exchange. URL <https://math.stackexchange.com/q/18086>. (version: 2011-01-19). Last Accessed 26/07/2019.
- [66] Howard C. Elman. *Iterative Methods for Large, Sparse, Nonsymmetric Systems of Linear Equations*. PhD thesis, Yale University, 1982.
- [67] Howard C. Elman, Christopher W. Miller, Eric T. Phipps, and Raymond S. Tuminaro. Assessment of Collocation and Galerkin Approaches to Linear Diffusion Equations with Random Data. *International Journal for Uncertainty Quantification*, 1(1):19–33, 2011.
- [68] Howard C. Elman, David J. Silvester, and Andrew J. Wathen. *Finite Elements and Fast Iterative Solvers, with Applications in Incompressible Fluid Dynamics*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, UK, second edition, 2014.
- [69] Oliver G. Ernst and Martin J. Gander. *Why it is Difficult to Solve Helmholtz Problems with Classical Iterative Methods*, chapter 10, pages 325–263. Number 83 in Lecture Notes in Computational Science and Engineering. Springer-Verlag, Berlin Heidelberg, 2012.
- [70] Oliver G. Ernst, Catherine E. Powell, David J. Silvester, and Elisabeth Ullmann. Efficient Solvers for a Linear Stochastic Galerkin Mixed Formulation of Diffusion Problems with Random Data. *SIAM Journal on Scientific Computing*, 31(2):1424–1447, 2009.

- [71] Paul Escapil-Inchauspé and Carlos Jerez-Hanckes. Helmholtz scattering by random domains: first-order sparse boundary elements approximation. *arXiv preprint arXiv:1908.11670*, 2019.
- [72] Azeddine Essai. Weighted FOM and GMRES for solving nonsymmetric linear systems. *Numerical Algorithms*, 18(3-4):277–292, 1998.
- [73] Sofi Esterhazy and Jens Markus Melenk. *On stability of discretizations of the Helmholtz equation*, volume 83 of *Lecture Notes in Computational Science and Engineering*, pages 285–324. Springer-Verlag, Berlin Heidelberg, 2012.
- [74] Lawrence Craig Evans and Ronald F. Gariepy. *Measure Theory and Fine Properties of Functions*. Studies in Advanced Mathematics. CRC Press, Boca Raton, FL, 1992.
- [75] Xiaobing H. Feng and Cody Lorton. An efficient Monte Carlo interior penalty discontinuous Galerkin method for elastic wave scattering in random media. *Computer Methods in Applied Mechanics and Engineering*, 315:141–168, 2017.
- [76] Xiaobing H. Feng and Dongwoo Sheen. An elliptic regularity coefficient estimate for a problem arising from a frequency domain treatment of waves. *Transactions of the American Mathematical Society*, 346(2):475–487, 1994.
- [77] Xiaobing H. Feng and Haijun Wu. Discontinuous Galerkin Methods for the Helmholtz Equation with Large Wave Number. *arXiv preprint arXiv:0810.1475*, 2008.
- [78] Xiaobing H. Feng and Haijun Wu. Discontinuous Galerkin Methods for the Helmholtz Equation with Large Wave Number. *SIAM Journal on Numerical Analysis*, 47(4):2872–2896, 2009.
- [79] Xiaobing H. Feng and Haijun Wu.  $hp$ -discontinuous Galerkin methods for the Helmholtz equation with large wave number. *Mathematics of Computation*, 80(276):1997–2024, 2011.
- [80] Xiaobing H. Feng, Junshan Lin, and Cody Lorton. An Efficient Numerical Method for Acoustic Wave Scattering in Random Media. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):790–822, 2015.
- [81] Xiaobing H. Feng, Junshan Lin, and David P. Nicholls. An Efficient Monte Carlo-Transformed Field Expansion Method for Electromagnetic Wave Scattering by Random Rough Surfaces. *Communications in Computational Physics*, 23(3):685–705, 2018.
- [82] Jeffrey Galkowski, Eike H. Müller, and Euan A. Spence. Wavenumber-explicit analysis for the Helmholtz  $h$ -BEM: error estimates and iteration counts for the Dirichlet problem. *Numerische Mathematik*, 142(2):329–357, 2019.
- [83] Jeffrey Galkowski, Euan A. Spence, and Jared Wunsch. Optimal constants in nontrapping resolvent estimates. *Pure and Applied Analysis, to appear*, 2019.

- [84] Irene Gamba, Leslie Greengard, Kevin R. Payne, Tonatiuh Sánchez-Vizuet, Christina Sormani, and Terence Tao. The Mathematics of Cathleen Synge Morawetz. *Notices of the American Mathematical Society*, 65(7):764–778, 2018.
- [85] Martin J. Gander and Hui 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.
- [86] Martin J. Gander, Ivan G. Graham, and Euan A. Spence. Applying GMRES to the Helmholtz equation with shifted Laplacian preconditioning: What is the largest shift for which wavenumber-independent convergence is guaranteed? *Numerische Mathematik*, 131(3):567–614, 2015.
- [87] Mahadevan Ganesh and Stuart C. Hawkins. A High Performance Computing and Sensitivity Analysis Algorithm for Stochastic Many-Particle Wave Scattering. *SIAM Journal on Scientific Computing*, 37(3):A1475–A1503, 2015.
- [88] Mahadevan Ganesh, Frances Y. Kuo, and Ian H. Sloan. Quasi-Monte Carlo finite element wave propagation in heterogeneous random media. *In preparation*.
- [89] Nicola Garofalo and Fang-Hua Lin. Unique Continuation for Elliptic Operators: A Geometric-Variational Approach. *Communications on Pure and Applied Mathematics*, 40(3):347–366, 1987.
- [90] Roger G. Ghanem and Robert M. Kruger. Numerical solution of spectral stochastic finite element systems. *Computer Methods in Applied Mechanics and Engineering*, 129(3):289–303, 1996.
- [91] Roger G. Ghanem and Pol D. Spanos. *Stochastic Finite Elements: A Spectral Approach, Revised Edition*. Dover, 2012.
- [92] Alec D. Gilbert, Ivan G. Graham, Frances Y. Kuo, Robert Scheichl, and Ian H. Sloan. Analysis of quasi-Monte Carlo methods for elliptic eigenvalue problems with stochastic coefficients. *Numerische Mathematik*, 142(4):863–915, 2019.
- [93] Michael B. Giles. Multilevel Monte Carlo research. URL [http://people.maths.ox.ac.uk/~gilesm/mlmc\\_community.html](http://people.maths.ox.ac.uk/~gilesm/mlmc_community.html). Last accessed 04/09/2019.
- [94] Michael B. Giles. Multilevel Monte Carlo Path Simulation. *Operations Research*, 56(3):607–617, 2008.
- [95] Michael B. Giles. Multilevel Monte Carlo methods. *Acta Numerica*, 24:259–328, 2015.
- [96] Claude Jeffrey Gittelsohn. Stochastic Galerkin discretization of the log-normal isotropic diffusion problem. *Mathematical Models and Methods in Applied Sciences*, 20(02):237–263, 2010.

- [97] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The John Hopkins University Press, Baltimore, MD, fourth edition, 2013.
- [98] Andrew D. Gordon and Catherine E. Powell. On solving stochastic collocation systems with algebraic multigrid. *IMA Journal of Numerical Analysis*, 32(3):1051–1070, 2012.
- [99] Ivan G. Graham and Stefan A. Sauter. Stability and error analysis for the Helmholtz equation with variable coefficients. *Mathematics of Computation*, 89(321):105–138, 2019.
- [100] Ivan G. Graham, Frances Y. Kuo, Dirk Nuyens, Robert Scheichl, and Ian H. Sloan. Quasi-Monte Carlo methods for elliptic PDEs with random coefficients and applications. *Journal of Computational Physics*, 230(10):3668–3694, 2011.
- [101] Ivan G. Graham, Maike Löhndorf, Jens Markus Melenk, and Euan A. Spence. When is the error in the  $h$ -BEM for solving the Helmholtz equation bounded independently of  $k$ ? *BIT Numerical Mathematics*, 55(1):171–214, 2014.
- [102] Ivan G. Graham, Euan A. Spence, and Eero Vainikko. Domain decomposition preconditioning for high-frequency Helmholtz problems with absorption. *Mathematics of Computation*, 86(307):2089–2127, 2017.
- [103] Ivan G. Graham, Matthew J. Parkinson, and Robert Scheichl. Modern Monte Carlo Variants for Uncertainty Quantification in Neutron Transport. In Josef Dick, Frances Y. Kuo, and Henryk Woźniakowski, editors, *Contemporary Computational Mathematics — A Celebration of the 80th Birthday of Ian Sloan*, pages 455–481, Cham, Switzerland, 2018. Springer International Publishing AG. Corrected publication.
- [104] Ivan G. Graham, Matthew J. Parkinson, and Robert Scheichl. Error Analysis and Uncertainty Quantification for the Heterogeneous Transport Equation in Slab Geometry. *arXiv preprint arXiv:1903.11838*, 2019.
- [105] Ivan G. Graham, Owen R. Pembery, and Euan A. Spence. The Helmholtz equation in heterogeneous media: a priori bounds, well-posedness, and resonances. *Journal of Differential Equations*, 266(6):2869–2923, 2019.
- [106] Ivan G. Graham, Euan A. Spence, and Jun Zou. Domain Decomposition with local impedance condition for the Helmholtz equation. *arXiv preprint arXiv:1806.03731*, 2019.
- [107] Pierre Grisvard. *Elliptic Problems in Nonsmooth Domains*. Pitman, Boston, 1985.
- [108] Max D. Gunzburger, Clayton G. Webster, and Guannan Zhang. Stochastic finite element methods for partial differential equations with random input data. *Acta Numerica*, 23: 521–650, 2014.
- [109] Stefan Güttel and Jennifer Pestana. Some observations on weighted GMRES. *Numerical Algorithms*, 67(4):733–752, 2014.

- [110] Paul R. Halmos. *Measure Theory*, volume 18 of *Graduate Texts in Mathematics*. Springer Science+Business Media, New York, 1974.
- [111] Stefan Heinrich. Monte Carlo Complexity of Global Solution of Integral Equations. *Journal of Complexity*, 14(2):151–175, 1998.
- [112] Stefan Heinrich. Multilevel Monte Carlo Methods. In Svetozar Margenov, Jerzy Waśniewski, and Plamen Yalamov, editors, *Large-Scale Scientific Computing*, number 2179 in Lecture Notes in Computer Science, pages 58–67, Berlin Heidelberg, 2001. Springer-Verlag.
- [113] Bruce Hendrickson and Robert Leland. A multilevel algorithm for partitioning graphs. In *Supercomputing '95: Proceedings of the 1995 ACM/IEEE Conference on Supercomputing (CDROM)*, page 28, New York, 1995. ACM Press. ISBN 0-89791-816-9.
- [114] Lukas Herrmann, Annika Lang, and Christoph Schwab. Numerical analysis of lognormal diffusions on the sphere. *Stochastics and Partial Differential Equations: Analysis and Computations*, 6(1):1–44, 2018.
- [115] Ulrich Hetmaniuk. Stability estimates for a class of Helmholtz problems. *Communications in Mathematical Sciences*, 5(3):665–678, 2007.
- [116] Ralf Hiptmair and Patrick Meury. Stabilized FEM–BEM Coupling for Helmholtz Transmission Problems. *SIAM Journal on Numerical Analysis*, 44(5):2107–2130, 2006.
- [117] Ralf Hiptmair, Laura Scarabosio, Claudia Schillings, and Christoph Schwab. Large deformation shape uncertainty quantification in acoustic scattering. *Advances in Computational Mathematics*, 44(5):1475–1518, 2018.
- [118] Frank Ihlenburg. *Finite Element Analysis of Acoustic Scattering*, volume 132 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1998.
- [119] Frank Ihlenburg and Ivo M. Babuška. Finite Element Solution of the Helmholtz Equation with High Wave Number Part I: The h-Version of the FEM. *Computers & Mathematics with Applications*, 30(9):9–37, 1995.
- [120] Frank Ihlenburg and Ivo M. Babuška. Dispersion analysis and error estimation of Galerkin finite element methods for the helmholtz equation. *International Journal for Numerical Methods in Engineering*, 38(22):3745–3774, 1995.
- [121] Frank Ihlenburg and Ivo M. Babuška. Finite element solution of the Helmholtz equation with high wave number part II: the h-p version of the FEM. *SIAM Journal on Numerical Analysis*, 34(1):315–358, 1997.
- [122] Carlos Jerez-Hanckes and Christoph Schwab. Electromagnetic wave scattering by random surfaces: uncertainty quantification via sparse tensor boundary elements. *IMA Journal of Numerical Analysis*, 37(3):1175–1210, 2016.

- [123] Carlos Jerez-Hanckes, Christoph Schwab, and Jakob Zech. Electromagnetic wave scattering by random surfaces: Shape holomorphy. *Mathematical Models and Methods in Applied Sciences*, 27(12):2229–2259, 2017.
- [124] David Jerison and Carlos E. Kenig. Unique continuation and absence of positive eigenvalues for Schrödinger operators. *Annals of Mathematics*, 121(3):463–488, 1985.
- [125] Chao Jin and Xiao-Chuan Cai. A Preconditioned Recycling GMRES Solver for Stochastic Helmholtz Problems. *Communications in Computational Physics*, 6(2):342–353, 2009.
- [126] Wenjia Jing and Olivier Pinaud. A backscattering model based on corrector theory of homogenization for the random Helmholtz equation. *Discrete and Continuous Dynamical Systems — Series B*, 24(10):5377–5407, 2019.
- [127] Andreas Keese. *Numerical Solution of Systems with Stochastic Uncertainties—A General Purpose Framework for Stochastic Finite Elements*. PhD thesis, Technischen Universität Braunschweig, 2004.
- [128] Arbaz Khan, Catherine E. Powell, and David J. Silvester. Robust Preconditioning for Stochastic Galerkin Formulations of Parameter-Dependent Nearly Incompressible Elasticity Equations. *SIAM Journal on Scientific Computing*, 41(1):A402–A421, 2019.
- [129] Boris N. Khoromskij and Christoph Schwab. Tensor-Structured Galerkin Approximation of Parametric and Stochastic Elliptic PDEs. *SIAM Journal on Scientific Computing*, 33(1):364–385, 2011.
- [130] Frances Kuo, Robert Scheichl, Christoph Schwab, Ian Sloan, and Elisabeth Ullmann. Multilevel Quasi-Monte Carlo methods for lognormal diffusion problems. *Mathematics of Computation*, 86(308):2827–2860, 2017.
- [131] Frances Y. Kuo and Dirk Nuyens. Application of Quasi-Monte Carlo Methods to Elliptic PDEs with Random Diffusion Coefficients: A Survey of Analysis and Implementation. *Foundations of Computational Mathematics*, 16(6):1631–1696, 2016.
- [132] Frances Y. Kuo and Dirk Nuyens. Application of Quasi-Monte Carlo Methods to PDEs with Random Coefficients — An Overview and Tutorial. In Art B. Owen and Peter W. Glynn, editors, *Monte Carlo and Quasi-Monte Carlo Methods*, volume 241 of *Springer Proceedings in Mathematics & Statistics*, pages 53–71, Cham, Switzerland, 2018. Springer International Publishing AG.
- [133] Frances Y. Kuo and Dirk Nuyens. Hot New Directions for Quasi-Monte Carlo Research in Step with Applications. In Art B. Owen and Peter W. Glynn, editors, *Monte Carlo and Quasi-Monte Carlo Methods*, volume 241 of *Springer Proceedings in Mathematics & Statistics*, pages 123–144, Cham, Switzerland, 2018. Springer International Publishing AG.

- [134] David Lafontaine, Euan A. Spence, and Jared Wunsch. For most frequencies, strong trapping has a weak effect in frequency-domain scattering. *arXiv preprint arXiv:1903.12172*, 2019.
- [135] David Lafontaine, Euan A. Spence, and Jared Wunsch. A sharp relative-error bound for the Helmholtz  $h$ -FEM at high frequency. *arXiv preprint arXiv:1911.11093*, 2019.
- [136] Peter D Lax and Ralph S Phillips. *Scattering Theory*, volume 26 of *Pure and Applied Mathematics*. Academic Press Inc., Boston, MA, second edition, 1989. with appendices by Cathleen S. Morawetz and Georg Schmidt.
- [137] Rolf Leis. *Initial Boundary Value Problems in Mathematical Physics*. Wiley, Chichester, UK, 1986.
- [138] Jingshi Li, Xiaoshen Wang, and Kai Zhang. An efficient alternating direction method of multipliers for optimal control problems constrained by random Helmholtz equations. *Numerical Algorithms*, 78(1):161–191, 2018.
- [139] Yonglin Li and Haijun Wu. FEM and CIP-FEM for Helmholtz Equation with High Wave Number and Perfectly Matched Layer Truncation. *SIAM Journal on Numerical Analysis*, 57(1):96–129, 2019.
- [140] Jacques-Louis Lions and Enrico Magenes. *Non-Homogeneous Boundary Value Problems and Applications, Volume I*, volume 181 of *Grundlehren der mathematischen Wissenschaften*. Springer-Verlag, Berlin Heidelberg, 1972.
- [141] Anna Lischke, Guofei Pang, Mamikon Gulian, Fangying Song, Christian Glusa, Xiaoning Zheng, Zhiping Mao, Wei Cai, Mark M. Meerschaert, Mark Ainsworth, and George Em Karniadakis. What Is the Fractional Laplacian? A Comparative Review with New Results. *arXiv preprint arXiv:1801.09767*, 2018.
- [142] Michel Loève. *Probability Theory I*, volume 45 of *Graduate Texts in Mathematics*. Springer-Verlag, New York Heidelberg Berlin, 4th edition, 1977.
- [143] Gabriel J. Lord, Catherine E. Powell, and Tony Shardlow. *An Introduction to Computational Stochastic PDEs*. Cambridge Texts in Applied Mathematics. Cambridge University Press, New York, 2014.
- [144] Fabio Luporini, Ana Lucia Varbanescu, Florian Rathgeber, Gheorghe-Teodor Bercea, J. Ramanujam, David A. Ham, and Paul H. J. Kelly. Cross-Loop Optimization of Arithmetic Intensity for Finite Element Local Assembly. *ACM Transactions on Architecture and Code Optimization*, 11(4):57:1–57:25, 2015.
- [145] Charalambos Makridakis and Ricardo H. Nochetto. Elliptic Reconstruction and a Posteriori Error Estimates for Parabolic Problems. *SIAM Journal on Numerical Analysis*, 41(4):1584–1594, 2003.

- [146] William C. H. McLean. *Strongly Elliptic Systems and Boundary Integral Equations*. Cambridge University Press, 2000.
- [147] Jens Markus Melenk. *On Generalized Finite Element Methods*. PhD thesis, The University of Maryland, 1995.
- [148] Jens Markus Melenk and Stefan Sauter. Convergence analysis for finite element discretizations of the Helmholtz equation with Dirichlet-to-Neumann boundary conditions. *Mathematics of Computation*, 79(272):1871–1914, 2010.
- [149] Jens Markus Melenk and Stefan Sauter. Wavenumber Explicit Convergence Analysis for Galerkin Discretizations of the Helmholtz Equation. *SIAM Journal on Numerical Analysis*, 49(3):1210–1243, 2011.
- [150] Richard B. Melrose and Johannes Sjöstrand. Singularities of boundary value problems. II. *Communications on Pure and Applied Mathematics*, 35(2):129–168, 1982.
- [151] Siddhartha Mishra, Christoph Schwab, and Jonas Šukys. Multi-level Monte Carlo finite volume methods for uncertainty quantification of acoustic wave propagation in random heterogeneous layered medium. *Journal of Computational Physics*, 312:192–217, 2016.
- [152] Andrea Moiola and Euan A. Spence. Acoustic transmission problems: Wavenumber-explicit bounds and resonance-free regions. *Mathematical Models and Methods in Applied Sciences*, 29(2):317–354, 2019.
- [153] Cathleen S. Morawetz. The Decay of Solutions of the Exterior Initial-Boundary Value Problem for the Wave Equation. *Communications on Pure and Applied Mathematics*, 14(3): 561–568, 1961.
- [154] Cathleen S. Morawetz. Decay for Solutions of the Exterior Problem for the Wave Equation. *Communications on Pure and Applied Mathematics*, 28(2):229–264, 1975.
- [155] Cathleen S. Morawetz and Donald Ludwig. An Inequality for the Reduced Wave Operator and the Justification of Geometrical Optics. *Communications on Pure and Applied Mathematics*, 21(2):187–203, 1968.
- [156] Cathleen S. Morawetz, James V. Ralston, and Walter A. Strauss. Decay of Solutions of the Wave Equation Outside Nontrapping Obstacles. *Communications on Pure and Applied Mathematics*, 30(4):447–508, 1977.
- [157] Antje Mugler and Hans-Jörg Starkloff. On elliptic partial differential equations with random coefficients. *Studia Universitatis Babeş-Bolyai Mathematica*, 56(2):473–487, 2011.
- [158] Jean Claude Nédélec. *Acoustic and Electromagnetic Equations: Integral Representations for Harmonic Problems*, volume 144 of *Applied Mathematical Sciences*. Springer Science+Business Media, New York, 2001.

- [159] Joachim A. Nitsche. Ein Kriterium für die Quasi-Optimalität des Ritzschen Verfahrens. *Numerische Mathematik*, 11(4):346–348, 1986.
- [160] Fabio Nobile, Raul Tempone, and Clayton G. Webster. A Sparse Grid Stochastic Collocation Method for Partial Differential Equations with Random Input Data. *SIAM Journal on Numerical Analysis*, 46(5):2309–2345, 2008.
- [161] Dirk Nuyens. The ‘Magic Point Shop’ of QMC point generators and generating vectors, . URL <https://people.cs.kuleuven.be/~dirk.nuyens/qmc-generators/>. Last accessed 7/8/19.
- [162] Dirk Nuyens. QMC point generators in C++ README, . URL <https://bitbucket.org/dnuyens/qmc-generators/src/cb0f2fb10fa9c9f2665e41419097781b611daa1e/cpp/README.md>. Last accessed 23/09/2019.
- [163] Mario Ohlberger and Barbara Verfürth. A new Heterogeneous Multiscale Method for the Helmholtz Equation with High Contrast. *Multiscale Modeling & Simulation*, 16(1):385–411, 2018.
- [164] Bernt Øksendal. *Stochastic Differential Equations*. Universitext. Springer-Verlag, Berlin Heidelberg, sixth corrected printing, sixth edition, 2013.
- [165] Matthew J. Parkinson. *Uncertainty Quantification in Radiative Transport*. PhD thesis, University of Bath, 2018.
- [166] Michael L. Parks, Eric de Sturler, Greg Mackey, Duane D. Johnson, and Spandan Maiti. Recycling Krylov Subspaces for Sequences of Linear Systems. *SIAM Journal on Scientific Computing*, 28(5):1651–1674, 2006.
- [167] Manuel F. Pellissetti and Roger G. Ghanem. Iterative solution of systems of linear equations arising in the context of stochastic finite elements. *Advances in Engineering Software*, 31(8–9):607–616, 2000.
- [168] Owen Pembedy. orpembedy/thesis, February 2020. URL <https://doi.org/10.5281/zenodo.3666163>.
- [169] Owen R. Pembedy. orpembedy/helmholtz-firedrake, February 2020. URL <https://doi.org/10.5281/zenodo.3665834>.
- [170] Owen R. Pembedy. orpembedy/helmholtz-monte-carlo, February 2020. URL <https://doi.org/10.5281/zenodo.3665836>.
- [171] Owen R. Pembedy. orpembedy/helmholtz-nearby-preconditioning, February 2020. URL <https://doi.org/10.5281/zenodo.3665838>.
- [172] Owen R. Pembedy. orpembedy/prob-gmres-examples, February 2020. URL <https://doi.org/10.5281/zenodo.3665841>.

- [173] Owen R. Pembery. `orpembery/running-helmholtz-monte-carlo`, February 2020. URL <https://doi.org/10.5281/zenodo.3665844>.
- [174] Owen R. Pembery. Data for ‘The Helmholtz Equation in Heterogeneous and Random Media: Analysis and Numerics’, February 2020. URL <https://doi.org/10.5281/zenodo.3665798>.
- [175] Owen R. Pembery. `orpembery/running-nbpc`, February 2020. URL <https://doi.org/10.5281/zenodo.3665846>.
- [176] Owen R. Pembery and Euan A. Spence. The Helmholtz equation in random media: well-posedness and a priori bounds. *SIAM/ASA Journal on Uncertainty Quantification*, 8(1):58–87, 2020.
- [177] Benoit Perthame and Luis Vega. Morrey–Campanato Estimates for Helmholtz Equations. *Journal of Functional Analysis*, 164(2):340–355, 1999.
- [178] Georgi Popov and Georgi Vodev. Resonances near the real axis for transparent obstacles. *Communications in Mathematical Physics*, 207(2):411–438, 1999.
- [179] Georgi Popov and Georgi Vodev. Distribution of the resonances and local energy decay in the transmission problem. *Asymptotic Analysis*, 19(3–4):253–265, 1999.
- [180] Catherine E. Powell and Howard C. Elman. Block-diagonal preconditioning for spectral stochastic finite-element systems. *IMA Journal of Numerical Analysis*, 29(2):350–375, 2009.
- [181] Catherine E. Powell and David J. Silvester. Preconditioning Steady-State Navier–Stokes Equations with Random Data. *SIAM Journal on Scientific Computing*, 34(5):A2482–A2506, 2012.
- [182] Catherine E. Powell and Elisabeth Ullmann. Preconditioning Stochastic Galerkin Saddle Point Systems. *SIAM Journal on Matrix Analysis and Applications*, 31(5):2813–2840, 2010.
- [183] Florian Rathgeber, David A. Ham, Lawrence Mitchell, Michael Lange, Fabio Luporini, Andrew T. T. McRae, Gheorghe-Teodor Bercea, Graham R. Markall, and Paul H. J. Kelly. Firedrake: Automating the Finite Element Method by Composing Abstractions. *ACM Transactions on Mathematical Software*, 43(3):24:1–24:27, 2016. ISSN 0098-3500.
- [184] Kenneth F. Riley, Michael P. Hobson, and Stephen J. Bence. *Mathematical methods for physics and engineering*. Cambridge University Press, Cambridge, U.K., 1997.
- [185] Eveline Rosseel and Stefan Vandewalle. Iterative Solvers for the Stochastic Finite Element Method. *SIAM Journal on Scientific Computing*, 32(1):372–397, 2010.
- [186] Raymond A. Ryan. *Introduction to Tensor Products of Banach Spaces*. Springer Monogr. Math. Springer-Verlag, London, 2002.

- [187] Youcef Saad and Martin H. Schultz. GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems. *SIAM Journal on Scientific and Statistical Computing*, 7(3):856–869, 1986.
- [188] Stefan A. Sauter. A refined finite element convergence theory for highly indefinite Helmholtz problems. *Computing*, 78(2):101–115, 2006.
- [189] Stefan A. Sauter and Christoph Schwab. *Boundary Element Methods*, volume 39 of *Springer Ser. Comput. Math.* Springer-Verlag, Berlin Heidelberg, 2011.
- [190] Stefan A. Sauter and Céline Torres. Stability estimate for the Helmholtz equation with rapidly jumping coefficients. *Zeitschrift für angewandte Mathematik und Physik*, 69(6):139, 2018.
- [191] Laura Scarabosio. Multilevel Monte Carlo on a high-dimensional parameter space for transmission problems with geometric uncertainties. *International Journal for Uncertainty Quantification*, 9(6):515–541, 2019.
- [192] Alfred H. Schatz. An Observation Concerning Ritz–Galerkin Methods with Indefinite Bilinear Forms. *Mathematics of Computation*, 28(128):959–962, 1974.
- [193] Schlumberger. Seismic Wave. In *Oilfield Glossary*. URL [http://www.glossary.oilfield.slb.com/Terms/s/seismic\\_wave.aspx](http://www.glossary.oilfield.slb.com/Terms/s/seismic_wave.aspx). last accessed 18/09/2019.
- [194] Christoph Schwab.  *$p$ - and  $hp$ - Finite Element Methods*. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 1998.
- [195] Claude E. Shannon. Communication in the Presence of Noise. *Proceedings of the IRE*, 37(1):10–21, 1949.
- [196] Jacob Shapiro. Local energy decay for Lipschitz wavespeeds. *Communications in Partial Differential Equations*, 43(5):839–858, 2018.
- [197] Jacques Simon. Compact Sets in the Space  $L^p(0, T; B)$ . *Annali di Matematica Pura ed Applicata*, 146(1):65–96, 1986.
- [198] Bedřich Sousedík and Howard C. Elman. Stochastic Galerkin methods for the steady-state Navier–Stokes equations. *Journal of Computational Physics*, 316:435–452, 2016.
- [199] Euan A. Spence. Wavenumber-Explicit Bounds in Time-Harmonic Acoustic Scattering. *SIAM Journal on Mathematical Analysis*, 46(4):2987–3024, 2014.
- [200] Euan A. Spence. Overview of Variational Formulations for Linear Elliptic PDEs. In Athanassios S. Fokas and Beatrice Pelloni, editors, *Unified Transform Method for Boundary Value Problems: Applications and Advances*, pages 93–159. SIAM, 2015.
- [201] Jonas Šukys. *Robust multi-level Monte Carlo Finite Volume methods for systems of hyperbolic conservation laws with random input data*. PhD thesis, ETH Zürich, 2014.

- [202] Jonas Šukys, Siddhartha Mishra, and Christoph Schwab. Multi-level Monte Carlo Finite Difference and Finite Volume Methods for Stochastic Linear Hyperbolic Systems. In Josef Dick, Frances Y. Kuo, Gareth W. Peters, and Ian H. Sloan, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2012*, number 65 in Springer Proceedings in Mathematics & Statistics, pages 649–666, Berlin Heidelberg, 2013. Springer-Verlag.
- [203] Matthias Taus, Leonardo Zepeda-Núñez, Russel J. Hewett, and Laurent Demanet. L-Sweeps: A scalable parallel preconditioner for the high-frequency Helmholtz equation. In Manfred Kaltenbacher, Jens Markus Melenk, Lothar Nannen, and Florian Toth, editors, *14th International Conference on Mathematical and Numerical Aspects of Wave Propagation: Book of Abstracts*, pages 250–251, Vienna, Austria, 2019. Institute of Mechanics and Mechatronics, Faculty of Mechanical and Industrial Engineering and Institute of Analysis and Scientific Computing, Faculty of Mathematics and Geoinformation, TU Wien. doi: 10.34726/waves2019. URL <https://repositum.tuwien.ac.at/obvutwoa/download/pdf/4117715>.
- [204] Aretha L. Teckentrup, Peter Jantsch, Clayton G. Webster, and Max Gunzburger. A Multilevel Stochastic Collocation Method for Partial Differential Equations with Random Input Data. *SIAM/ASA Journal on Uncertainty Quantification*, 3(1):1046–1074, 2015.
- [205] Martin Thomas. *Analysis of Rough Surface Scattering Problems*. PhD thesis, University of Reading, 2006.
- [206] Vidar Thomée. Negative Norm Estimates and Superconvergence in Galerkin Methods for Parabolic Problems. *Mathematics of Computation*, 34(149):93–113, 1980.
- [207] Ramakrishna Tipireddy, Eric T. Phipps, and Roger G. Ghanem. A Comparison of Solution Methods for Stochastic Partial Differential Equations. In Eric C. Cyr and S. Scott Collis, editors, *CSRI Summer Proceedings 2010*, pages 79–90. Sandia National Laboratories, 2010.
- [208] Andrea Toselli and Olof B. Widlund. *Domain Decomposition Methods — Algorithms and Theory*, volume 34 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin Heidelberg, 2005.
- [209] Paul Tsuji, Dongbin Xiu, and Lexing Ying. Fast method for high-frequency acoustic scattering from random scatterers. *International Journal for Uncertainty Quantification*, 1(2):99–117, 2011.
- [210] Elisabeth Ullmann, Howard C. Elman, and Oliver G. Ernst. Efficient Iterative Solvers for Stochastic Galerkin Discretizations of Log-Transformed Random Diffusion Problems. *SIAM Journal on Scientific Computing*, 34(2):A659–A682, 2012.
- [211] Boris R. Vainberg. On the short wave asymptotic behaviour of solutions of stationary problems and the asymptotic behaviour as  $t \rightarrow \infty$  of solutions of non-stationary problems. *Russian Mathematical Surveys*, 30(2):1–58, 1975.

- [212] Georgi Vodev. On the uniform decay of the local energy. *Serdica Mathematical Journal*, 25(3):191–206, 1999.
- [213] Guanjie Wang and Qifeng Liao. Efficient Spectral Stochastic Finite Element Methods for Helmholtz Equations with Random Inputs. *East Asian Journal on Applied Mathematics*, 9(3):601–621, 2019.
- [214] Mary Fanett Wheeler. A Priori  $L_2$  Error Estimates for Galerkin Approximations to Parabolic Partial Differential Equations. *SIAM Journal on Numerical Analysis*, 10(4):723–759, 1973.
- [215] Stephen Willard. *General Topology*. Addison–Wesley, Reading, Massachusetts, 1970.
- [216] Haijun Wu. Pre-asymptotic error analysis of CIP-FEM and FEM for the Helmholtz equation with high wave number. part I: linear version. *IMA Journal of Numerical Analysis*, 34(3):1266–1288, 2014.
- [217] Haijun Wu and Jun Zou. Finite Element Method and its Analysis for a Nonlinear Helmholtz Equation with High Wave Numbers. *SIAM Journal on Numerical Analysis*, 56(3):1338–1359, 2018.
- [218] Dongbin Xiu and Jan S. Hesthaven. High-Order Collocation Methods for Differential Equations with Random Inputs. *SIAM Journal on Scientific Computing*, 27(3):1118–1139, 2005.
- [219] Dongbin Xiu and George Em Karniadakis. The Wiener–Askey Polynomial Chaos for Stochastic Differential Equations. *SIAM Journal on Scientific Computing*, 24(2):619–644, 2002.
- [220] Dongbin Xiu and Jie Shen. An Efficient Spectral Method for Acoustic Scattering from Rough Surfaces. *Communications in Computational Physics*, 2(1):54–72, 2007.
- [221] Leonardo Zepeda-Núñez, Adrian Scheuer, Russel J. Hewett, and Laurent Demanet. The method of polarized traces for the 3D Helmholtz equation. *Geophysics*, 84(4):T133–T333, 2019.
- [222] Lingxue Zhu and Haijun Wu. Preasymptotic error analysis of CIP-FEM and FEM for Helmholtz equation with high wave number. part II:  $hp$  version. *SIAM Journal on Numerical Analysis*, 51(3):1828–1852, 2013.