# A REDUCED BASIS METHOD FOR MATRIX EQUATIONS ASSOCIATED WITH STOCHASTIC GALERKIN APPROXIMATION OF PARAMETER-DEPENDENT SADDLE POINT PROBLEMS 

A thesis submitted to The University of Manchester for the degree of Doctor of Philosophy in the Faculty of Science and Engineering

Ying Liu<br>Department of Mathematics

## Contents

List of Tables ..... 4
List of Figures ..... 12
Abstract ..... 15
Declaration ..... 16
Copyright Statement ..... 17
Acknowledgements ..... 18
1 Introduction ..... 19
1.1 Deterministic saddle point problems ..... 20
1.2 Parameter-dependent saddle point problems ..... 23
1.3 Methods for solving matrix equations ..... 28
1.4 Outline ..... 33
2 Parameter-dependent Linear Elasticity Problem ..... 36
2.1 Random fields and the Karhunen-Loève expansion ..... 37
2.2 Stochastic Galerkin approximation ..... 46
2.2.1 Weak formulation ..... 46
2.2.2 Finite-dimensional problem ..... 48
2.3 Preconditioned MINRES ..... 57
2.3.1 Preconditioning ..... 59
2.3.2 Numerical experiments ..... 65
2.3.3 Conclusions ..... 73
3 A Reduced Basis Method ..... 74
3.1 Low rank analysis ..... 75
3.2 Multi-term reduced basis method ..... 80
3.2.1 Pre-processing phase ..... 81
3.2.2 Iterative solution phase ..... 89
3.2.3 Stopping criteria ..... 94
3.2.4 Multi-RB algorithm ..... 95
4 Multi-RB with Symmetric Preconditioning ..... 98
4.1 Eigenvalues and choice of shifts ..... 98
4.2 Numerical experiments ..... 102
4.3 Conclusions ..... 116
5 Multi-RB with Left Preconditioning ..... 118
5.1 Eigenvalues and choice of shifts ..... 118
5.2 Numerical experiments ..... 120
5.3 Conclusions ..... 129
6 Groundwater flow problem ..... 131
6.1 Symmetric preconditioning ..... 136
6.1.1 Pre-processing phase ..... 136
6.1.2 Iterative solution phase ..... 138
6.2 Left preconditioning ..... 147
6.2.1 Pre-processing phase ..... 147
6.2.2 Iterative solution phase ..... 148
6.3 Conclusions ..... 154
7 Conclusions ..... 157
Bibliography ..... 159

## List of Tables

### 2.1 The smallest number of terms $M$ needed to ensure $\varepsilon_{M} \leqslant \epsilon$ or $\beta_{M} \geqslant \vartheta$. <br> 42

2.2 The proportion $\beta_{M}$ and the relative error $\varepsilon_{M}$ for $M=1, \ldots, 5$ ..... 44
2.3 Numerical values of the minimum and maximum of the Young modulus. ..... 48
2.4 Iteration counts and timings (in brackets) using preconditioned MIN- RES for Test problem 1 when $\sigma=0.17, \nu=0.4999, M=7$ and $d=4$. ..... 68
2.5 Iteration counts and timings (in brackets) using preconditioned MIN- RES for the slow decay case in Test problem 2 when $\nu=0.4999$, $M=5$ and $d=4$. ..... 68
2.6 Iteration counts and timings (in brackets) using preconditioned MIN- RES for the fast decay case in Test problem 2 when $\nu=0.4999$, $M=5$ and $d=4$. ..... 68
2.7 Iteration counts and timings (in brackets) using preconditioned MIN- RES for Test problem 1 with varying $\nu$ when $\sigma=0.17, l=6, M=7$ and $d=4$. ..... 69
2.8 Iteration counts and timings (in brackets) using preconditioned MIN- RES for the slow decay case in Test problem 2 with varying $\nu$ when $l=6, M=5$ and $d=4$. ..... 69
2.9 Iteration counts and timings (in brackets) using preconditioned MIN- RES for the fast decay case in Test problem 2 with varying $\nu$ when $l=6, M=5$ and $d=4$. ..... 69
2.10 Iteration counts and timings (in brackets) using preconditioned MIN- RES for Test problem 1 with varying $d$ when $\sigma=0.17, l=6$, $\nu=0.4999$ and $M=7$. ..... 70
2.11 Iteration counts and timings (in brackets) using preconditioned MINRES for the slow decay case in Test problem 2 with varying $d$ when $l=6, \nu=0.4999$ and $M=5$.
2.12 Iteration counts and timings (in brackets) using preconditioned MINRES for the fast decay case in Test problem 2 with varying $d$ when $l=6, \nu=0.4999$ and $M=5$.
2.13 Iteration counts and timings (in brackets) using preconditioned MIN-
RES for Test problem 1 with varying $M$ when $\sigma=0.17, l=6$,
$\nu=0.4999$ and $d=4 . \ldots . . . . . . . . . . . . . . . . . . . . . .$.
2.14 Iteration counts and timings (in brackets) using preconditioned MIN-
RES for the slow decay case in Test problem 2 with varying $M$ when
$l=6, \nu=0.4999$ and $d=4 . . . . . . . . . . . . . . . . . . . . . .$.
71
2.15 Iteration counts and timings (in brackets) using preconditioned MINRES for the fast decay case in Test problem 2 with varying $M$ when $l=6, \nu=0.4999$ and $d=4$.

2.16 Iteration counts and timings (in brackets) in seconds using precondi
tioned MINRES for Test problem 1 with varying $\sigma$ when $l=6$,
$\nu=0.4999, M=7$ and $d=4$. ..... 72
2.17 Sizes of some large problems. ..... 72
2.18 Iteration counts and timings (in brackets) using preconditioned MIN- RES for Test problem 1 when $\sigma=0.085, \nu=0.4$; the stopping tolerance is $10^{-6}$. ..... 73

3.1 The smallest $p$ needed to achieve $\varepsilon_{p}<10^{-6}$ for Test problem 1 with
$n_{x}=2562$. The actual rank of $X$ is given in brackets. ..... 77
3.2 The smallest $p$ needed to achieve $\varepsilon_{p}<10^{-6}$ for Test problem 1 with $n_{x}=10242$. The actual rank of $X$ is given in brackets.77

3.3 The smallest $p$ needed to achieve $\varepsilon_{p}<10^{-6}$ for Test problem 2 with
$n_{x}=2562$. The actual rank of $X$ is given in brackets. ..... 78
3.4 The smallest $p$ needed to achieve $\varepsilon_{p}<10^{-6}$ for Test problem 2 with $n_{x}=10242$. The actual rank of $X$ is given in brackets.
4.1 Computed theoretical eigenvalue bounds for $\widehat{K}_{r}$, for $r=1,2, \ldots, 7$, in
Test problem 1. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 99
4.2 Computed theoretical eigenvalue bounds for $\hat{K}_{r}$, for $r=1,2, \ldots, 5$, in Test problem 2.
4.3 Computed theoretical eigenvalue bounds for $P_{r}=\widehat{K}_{r-1}+\alpha_{r-1} I$, for $r=2, \ldots, 7,8$, in Test problem 1.
4.4 Computed theoretical eigenvalue bounds for $P_{r}=\widehat{K}_{r-1}+\alpha_{r-1} I$, for $r=2, \ldots, 5,6$, in Test problem 2. 102
4.5 Numerical results for Test problem 1 with fixed grid level $6, \nu=$ 0.4999 and $\sigma=0.085$ using the symmetric preconditioning strategy; Multi-RB solver vs preconditioned MINRES. 104
4.6 Numerical results for Test problem 1 with fixed grid level $6, \nu=$ 0.4999 and $\sigma=0.17$ using the symmetric preconditioning strategy; Multi-RB solver vs standard MINRES.
4.7 Numerical results for the slow decay case in Test problem 2 with fixed grid level 6 and $\nu=0.4999$ using the symmetric preconditioning strategy; Multi-RB solver vs standard MINRES.
4.8 Numerical results for the fast decay case in Test problem 2 with fixed grid level 6 and $\nu=0.4999$ using the symmetric preconditioning strategy; Multi-RB solver vs standard MINRES. Bold numbers indicate unexpectedly large iteration counts.
4.9 Numerical results for Test problem 1 with varying grid level $l$ when $\nu=0.4999, \sigma=0.17, M=7$ and $d=4$ using the symmetric preconditioning strategy.
4.10 Numerical results for the slow decay case of Test problem 2 with varying grid level $l$ when $\nu=0.4999, M=7$ and $d=4$ using the symmetric preconditioning strategy.
4.11 Numerical results for Test problem 1 with varying Poisson ratio $\nu$ when $l=6, \sigma=0.17, M=7$ and $d=4$ using the symmetric preconditioning strategy.
4.12 Numerical results for the slow decay case of Test problem 2 with varying Poisson ratio $\nu$ when $l=6, M=7$ and $d=4$ using the symmetric preconditioning strategy.
4.13 Numerical results for Test problem 1 with varying $\vartheta$ when $l=6, \sigma=$ $0.17, \nu=0.4999 M=7$ and $d=4$ using the symmetric preconditioning strategy.
4.14 Numerical results for the slow decay case of Test problem 2 with varying $\vartheta$ when $l=6, \nu=0.4999, M=7$ and $d=4$ using the symmetric preconditioning strategy.
4.15 Numerical results for the fast decay case of Test problem 2 with varying $\vartheta$ when $l=6, \nu=0.4999, M=7$ and $d=4$ using the symmetric preconditioning strategy.
4.16 Numerical results for Test problem 1 with varying $M$ when $l=6$, $\sigma=0.17, \nu=0.4999$ and $d=4$ using the multi-parameter strategy and the symmetric preconditioning strategy.
4.17 Numerical results for the slow decay case of Test problem 2 with varying $M$ when $l=6, \nu=0.4999$ and $d=4$ using the multi-parameter strategy and the symmetric preconditioning strategy.
4.18 Numerical results for the fast decay case of Test problem 2 with varying $M$ when $l=6, \nu=0.4999$ and $d=4$ using the multi-parameter strategy and the symmetric preconditioning strategy. Bold numbers indicate unexpectedly large iteration counts.
4.19 Numerical results for Test problem 1 with varying $M$ when $l=6$, $\nu=0.4999, \sigma=0.17$ and $d=4$ when we apply the Petrov-Galerkin condition.
4.20 Numerical results for the slow decay case of Test problem 2 with varying $M$ when $l=6, \nu=0.4999$ and $d=4$ when we apply the Petrov-Galerkin condition.
4.21 Numerical results for the fast decay case of Test problem 2 with varying $M$ when $l=6, \nu=0.4999$ and $d=4$ when we apply the Petrov-Galerkin condition.
4.22 Numerical results for problems $P 1, P 2$ and $P 3$ when we apply the Petrov-Galerkin condition.114
5.1 Numerical eigenvalue bounds for $\widetilde{K}_{r}$, for $r=1, \ldots, 7$, in Test problem1.119
5.2 Numerical eigenvalue bounds for $\widetilde{K}_{r}$, for $r=1, \ldots, 5$, in Test problem 2. ..... 1195.3 Numerical results for Test problem 1 with fixed grid level $6, \nu=$0.4999 and $\sigma=0.085$ using Multi-RB with the left preconditioningstrategy.122
5.4 Numerical results for Test problem 1 with fixed grid level $6, \nu=$ 0.4999 and $\sigma=0.17$ using Multi-RB with the left preconditioning strategy. 122
5.5 Numerical results for the slow decay case in Test problem 2 with fixed grid level 6 and $\nu=0.4999$ using Multi-RB with the left preconditioning strategy. ..... 123
5.6 Numerical results for the fast decay case in Test problem 2 with fixed grid level 6 and $\nu=0.4999$ using Multi-RB with the left preconditioning strategy. ..... 123
5.7 Numerical results for Test problem 1 with varying grid level $l$ when $\nu=0.4999, \sigma=0.17, M=7$ and $d=4$ using the left preconditioning strategy. ..... 124
5.8 Numerical results for the slow decay case in Test problem 2 with varying grid level $l$ when $\nu=0.4999, M=7$ and $d=4$ using the left preconditioning strategy. ..... 124
5.9 Numerical results for Test problem 1 with varying Poisson ratio $\nu$ when $l=6, \sigma=0.17, M=7$ and $d=4$ using the left preconditioning strategy. ..... 125
5.10 Numerical results for the slow decay case in Test problem 2 with varying Poisson ratio $\nu$ when $l=6, M=7$ and $d=4$ using the left preconditioning strategy. ..... 125
5.11 Numerical results for the slow decay case of Test problem 2 with varying $\vartheta$ when $l=6, \nu=0.4999, M=7$ and $d=4$ using the left preconditioning strategy. ..... 125
5.12 Numerical results for the fast decay case of Test problem 2 with varying $\vartheta$ when $l=6, \nu=0.4999, M=7$ and $d=4$ using the left preconditioning strategy. ..... 126
5.13 Numerical results for Test problem 1 with varying $M$ when $l=6$, $\sigma=0.085, \nu=0.4999$ and $d=4$ using the multi-parameter strategy and the left preconditioning strategy. ..... 126
5.14 Numerical results for Test problem 1 with varying $M$ when $l=6$, $\sigma=0.17, \nu=0.4999$ and $d=4$ using the multi-parameter strategy and the left preconditioning strategy. ..... 127
5.15 Numerical results for the slow decay case of Test problem 2 with varying $M$ when $l=6, \nu=0.4999$ and $d=4$ using the multi-parameter strategy and the left preconditioning strategy. ..... 127
5.16 Numerical results for the fast decay case of Test problem 2 with varying $M$ when $l=6, \nu=0.4999$ and $d=4$ using the multi-parameter strategy and the left preconditioning strategy. ..... 127
5.17 Numerical results obtained using Multi-RB with the left preconditioning strategy for test problems with $l=9, M=15, d=4$ and $\nu=0.4999$. ..... 129
6.1 Numerical eigenvalue bounds for $\widehat{K}_{r}$ in Test problems 3-4. ..... 137
6.2 Numerical results for case 1 in Test problem 3 with fixed grid level 6 (giving $n_{x}=12288$ ) using Multi-RB with the symmetric precondition- ing strategy. ..... 139
6.3 Numerical results for case 2 in Test problem 3 with fixed grid level 6 (giving $n_{x}=12416$ ) using Multi-RB with the symmetric precondition- ing strategy. ..... 140
6.4 Numerical results for case 1 in Test problem 4 with fixed grid level 6 (giving $n_{x}=12288$ ) using Multi-RB with the symmetric precondition- ing strategy. ..... 140
6.5 Numerical results for case 2 in Test problem 4 with fixed grid level 6(giving $n_{x}=12416$ ) using Multi-RB with the symmetric precondition-ing strategy.1416.6 Numerical results for case 1 in Test problem 3 with varying grid level$l$ when $M=7$ and $d=4$ using the symmetric preconditioning strategy.142
6.7 Numerical results for case 2 in Test problem 3 with varying grid level $l$ when $M=7$ and $d=4$ using the symmetric preconditioning strategy. ..... 142
6.8 Numerical results for case 1 in Test problem 4 with varying grid level $l$ when $M=7$ and $d=4$ using the symmetric preconditioning strategy. ..... 142
6.9 Numerical results for case 2 in Test problem 4 with varying grid level $l$ when $M=7$ and $d=4$ using the symmetric preconditioning strategy. ..... 142
6.10 Numerical results for case 1 in Test problem 3 with varying $M$ when $l=6$ and $d=4$ when we apply the Petrov-Galerkin condition. ..... 146
6.11 Numerical results for case 2 in Test problem 3 with varying $M$ when $l=6$ and $d=4$ when we apply the Petrov-Galerkin condition. ..... 146
6.12 Numerical results for case 1 in Test problem 4 with varying $M$ when $l=6$ and $d=4$ when we apply the Petrov-Galerkin condition. ..... 146
6.13 Numerical results for case 2 in Test problem 4 with varying $M$ when $l=6$ and $d=4$ when we apply the Petrov-Galerkin condition. ..... 147
6.14 Numerical eigenvalue bounds for $\widetilde{K}_{r}$ in Test problems 3-4. ..... 148
6.15 Numerical results for case 1 in Test problem 3 with fixed grid level 6 (giving $n_{x}=12288$ ) using Multi-RB with the left preconditioning strategy. ..... 150
6.16 Numerical results for case 2 in Test problem 3 with fixed grid level 6 (giving $n_{x}=12416$ ) using Multi-RB with the left preconditioning strategy ..... 150
6.17 Numerical results for case 1 in Test problem 4 with fixed grid level 6 (giving $n_{x}=12288$ ) using Multi-RB with the left preconditioning strategy. ..... 151
6.18 Numerical results for case 2 in Test problem 4 with fixed grid level 6 (giving $n_{x}=12416$ ) using Multi-RB with the left preconditioning strategy. ..... 151
6.19 Numerical results for case 1 in Test problem 3 with varying grid level $l$ when $M=7$ and $d=4$ using the left preconditioning strategy. ..... 152
6.20 Numerical results for case 2 in Test problem 3 with varying grid level $l$ when $M=7$ and $d=4$ using the left preconditioning strategy. ..... 152
6.21 Numerical results for case 1 in Test problem 4 with varying grid level $l$ when $M=7$ and $d=4$ using the left preconditioning strategy. . . . . 152
6.22 Numerical results for case 2 in Test problem 4 with varying grid level $l$ when $M=7$ and $d=4$ using the left preconditioning strategy. . . . . 152
6.23 Numerical results obtained using Multi-RB with the left preconditioning strategy for Test problems 3-4 with $l=9, M=15$ and $d=4 \ldots \ldots 154$

## List of Figures

2.1 One realisation of $u_{M}(\boldsymbol{x}, \omega)$ with the covariance (2.15) when $\sigma=0.085$. ..... 42
2.2 Four realisations of $u_{M}(\boldsymbol{x}, \omega)$ with the covariance (2.15) when $\sigma=0.085$ and $M=13$. ..... 43
2.3 Variance of $u_{M}(\boldsymbol{x}, \omega)$ when $M=3$. ..... 44
2.4 One realisation of $u_{M}(\boldsymbol{x}, \omega)$ for the slow decay case. ..... 45
2.5 One realisation of $u_{M}(\boldsymbol{x}, \omega)$ for the fast decay case. ..... 45
2.6 Four realisations of $u_{M}(\boldsymbol{x}, \omega)$ for the slow decay case when $M=5$ ..... 45
2.7 Four realisations of $u_{M}(\boldsymbol{x}, \omega)$ for the fast decay case when $M=5$. ..... 45
2.8 Expectations and variances of components of the displacement $\boldsymbol{u}$ and Herrmann pressure $p$ when $l=6, \nu=0.4999, M=13, d=4$ and $\sigma=0.17$ for Test problem 1 ..... 66
2.9 Expectations and variances of components of the displacement $\boldsymbol{u}$ and Herrmann pressure $p$ when $l=6, \nu=0.4999, M=13$ and $d=4$ for the fast decay case in Test problem 2. ..... 66
3.1 The normalised singular values of $X$ (left) and relative error $\varepsilon_{p}$ (right) of the best rank $p$ approximation for Test problem 1. ..... 79
3.2 The normalised singular values of $X$ (left) and relative error $\varepsilon_{p}$ (right) of the best rank $p$ approximation for Test problem 2. ..... 79
4.1 Numerical eigenvalues of $\widehat{K}_{r}$ for $r=1,2, \ldots, 7$ in Test problem 1, computed on a spatial mesh with grid level 5 . ..... 100
4.2 Numerical eigenvalues of $\widehat{K}_{r}$ for $r=1,2, \ldots, 5$ in Test problem 2, computed on a spatial mesh with grid level 5 . ..... 100
4.3 Inner iteration counts for the fast decay case of Test problem 2 when $l=6, \nu=0.4999$ and $d=4$ using multi-parameter strategy. ..... 110
4.4 Convergence history for Test problem 1 when we apply the Galerkin condition with $\nu=0.4999, l=5, M=7$ and $d=4$. ..... 111
4.5 Convergence history for Test problem 2 when we apply the Galerkin condition with $\nu=0.4999, l=5, M=7$ and $d=4$. ..... 111
4.6 Convergence history for Test problem 1 when we apply the Petrov- Galerkin condition with $\nu=0.4999, l=5, M=7$ and $d=4$. ..... 112
4.7 Convergence history for Test problem 2 when we apply the Petrov- Galerkin condition with $\nu=0.4999, l=5, M=7$ and $d=4$. ..... 112
4.8 MINRES inner iteration counts for $P 2$ and $P 3$ when we apply the Petrov-Galerkin condition. ..... 114
4.9 The relative difference $\Delta_{k}$ against the dimension $n_{k}$ for Test problem 1when $\sigma=0.085, \nu=0.4999, l=6, M=5$ and $d=3$.115
5.1 Numerical eigenvalues of $\widetilde{K}_{r}$ for $r=1,2, \ldots, 7$ in Test problem 1computed on a spatial mesh with $l=5$.120
5.2 Numerical eigenvalues of $\widetilde{K}_{r}$ for $r=1,2, \ldots, 5$ in Test problem 2 computed on a spatial mesh with grid $l=5$. ..... 120
5.3 Convergence history for Test problem 1 with $\nu=0.4999, l=5$,$M=7$ and $d=4$ using the left preconditioning strategy.128
5.4 Convergence history for Test problem 2 with $\nu=0.4999, l=5$,$M=7$ and $d=4$ using the left preconditioning strategy.128
6.1 Expectations and variances of components of the pressure $p$ and thevelocity $\boldsymbol{u}$ when $l=6, M=7$ and $d=4$ for case 1 in Test problem 3. 135
6.2 Convergence history of $\Delta_{k}$ for case 1 in Test problem 3 with $M=7$ and $d=4$.1436.3 Convergence history for Test problem $\mathbf{3}$ when we apply the Galerkincondition with $l=5, M=7$ and $d=4$.1446.4 Convergence history for Test problem 4 when we apply the Galerkincondition with $l=5, M=7$ and $d=4$.144
6.5 Convergence history for Test problem 3 when we apply the Petrov- Galerkin condition with $l=5, M=7$ and $d=4$ ..... 145
6.6 Convergence history for Test problem 4 when we apply the Petrov- Galerkin condition with $l=5, M=7$ and $d=4$ ..... 145
6.7 Convergence history for Test problem 3 with $l=5, M=7$ and $d=4$ using the left preconditioning strategy. ..... 153
6.8 Convergence history for Test problem 4 with $l=5, M=7$ and $d=4$ using the left preconditioning strategy. ..... 153

## Abstract

We are interested in the efficient numerical solution of the structured linear systems that arise when we apply stochastic Galerkin mixed finite element methods (SGMFEMs) to systems of partial differential equations (PDEs) with inputs that depend on a possibly large number of uncertain parameters. In particular, we are interested in PDEs arising in engineering applications for which SGMFEMs give rise to saddle point systems, such as linear elasticity and fluid flow problems.

Despite being highly structured, saddle point systems associated with SGMFEMs are challenging to solve due to their extremely large size. The number of equations is the product of (i) the number of degrees of freedom associated with the chosen mixed finite element method on the spatial domain and (ii) the dimension of the polynomial space associated with the parameter domain. When we refine the finite element mesh and/or increase the degree of the parametric polynomial approximation to improve accuracy, the dimension of the associated linear system increases. When working on standard desktop computers, one cannot use conventional Krylov subspace methods for very fine SGMFEM discretisations because storing the required matrices and vectors quickly exhausts available memory. One potential remedy is to recast the discrete problem as a linear multi-term matrix equation (LMTME) and use reduced basis methods. Such methods construct low rank factored approximations to the solution matrix by projecting the problem onto a lower-dimensional space.

Our main aim is to develop a memory efficient solver for the discrete problems that arise when we apply SGMFEMs to a three-field linear elasticity model with parameterdependent Young modulus. The starting point is a reduced basis method known as Multi-RB that was recently proposed for LMTMEs associated with symmetric and positive definite problems. After reformulating the matrix equation, the scheme iteratively constructs a reduced basis using a strategy inspired by rational Krylov subspace approximation, then applies Galerkin projection and solves a reduced problem. When the problem is not positive definite, the best choice of preconditioner and projection technique to apply is not clear. For large-scale problems with solutions that cannot be approximated well by very low rank matrices, selecting a computationally feasible stopping condition is also problematic.

We modify the Multi-RB method to solve LMTMEs associated with symmetric and indefinite problems. For the linear elasticity problem of main interest, we critically assess the performance of the method using two distinct preconditioning strategies and two projection techniques. We also provide new eigenvalue analysis for the preconditioned systems. For smaller problems, we examine the convergence of the method by tracking the preconditioned and unpreconditioned residuals. Finally, to compare how the solver behaves on a different LMTME with the same structure, but whose solution matrix has different rank characteristics, we also consider a two-field groundwater flow model with parameter-dependent permeability coefficient.

## Declaration

No portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

## Copyright Statement

i. The author of this thesis (including any appendices and/or schedules to this thesis) owns certain copyright or related rights in it (the "Copyright") and s/he has given The University of Manchester certain rights to use such Copyright, including for administrative purposes.
ii. Copies of this thesis, either in full or in extracts and whether in hard or electronic copy, may be made only in accordance with the Copyright, Designs and Patents Act 1988 (as amended) and regulations issued under it or, where appropriate, in accordance with licensing agreements which the University has from time to time. This page must form part of any such copies made.
iii. The ownership of certain Copyright, patents, designs, trade marks and other intellectual property (the "Intellectual Property") and any reproductions of copyright works in the thesis, for example graphs and tables ("Reproductions"), which may be described in this thesis, may not be owned by the author and may be owned by third parties. Such Intellectual Property and Reproductions cannot and must not be made available for use without the prior written permission of the owner(s) of the relevant Intellectual Property and/or Reproductions.
iv. Further information on the conditions under which disclosure, publication and commercialisation of this thesis, the Copyright and any Intellectual Property and/or Reproductions described in it may take place is available in the University IP Policy (see http://documents.manchester.ac.uk/DocuInfo.aspx?DocID=487), in any relevant Thesis restriction declarations deposited in the University Library, The University Library's regulations (see http://www.manchester.ac.uk/library/aboutus/regulations) and in The University's Policy on Presentation of Theses.

## Acknowledgements

I thank my supervisor, Catherine Powell, for her guidance and support over the past four years. This thesis could not have been completed without her corrections and suggestions. I've learnt so much from her. I thank my co-supervisor, David Silvester, for his introduction to the basics of mixed finite element methods and matrix analysis during my first year. I also thank Valeria Simoncini for the discussion of the multi-term reduced basis method.

I thank my parents for their love all the time. I also thank them for allowing me to do my PhD in the UK. It has been a long time since I last left home. They keep in touch with me every day so let me not be too homesick and focus on my project.

Finally, I thank all my friends, from back home and from Manchester, for their company. Research life is challenging and sometimes lonely, especially during the Covid years. With their help and care, most of my four-year research life has been colourful.

## Chapter 1

## Introduction

Many important physical processes such as the deformation of elastic materials and fluid flows can be modelled using partial differential equations (PDEs). Since it is often too difficult to solve PDEs exactly, the study of numerical approximation schemes is important. For models consisting of systems of deterministic PDEs with inputs (such as coefficients, boundary conditions, etc.) that are known, we may approximate their solutions using standard finite element methods, see [11, 26]. For PDEs with more than one solution variable, Galerkin mixed finite element approximation is a popular strategy. Such approximation schemes often lead to discrete problems with saddle point structure that can be solved using standard Krylov subspace methods, see [26].

In this thesis, we are concerned with PDEs with more than one solution variable that have uncertain inputs that are modelled as random fields. Such stochastic problems can be reformulated as so-called parametric PDEs and solved using stochastic Galerkin mixed finite element methods, see [25, 27, 44, 58]. The resulting discrete problems have both saddle point and Kronecker product structures and are commonly of extremely large size. On a standard desktop computer, standard Krylov subspace methods can no longer be applied because storing the required matrices and vectors quickly exhausts available memory. This deficiency motivates our need in this thesis to develop a memory-efficient solver for high-dimensional problems. Using properties of the Kronecker product, we will exploit the fact that the linear systems of interest can be reformulated as matrix equations. Our new solver is an extension of a reduced basis method called Multi-RB that was introduced in [64] for matrix equations associated with a parameter-dependent PDE problem with one solution variable for which
stochastic Galerkin approximation yields a symmetric and positive definite system matrix. The key difference is that the problems we want to investigate have a more complex structure. The associated matrices are symmetric and indefinite with saddle point structure.

In this chapter, we introduce two PDE problems with uncertain inputs that can be formulated as parametric PDEs. In Section 1.1, for the benefit of readers who are not familiar with parametric PDEs, we first give a brief introduction to the associated deterministic formulations of these two problems and briefly explain how to apply Galerkin mixed finite element approximation. Then, in Section 1.2, we introduce the parameter-dependent formulations and outline the structure of the linear systems obtained when stochastic Galerkin mixed finite element approximation is applied. We also explain how to reformulate the discrete problems as matrix equations. In Section 1.3, we review some existing methods in the literature for solving matrix equations and motivate our chosen solution approach that will be developed later.

### 1.1 Deterministic saddle point problems

Let $D \subset \mathbb{R}^{2}$ be a spatial domain that is a Lipschitz polygon with boundary $\partial D$ that can be decomposed into two parts as

$$
\partial D=\partial D_{D} \cup \partial D_{N},
$$

where $\partial D_{D} \cap \partial D_{N}=\varnothing, \partial D \neq \varnothing$ and $\partial D_{N} \neq \varnothing$. First, we consider a linear elasticity model, which involves the unknown displacement $\boldsymbol{u}: D \rightarrow \mathbb{R}^{2}$ of an elastic material, a body force $\boldsymbol{f}: D \rightarrow \mathbb{R}^{2}$ and a stress tensor $\boldsymbol{\sigma}: D \rightarrow \mathbb{R}^{2 \times 2}$ of the form

$$
\begin{equation*}
\boldsymbol{\sigma}(\boldsymbol{x}):=2 \mu(\boldsymbol{x}) \boldsymbol{\epsilon}(\boldsymbol{x})+\lambda(\boldsymbol{x}) \nabla \cdot \boldsymbol{u}(\boldsymbol{x}) \boldsymbol{I}, \tag{1.1}
\end{equation*}
$$

which is defined in terms of the strain tensor

$$
\boldsymbol{\epsilon}(\boldsymbol{x})=\frac{1}{2}\left(\nabla \boldsymbol{u}(\boldsymbol{x})+(\nabla \boldsymbol{u}(\boldsymbol{x}))^{\top}\right),
$$

which itself depends on $\boldsymbol{u}$. In (1.1), $\boldsymbol{I} \in \mathbb{R}^{2 \times 2}$ is the identity matrix, and $\mu: D \rightarrow \mathbb{R}$ and $\lambda: D \rightarrow \mathbb{R}$ are the Lamé coefficients arising in strain-stress relationships, see $[48,73]$. These coefficients depend on the Young modulus $E: D \rightarrow \mathbb{R}$ and the Poisson
ratio $\nu$ via the following relations:

$$
\mu(\boldsymbol{x})=\frac{E(\boldsymbol{x})}{2(1+\nu)}, \quad \lambda(\boldsymbol{x})=\frac{E(\boldsymbol{x}) \nu}{(1+\nu)(1-2 \nu)} .
$$

Note that here, $E$ is spatially varying whereas $\nu \in(0,0.5)$ is assumed to be constant. The standard linear elasticity model with mixed boundary conditions is written as

$$
\begin{align*}
-\nabla \cdot \boldsymbol{\sigma}(\boldsymbol{x}) & =\boldsymbol{f}(\boldsymbol{x}), & & \boldsymbol{x} \in D  \tag{1.2}\\
\boldsymbol{u}(\boldsymbol{x}) & =\boldsymbol{g}(\boldsymbol{x}), & & \boldsymbol{x} \in \partial D_{D}  \tag{1.3}\\
\boldsymbol{\sigma}(\boldsymbol{x}) \cdot \boldsymbol{n} & =0, & & \boldsymbol{x} \in \partial D_{N} \tag{1.4}
\end{align*}
$$

where the vector $\boldsymbol{n}$ is the unit normal vector to the boundary pointing outward and $\boldsymbol{g}: D \rightarrow \mathbb{R}^{2}$ is a known function. When the Poisson ratio $\nu \rightarrow \frac{1}{2}, \lambda \rightarrow \infty$ and the material becomes less compressible.

If we apply standard finite element methods to approximate $\boldsymbol{u}$ in the linear elasticity problem (1.2)-(1.4), a priori error estimates are not uniformly convergent for all $\nu$. The approximation error in actual computations does not decrease at the predicted rate in the nearly incompressible case for most practical levels of discretisation. In this case, numerical approximations of the displacement $\boldsymbol{u}$ will deteriorate when $\nu \rightarrow 1 / 2$. This phenomenon is called locking, see [3, 86]. The underlying issue is that the coefficient $\lambda \rightarrow \infty$ and $\boldsymbol{u}$ must satisfy the constraint $\nabla \cdot \boldsymbol{u}=0$ in the incompressible limit. A standard way to avoid locking is to introduce an additional variable $p: D \rightarrow \mathbb{R}$

$$
\begin{equation*}
p(\boldsymbol{x}):=-\lambda(\boldsymbol{x}) \nabla \cdot \boldsymbol{u}(\boldsymbol{x}), \tag{1.5}
\end{equation*}
$$

which is often called the Herrmann pressure (see [35] for more details). If we substitute (1.5) into (1.1), the stress tensor $\boldsymbol{\sigma}$ now becomes

$$
\begin{equation*}
\boldsymbol{\sigma}(\boldsymbol{x}):=2 \mu(\boldsymbol{x}) \boldsymbol{\epsilon}(\boldsymbol{x})-p(\boldsymbol{x}) \boldsymbol{I} . \tag{1.6}
\end{equation*}
$$

This leads to a two-field linear elasticity problem

$$
\begin{align*}
-\nabla \cdot \boldsymbol{\sigma}(\boldsymbol{x}) & =\boldsymbol{f}(\boldsymbol{x}), & & \boldsymbol{x} \in D  \tag{1.7}\\
\nabla \cdot \boldsymbol{u}(\boldsymbol{x})+\lambda^{-1}(\boldsymbol{x}) p(\boldsymbol{x}) & =0, & & \boldsymbol{x} \in D \tag{1.8}
\end{align*}
$$

with the same mixed boundary conditions as in (1.3)-(1.4). This mixed formulation (1.7)-(1.8) is valid for all possible values of the Poisson ratio. When $\nu \rightarrow 1 / 2$, this
mixed formulation becomes the Stokes equations, see [14, 26, 31]. We can apply mixed approximations that are stable for the Stokes problem to solve this mixed formulation of the linear elasticity problem. In [43], several a posteriori error estimators for mixed approximations are proved to be robust in the incompressible limit for the linear elasticity problem (1.7)-(1.8).

Second, we consider the so-called Darcy model for groundwater flow. This combines Darcy's law

$$
\begin{equation*}
\boldsymbol{u}(\boldsymbol{x})=-A(\boldsymbol{x}) \nabla p(\boldsymbol{x}) \tag{1.9}
\end{equation*}
$$

which relates fluid velocity $\boldsymbol{u}: D \rightarrow \mathbb{R}^{2}$ to the pressure $p: D \rightarrow \mathbb{R}$ and permeability coefficient $A: D \rightarrow \mathbb{R}$, and a mass conservation law

$$
\nabla \cdot \boldsymbol{u}(\boldsymbol{x})=f(\boldsymbol{x}),
$$

where $f: D \rightarrow \mathbb{R}$ describes any sources/sinks. When the fluid is incompressible and there are no sources/sinks, then $f(\boldsymbol{x})=0$. Combining these equations gives the two-field model

$$
\begin{align*}
A^{-1}(\boldsymbol{x}) \boldsymbol{u}(\boldsymbol{x})+\nabla p(\boldsymbol{x}) & =\mathbf{0}, & & \boldsymbol{x} \in D,  \tag{1.10}\\
\nabla \cdot \boldsymbol{u}(\boldsymbol{x}) & =f(\boldsymbol{x}), & & \boldsymbol{x} \in D, \tag{1.11}
\end{align*}
$$

with mixed boundary conditions

$$
\begin{array}{ll}
p(\boldsymbol{x})=g(\boldsymbol{x}), & \boldsymbol{x} \in \partial D_{D} \\
\boldsymbol{u}(\boldsymbol{x}) \cdot \boldsymbol{n}=0, & \boldsymbol{x} \in \partial D_{N} \tag{1.13}
\end{array}
$$

See $[62,74]$ for further discussion.
We now apply Galerkin mixed finite element approximation to approximate the displacement $\boldsymbol{u}$ and the Herrmann pressure $p$ for the elasticity problem and the velocity $\boldsymbol{u}$ and the pressure $p$ for the groundwater flow problem. In this case, we choose two finite element subspaces for $\boldsymbol{u}$ and $p$ in these two problems that satisfy the inf-sup condition, as described in $[10,14]$.

The weak formulations of the two-field linear elasticity problem (1.7)-(1.8) with boundary conditions (1.3)-(1.4) and the two-field groundwater flow problem (1.10)(1.13) are saddle point problems. In abstract form, we seek $\boldsymbol{u} \in \boldsymbol{V}$ and $p \in Q$ such
that

$$
\begin{align*}
a(\boldsymbol{u}, \boldsymbol{v})+b(\boldsymbol{v}, p)=h(\boldsymbol{v}), & \forall \boldsymbol{v} \in \boldsymbol{V},  \tag{1.14}\\
b(\boldsymbol{u}, q)-c(p, q)=\ell(q), & \forall q \in Q, \tag{1.15}
\end{align*}
$$

where $\boldsymbol{V}$ and $Q$ are appropriate Hilbert spaces on $D, a: \boldsymbol{V} \times \boldsymbol{V} \rightarrow \mathbb{R}, b: Q \times \boldsymbol{V} \rightarrow \mathbb{R}$ and $c: Q \times Q \rightarrow \mathbb{R}$ are bilinear forms and $h: V \rightarrow \mathbb{R}$ and $\ell: Q \rightarrow \mathbb{R}$ are linear functionals. To approximate the solution in each case, we must choose compatible finite-dimensional spaces $\boldsymbol{V}_{h} \subset \boldsymbol{V}, Q_{h} \subset Q$ and find $\boldsymbol{u}_{h} \in \boldsymbol{V}_{h}, p_{h} \in Q_{h}$ by solving

$$
\begin{aligned}
a\left(\boldsymbol{u}_{h}, \boldsymbol{v}_{h}\right)+b\left(\boldsymbol{v}_{h}, p_{h}\right)=h\left(\boldsymbol{v}_{h}\right), & \forall \boldsymbol{v}_{h} \in \boldsymbol{V}_{h}, \\
b\left(\boldsymbol{u}_{h}, q_{h}\right)-c\left(p_{h}, q_{h}\right)=\ell\left(q_{h}\right), & \forall q_{h} \in Q_{h} .
\end{aligned}
$$

In both cases, this yields a discrete saddle point system of the form

$$
\left[\begin{array}{cc}
\mathcal{A} & \mathcal{B}^{\top}  \tag{1.16}\\
\mathcal{B} & -\mathcal{C}
\end{array}\right]\left[\begin{array}{l}
\mathbf{u} \\
\mathbf{p}
\end{array}\right]=\left[\begin{array}{l}
\mathbf{f}_{1} \\
\mathbf{f}_{2}
\end{array}\right] .
$$

If $E(\boldsymbol{x})$ and $A^{-1}(\boldsymbol{x})$ are strictly positive spatial functions, then the matrix $\mathcal{A}$ is symmetric and positive definite. The entries of the vectors $\mathbf{u}$ and $\mathbf{p}$ are the coefficients of $\boldsymbol{u}_{h}$ and $p_{h}$ when expanded in the chosen bases for $\boldsymbol{V}_{h}$ and $Q_{h}$. For the linear elasticity problem (1.7)-(1.8) with boundary conditions (1.3)-(1.4), the vectors $\mathbf{u}$ and $\mathbf{p}$ are associated with approximations of the displacement and the Herrmann pressure, respectively. For the groundwater flow problem (1.10)-(1.13), $\mathcal{C}=0$, and the vectors $\mathbf{u}$ and $\mathbf{p}$ are associated with approximations of the fluid velocity and the pressure, respectively.

### 1.2 Parameter-dependent saddle point problems

In real world applications, the inputs of PDEs, such as coefficients, boundary conditions and source terms, are often uncertain. For example, the Young modulus $E(\boldsymbol{x})$ in the linear elasticity problem (1.7)-(1.8) with boundary conditions (1.3)-(1.4) and the permeability coefficient $A(\boldsymbol{x})$ in the groundwater flow problem (1.10)-(1.13) are often uncertain. A common way to deal with uncertain quantities that are spatially varying is to model them as random fields with prescribed mean and spatial covariance functions. Such random fields can be expressed as Karhunen-Loève (KL) expansions,
see [4, 49, 55]. A KL expansion is an infinite series with random coefficients, which can be truncated after a finite number of terms for use in computations. We will discuss KL expansions in more detail in Chapter 2.

Consider again the linear elasticity problem (1.7)-(1.8) and suppose we model the Young modulus $E(\boldsymbol{x})$ as a random field of the form

$$
\begin{equation*}
E(\boldsymbol{x}, \omega):=e_{0}(\boldsymbol{x})+\sum_{r=1}^{M} e_{r}(\boldsymbol{x}) \xi_{r}(\omega), \quad \boldsymbol{x} \in D, \omega \in \Omega \tag{1.17}
\end{equation*}
$$

where $e_{0}(\boldsymbol{x})$ represents the mean, $\xi_{r}: \Omega \rightarrow \Gamma_{r}$, for $r=1,2, \ldots, M$, are independent mean zero real-valued random variables, $\Omega$ is a sample space associated with a probability space and $\Gamma_{r} \subset \mathbb{R}$. If we make the change of variable $y_{r}=\xi_{r}(\omega)$, we can rewrite the Young modulus as the parameter-dependent function

$$
\begin{equation*}
E(\boldsymbol{x}, \boldsymbol{y}):=e_{0}(\boldsymbol{x})+\sum_{r=1}^{M} e_{r}(\boldsymbol{x}) y_{r}, \quad \boldsymbol{x} \in D, \boldsymbol{y} \in \Gamma \tag{1.18}
\end{equation*}
$$

where $\boldsymbol{y}=\left(y_{1}, y_{2}, \ldots, y_{M}\right) \in \Gamma$, each $y_{r}$ is the image of the random variable $\xi_{r}$ appearing in (1.17), and the so-called parameter domain $\Gamma$ is defined as

$$
\Gamma:=\Gamma_{1} \times \cdots \times \Gamma_{M} \subset \mathbb{R}^{M}
$$

The associated parametric two-field linear elasticity problem is then written as

$$
\begin{align*}
-\nabla \cdot \boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{y}) & =\boldsymbol{f}(\boldsymbol{x}), & & \boldsymbol{x} \in D, \boldsymbol{y} \in \Gamma  \tag{1.19}\\
\nabla \cdot \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{y})+\lambda^{-1}(\boldsymbol{x}, \boldsymbol{y}) p(\boldsymbol{x}, \boldsymbol{y}) & =0, & & \boldsymbol{x} \in D, \boldsymbol{y} \in \Gamma \tag{1.20}
\end{align*}
$$

with the boundary conditions

$$
\begin{align*}
\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{y}) & =\boldsymbol{g}(\boldsymbol{x}), & & \boldsymbol{x} \in \partial D_{D}, \boldsymbol{y} \in \Gamma,  \tag{1.21}\\
\boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{y}) \cdot \boldsymbol{n} & =\mathbf{0}, & & \boldsymbol{x} \in \partial D_{N}, \boldsymbol{y} \in \Gamma . \tag{1.22}
\end{align*}
$$

Note that since $E$ is a function of $\boldsymbol{x} \in D$ and $\boldsymbol{y} \in \Gamma$, then so is the displacement $\boldsymbol{u}$ and the Herrmann pressure $p$. For each $\boldsymbol{y} \in \Gamma$, we obtain a different spatial solution.

In Chapter 2, we will approximate the solution of the parametric linear elasticity problem using a stochastic Galerkin mixed finite element method (SGMFEM). Such methods are not sampling methods. The idea is to find approximations to each solution field in tensor product spaces of the form $\boldsymbol{V}_{h} \otimes S_{d}$, where $\boldsymbol{V}_{h}$ is a finite-dimensional
space of functions on the spatial domain $D$ (a finite element space) and $S_{d}$ is a finitedimensional space of multivariate polynomials on the parameter domain $\Gamma$. As we shall see in Chapter 2, to implement stochastic Galerkin methods efficiently, we require that every term in the finite-dimensional weak formulation of the problem is separable. That is, the terms that depend on $\boldsymbol{x}$ can be separated from the terms that depend on $\boldsymbol{y}$. Unfortunately, in the Herrmann model, this is not the case because the coefficient $\lambda^{-1}(\boldsymbol{x}, \boldsymbol{y})$ appears in (1.20). We have

$$
\lambda^{-1}(\boldsymbol{x}, \boldsymbol{y})=\tilde{\lambda}^{-1} E^{-1}(\boldsymbol{x}, \boldsymbol{y})
$$

where we define the constant

$$
\begin{equation*}
\tilde{\lambda}:=\frac{\nu}{(1+\nu)(1-2 \nu)} . \tag{1.23}
\end{equation*}
$$

Although $E(\boldsymbol{x}, \boldsymbol{y})$ is separable, due to (1.18), $E^{-1}(\boldsymbol{x}, \boldsymbol{y})$ is not separable, and so $\lambda^{-1}(\boldsymbol{x}, \boldsymbol{y})$ is not separable either.

To remedy this, we introduce another auxiliary variable $\tilde{p}=p / E$. This allows us to eliminate $E^{-1}$ in the two-field problem (1.19)-(1.20). See [44] for more details. This leads to the following three-field parametric formulation

$$
\begin{align*}
-\nabla \cdot \boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{y}) & =\boldsymbol{f}(\boldsymbol{x}), & & \boldsymbol{x} \in D, \boldsymbol{y} \in \Gamma,  \tag{1.24}\\
\nabla \cdot \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{y})+\tilde{\lambda}^{-1} \tilde{p}(\boldsymbol{x}, \boldsymbol{y}) & =0, & & \boldsymbol{x} \in D, \boldsymbol{y} \in \Gamma,  \tag{1.25}\\
\tilde{\lambda}^{-1} p(\boldsymbol{x}, \boldsymbol{y})-\tilde{\lambda}^{-1} E(\boldsymbol{x}, \boldsymbol{y}) \tilde{p}(\boldsymbol{x}, \boldsymbol{y}) & =0, & & \boldsymbol{x} \in D, \boldsymbol{y} \in \Gamma, \tag{1.26}
\end{align*}
$$

with the same boundary conditions as in (1.21)-(1.22).
The associated weak formulation can be expressed as: find $\boldsymbol{u} \in \boldsymbol{V}, p \in Q$ and $\tilde{p} \in Q$ such that

$$
\begin{align*}
a(\boldsymbol{u}, \boldsymbol{v})+b(\boldsymbol{v}, p) & =h(\boldsymbol{v}), & & \forall \boldsymbol{v} \in \boldsymbol{V},  \tag{1.27}\\
b(\boldsymbol{u}, q)-c(\tilde{p}, q) & =0, & & \forall q \in Q,  \tag{1.28}\\
-c(p, \tilde{q})+d(\tilde{p}, \tilde{q}) & =0, & & \forall \tilde{q} \in Q, \tag{1.29}
\end{align*}
$$

where $\boldsymbol{V}$ and $Q$ are now appropriate function spaces on $D \times \Gamma$. If we define two new bilinear forms $a^{*}(\cdot, \cdot): \boldsymbol{W} \times \boldsymbol{W}$, and $b^{*}(\cdot, \cdot): \boldsymbol{W} \times Q$ as follows:

$$
\begin{aligned}
a^{*}(\boldsymbol{r}, \boldsymbol{w}) & :=a(\boldsymbol{u}, \boldsymbol{v})+d(\tilde{p}, \tilde{q}), \\
b^{*}(\boldsymbol{w}, p) & :=b(\boldsymbol{v}, p)-c(p, \tilde{q}),
\end{aligned}
$$

where $\boldsymbol{W}:=\boldsymbol{V} \times Q, \boldsymbol{r}=(\boldsymbol{u}, \tilde{p}), \boldsymbol{w}=(\boldsymbol{v}, \tilde{q})$ and add (1.29) to (1.27), we obtain the following saddle point problem

$$
\begin{aligned}
a^{*}(\boldsymbol{r}, \boldsymbol{w})+b^{*}(\boldsymbol{w}, p) & =h^{*}(\boldsymbol{w}), & & \forall \boldsymbol{w} \in \boldsymbol{W}, \\
b^{*}(\boldsymbol{r}, q) & =0, & & \forall q \in Q,
\end{aligned}
$$

where $h^{*}(\boldsymbol{w}):=h(\boldsymbol{v})$. Hence, we see that this has the same structure as the two-field saddle point problem (1.14)-(1.15).

To apply a stochastic Galerkin mixed finite element approximation, we choose finite dimensional spaces of the form

$$
\boldsymbol{V}_{h d}=\boldsymbol{V}_{h} \otimes S_{d} \subset \boldsymbol{V}, \quad Q_{h d}=Q_{h} \otimes S_{d} \subset Q
$$

and seek $\boldsymbol{u}_{h d} \in \boldsymbol{V}_{h d}, \tilde{p}_{h d} \in Q_{h d}$, and $p_{h d} \in Q_{h d}$ such that

$$
\begin{aligned}
a\left(\boldsymbol{u}_{h d}, \boldsymbol{v}_{h d}\right)+b\left(\boldsymbol{v}_{h d}, p_{h d}\right) & =h\left(\boldsymbol{v}_{h d}\right), & & \forall \boldsymbol{v}_{h d} \in \boldsymbol{V}_{h d}, \\
b\left(\boldsymbol{u}_{h d}, q_{h d}\right)-c\left(\tilde{p}_{h d}, q_{h d}\right) & =0, & & \forall q_{h d} \in Q_{h d} \\
-c\left(p_{h d}, \tilde{q}_{h d}\right)+d\left(\tilde{p}_{h d}, \tilde{q}_{h d}\right) & =0, & & \forall \tilde{q}_{h d} \in Q_{h d} .
\end{aligned}
$$

As we will show in Chapter 2, this yields a discrete saddle point system of the same form as (1.16), where $\mathcal{C}=0$, and $\mathcal{A}$ and $\mathcal{B}$ have the following block structures

$$
\begin{align*}
& \mathcal{A}:=\left[\begin{array}{cc|c}
\beta \sum_{r=0}^{M} G_{r} \otimes A_{r}^{(11)} & \beta \sum_{r=0}^{M} G_{r} \otimes A_{r}^{(21)} & 0 \\
\beta \sum_{r=0}^{M} G_{r} \otimes A_{r}^{(12)} & \beta \sum_{r=0}^{M} G_{r} \otimes A_{r}^{(22)} & 0 \\
\hline 0 & 0 & \tilde{\lambda}^{-1} \sum_{r=0}^{M} G_{r} \otimes D_{r}
\end{array}\right]  \tag{1.30}\\
& \mathcal{B}:=\left[\begin{array}{lll}
G_{0} \otimes B_{1} & G_{0} \otimes B_{2} & -\tilde{\lambda}^{-1} G_{0} \otimes C
\end{array}\right] . \tag{1.31}
\end{align*}
$$

Here, $\otimes$ denotes the matrix Kronecker product. For two matrices $A \in \mathbb{R}^{m \times n}$ and $B \in \mathbb{R}^{p \times q}$, the Kronecker product is defined as (see [45, 82])

$$
A \otimes B:=\left[\begin{array}{ccc}
a_{11} B & \cdots & a_{1 n} B \\
\vdots & \ddots & \vdots \\
a_{m 1} B & \cdots & a_{m n} B
\end{array}\right]
$$

In (1.30), the constant $\beta$ is defined as

$$
\begin{equation*}
\beta:=\frac{1}{1+\nu}, \tag{1.32}
\end{equation*}
$$

the symmetric matrices $G_{r}$ for $r=0,1, \ldots, M$ are associated with the chosen parametric approximation space $S_{d}$, and the matrices $A_{r}^{(i j)}$ for $i, j=1,2, r=0,1, \ldots, M$, as well as the matrices $B_{1}, B_{2}, C$ in (1.31) and $D_{r}$, for $0,1, \ldots, M$, are associated with the chosen spatial approximation spaces $\boldsymbol{V}_{h}$ and $Q_{h}$. Full details will be given in Chapter 2.

Analogously, we can model the inverse of the permeability coefficient $A^{-1}(\boldsymbol{x})$ in the groundwater flow problem in (1.10)-(1.13) as a parameter-dependent function of the form

$$
\begin{equation*}
A^{-1}(\boldsymbol{x}, \boldsymbol{y})=a_{0}(\boldsymbol{x})+\sum_{r=1}^{M} a_{r}(\boldsymbol{x}) y_{r}, \quad \boldsymbol{x} \in D, \boldsymbol{y} \in \Gamma \tag{1.33}
\end{equation*}
$$

Substituting (1.33) into (1.10)-(1.11) yields the parametric groundwater flow problem

$$
\begin{align*}
A^{-1}(\boldsymbol{x}, \boldsymbol{y}) \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{y})+\nabla p(\boldsymbol{x}, \boldsymbol{y}) & =0, & & \boldsymbol{x} \in D, \boldsymbol{y} \in \Gamma,  \tag{1.34}\\
\nabla \cdot \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{y}) & =f(\boldsymbol{x}), & & \boldsymbol{x} \in D, \boldsymbol{y} \in \Gamma, \tag{1.35}
\end{align*}
$$

with the mixed boundary conditions

$$
\begin{array}{rlrl}
p(\boldsymbol{x}, \boldsymbol{y}) & =g(\boldsymbol{x}), & \boldsymbol{x} \in \partial D_{D}, \boldsymbol{y} \in \Gamma, \\
\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{y}) \cdot \boldsymbol{n} & =0, & & \boldsymbol{x} \in \partial D_{N}, \boldsymbol{y} \in \Gamma . \tag{1.37}
\end{array}
$$

The weak formulation of (1.34)-(1.37) is also a saddle point problem. If we apply a SGMFEM to this problem, we again obtain a linear system of the form (1.16) with $\mathcal{C}=0$. As we will show in Chapter 6 , in this case, $\mathcal{A}$ and $\mathcal{B}$ have the simpler structures

$$
\mathcal{A}:=\sum_{r=0}^{M} G_{r} \otimes A_{r}, \quad \mathcal{B}:=G_{0} \otimes B .
$$

Again, the symmetric matrices $G_{r}$ are associated with the chosen parametric approximation space, and the matrices $A_{r}$, for $r=0,1, \ldots, M$, and $B$ are associated with the chosen spatial approximation spaces.

The key point is that for both the parameter-dependent three-field linear elasticity problem and the two-field groundwater flow problem, stochastic Galerkin mixed finite element approximation leads to a saddle point problem of the form (1.16) with
blocks that have Kronecker product structure. By reordering the degrees of freedom associated with the solution fields, these linear systems can also be expressed as

$$
\begin{equation*}
\left(\sum_{r=0}^{M} G_{r} \otimes K_{r}\right) \mathbf{x}=\mathbf{h} \tag{1.38}
\end{equation*}
$$

where the matrices $K_{r}$ for $r=0,1, \ldots, M$ are associated with the spatial approximation for all solution fields. We now introduce a standard property of the matrix Kronecker product, see [82]. For two matrices $A$ and $B$ and a vector $\mathbf{v}$ of appropriate sizes

$$
\begin{equation*}
(A \otimes B) \mathbf{v}=\operatorname{vec}\left(B V A^{\top}\right) \tag{1.39}
\end{equation*}
$$

where $V=\operatorname{array}(\mathbf{v})$, the function $\operatorname{vec}(\cdot)$ stacks the columns of a matrix one on top of another to form a vector and the function $\operatorname{array}(\cdot)$ is the inverse function of $\operatorname{vec}(\cdot)$. Using the property (1.39), the linear system (1.38) can also be reformulated as a linear multi-term matrix equation (LMTME) of the form

$$
\begin{equation*}
\sum_{r=0}^{M} K_{r} X G_{r}^{\top}=H \tag{1.40}
\end{equation*}
$$

where $X=\operatorname{array}(\mathbf{x})$ is the solution matrix and $H=\operatorname{array}(\mathbf{h})$. That is, if we stack the columns of $X$ on top of one another, we obtain the solution vector $\mathbf{x}$ of the linear system (1.38).

In this thesis, we focus on the memory-efficient numerical solution of the LMTMEs that arise when we apply SGMFEMs to parameter-dependent saddle point problems such as the ones introduced above. First, we briefly review some existing numerical methods for solving matrix equations.

### 1.3 Methods for solving matrix equations

As a starting point, we consider a simpler group of matrix equations with just two terms called Sylvester equations, which have the following form

$$
\begin{equation*}
K X+X G=H \tag{1.41}
\end{equation*}
$$

Here, $K \in \mathbb{R}^{n \times n}, G \in \mathbb{R}^{m \times m}$ and $H \in \mathbb{R}^{n \times m}$ are known. The solvability of (1.41) is discussed in [1]. A comprehensive overview of solution techniques for $X \in \mathbb{R}^{n \times m}$ is
provided in [78]. When both $n$ and $m$ are small, decomposition methods can be used to find the exact solution. For example, the Bartels-Stewart algorithm [7] was one of the first effective methods for solving Sylvester equations and the Hessenberg-Schur method [32] is a modification that reduces computational cost. The main strategy is to decompose $K^{\top}$ and $G$ by Schur decomposition or Hessenberg decomposition and split the matrix equation into a group of linear systems which can be solved by back substitution. When at least one of $K$ and $G$ is large, Smith's method (see [56, 79, 85]), the alternating direction implicit (ADI) method (see [8, 23, 24, 61, 85]) and iterative projection methods (see [38, 39, 41]) can be used instead to obtain approximations to the solution matrix $X$.

First, we briefly introduce Smith's method. For $p, q>0$, we can rewrite (1.41) as

$$
\begin{equation*}
\left(p I_{n}+K\right) X\left(q I_{m}+G\right)-\left(K-q I_{n}\right) X\left(G-p I_{m}\right)=(p+q) H \tag{1.42}
\end{equation*}
$$

If we assume the real parts of the eigenvalues of $K$ and $G$ are positive, then $p I_{n}+K$ and $q I_{m}+G$ are nonsingular and we can multiply (1.42) by their inverses on both sides to obtain the following matrix equation

$$
\begin{equation*}
X-\mathcal{K} X \mathcal{G}=\mathcal{H} \tag{1.43}
\end{equation*}
$$

where

$$
\begin{aligned}
\mathcal{K} & :=\left(p I_{n}+K\right)^{-1}\left(K-q I_{n}\right), \\
\mathcal{G} & :=\left(G-p I_{m}\right)\left(q I_{m}+G\right)^{-1}, \\
\mathcal{H} & :=(p+q)\left(p I_{n}+K\right)^{-1} H\left(q I_{m}+G\right)^{-1} .
\end{aligned}
$$

The matrices $\mathcal{K}$ and $\mathcal{G}$ have spectral radii less than one. Hence, we have $\mathcal{K}^{\ell} \rightarrow 0$ and $\mathcal{G}^{\ell} \rightarrow 0$ as $\ell \rightarrow \infty$. See [80] for more details about the transformation in (1.43). The equation (1.43) has a formal solution

$$
X=\sum_{k=1}^{\infty} \mathcal{K}^{k-1} \mathcal{H} \mathcal{G}^{k-1}
$$

Based on this, Barnett and Storey (see [6]) suggested the following iteration

$$
\begin{equation*}
X_{0}=\mathcal{H}, \quad X_{k+1}=\mathcal{H}+\mathcal{K} X_{k} \mathcal{G}, \quad k=0,1,2, \ldots, \tag{1.44}
\end{equation*}
$$

which can be expressed as

$$
X_{k}=\sum_{l=0}^{k} \mathcal{K}^{l} \mathcal{H} \mathcal{G}^{l}
$$

However, the rate of convergence for the iteration (1.44) is slow. Smith (see [79]) proposed the following iteration to accelerate the convergence

$$
\begin{equation*}
X_{0}=\mathcal{H}, \quad X_{k+1}=X_{k}+\mathcal{K}^{2^{k}} X_{k} \mathcal{G}^{2^{k}}, \quad k=0,1,2, \ldots, \tag{1.45}
\end{equation*}
$$

which also can be expressed as

$$
X_{k}=\sum_{l=0}^{2^{k}-1} \mathcal{K}^{l} \mathcal{H} \mathcal{G}^{l}
$$

The optimal choice of $p$ and $q$ is found by solving the minimax problem

$$
\begin{equation*}
\min _{p, q>0} \max _{s \in \Lambda(K), t \in \Lambda(G)}\left|\frac{(t-p)(s-q)}{(t+q)(s+p)}\right| \tag{1.46}
\end{equation*}
$$

where $\Lambda(\cdot)$ denotes the spectrum of a matrix.
Next, we outline ADI iteration for (1.41). Starting with $X_{0}=0$, the iteration is

$$
\begin{align*}
\left(K+p_{k} I_{n}\right) X_{k-\frac{1}{2}} & =H-X_{k-1}\left(G-p_{k} I_{m}\right),  \tag{1.47}\\
X_{k}\left(G+q_{k} I_{m}\right) & =H-\left(K-q_{k} I_{n}\right) X_{k-\frac{1}{2}}, \quad k=1,2, \ldots,
\end{align*}
$$

where the parameters $p_{k}$ and $q_{k}$ can be selected at each iteration to speed up convergence. In practice, we select a fixed number $J$ of parameters before the first iteration and cycle through these after every $J$ iterations. The choice of parameters $p_{k}$ and $q_{k}$, for $k=1, \ldots, J$, is similar to (1.46). The optimal choice is found by solving the following minimax problem, see [72, 83]

$$
\begin{equation*}
\min _{p_{k}, q_{k}>0} \max _{s \in \Lambda(K), t \in \Lambda(G)} \prod_{k=1}^{J}\left|\frac{\left(t-p_{k}\right)\left(s-q_{k}\right)}{\left(s+p_{k}\right)\left(t+q_{k}\right)}\right| . \tag{1.48}
\end{equation*}
$$

However, the spectra of $K$ and $G$ are usually not available. Therefore we generally approximate the problem with $\Lambda(K)$ and $\Lambda(G)$ replaced by their estimates. See [50, $72,84]$ for more details. In [85], an upper bound for the relative error

$$
e_{k}=\frac{\left\|X-X_{k}\right\|_{F}}{\|X\|_{F}}
$$

for Smith's method (1.45) and ADI iteration (1.47) is discussed and the computational costs per iteration are also analysed for these two methods.

Finally, we introduce iterative projection methods for the Sylvester equation (1.41). One can generate approximations of the form $X_{k}=V_{k} Y_{k} W_{k}^{\top}$ iteratively, where $Y_{k} \in$ $\mathbb{R}^{n_{k} \times m_{k}}$ needs to be determined and the columns of $V_{k} \in \mathbb{R}^{n \times n_{k}}$ and $W_{k} \in \mathbb{R}^{m \times m_{k}}$ are chosen to be orthonormal bases for some approximation spaces $\mathcal{V}_{k}$ and $\mathcal{W}_{k}$, respectively. In [69], standard Krylov subspaces are used and basis vectors are generated by the Arnoldi algorithm to approximate solutions of the Lyapunov equation, which is a special case of the Sylvester equation when $G=K^{\top}$. Apart from standard Krylov subspaces, one can also choose global Krylov subspaces (see [41]), rational Krylov subspaces (see $[17,19,66]$ ), or the extended Krylov subspace (see [12, 77]). Using the property (1.39) of the Kronecker product, we can determine $Y_{k}$ at each iteration by minimising some norm of the residual

$$
\mathbf{r}_{k}=\mathcal{M}\left(W_{k} \otimes V_{k}\right) \mathbf{y}_{k}-\mathbf{h},
$$

where $\mathbf{y}_{k}=\operatorname{vec}\left(Y_{k}\right)$ and

$$
\mathcal{M}:=I_{m} \otimes K+G^{\top} \otimes I_{n}
$$

For example, we can minimise the energy norm $\left\|\mathbf{r}_{k}\right\|_{\mathcal{M}}$ when $\mathcal{M}$ is symmetric and positive definite by imposing a Galerkin condition on $\mathbf{r}_{k}$. That is, by making $\mathbf{r}_{k}$ orthogonal to the columns of $W_{k} \otimes V_{k}$. Imposing the Galerkin condition ensures that the energy norm of the residual decreases monotonically and the convergence of the Conjugate Gradient (CG) method (see [36, 71]) is based on this property. In addition to imposing a Galerkin condition, other conditions to make $\mathbf{r}_{k}$ orthogonal to different constraint spaces are discussed in $[38,54,60,71]$. Imposing the Galerkin condition on $\mathbf{r}_{k}$ gives

$$
\left(W_{k}^{\top} \otimes V_{k}^{\top}\right) \mathcal{M}\left(W_{k} \otimes V_{k}\right) \mathbf{y}_{k}=\left(W_{k}^{\top} \otimes V_{k}^{\top}\right) \mathbf{h}
$$

which is equivalent to the matrix equation

$$
\begin{equation*}
\left(V_{k}^{\top} K V_{k}\right) Y_{k}+Y_{k}\left(W_{k}^{\top} G W_{k}\right)=V_{k}^{\top} H W_{k} . \tag{1.49}
\end{equation*}
$$

Note that $V_{k}^{\top} K V_{k} \in \mathbb{R}^{n_{k} \times n_{k}}$ and $W_{k}^{\top} G W_{k} \in \mathbb{R}^{m_{k} \times m_{k}}$. If $X$ can be well approximated by a low-rank matrix, then $n_{k}, m_{k} \ll \min (m, n)$ and solving the reduced problem (1.49) will be quicker and more memory-efficient than solving (1.41). In general, the solution
matrix $X$ can be well approximated by a low-rank matrix if it has rapidly decaying singular values. Let $\sigma_{1} \geqslant \cdots \geqslant \sigma_{\min (m, n)}$ denote the singular values of the solution matrix $X$ and let $p$ denote the rank of the right hand side matrix $H$. An upper bound for $\sigma_{p r+1} / \sigma_{1}$ for $1 \leqslant p r<n$ is given in [72, Theorem 2.1.1]. For fixed $K$ and $G$ in (1.41), the lower the rank of $H$, the faster the singular values of $X$ will decay and then applying projection methods to obtain a low rank approximation is appealing.

There is a large variety of projection methods. For example, standard Krylov subspace methods for solving linear systems are projection methods, see [71]. Recall that linear systems associated with the multi-term matrix equation (1.40) or the twoterm Sylvester equation (1.41) have Kronecker product structures. In [5, 9, 47, 51, 53], Krylov subspace methods are combined with low rank approximation to solve such linear systems. The procedure at each iteration is almost the same as a standard Krylov subspace method. The basic idea is to truncate the approximation $X_{k}$ and other matrices of the same size as the solution matrix $X$ based on the decay of their singular values at each iteration. That is, one fixes the rank at the start and computes the singular values of $X_{k}$, residuals, search directions, etc. in each iteration. Then $X_{k}$ and the corresponding matrices can be stored in low rank format. However, if the rank fixed at the start is too small, the error between $X_{k}$ and the actual solution could be too large.

We now consider linear multi-term matrix equations of the form (1.40). In this thesis, we consider a one-sided projection method known as Multi-RB [64] for solving (1.40). Unlike the low rank Krylov subspace methods we mentioned before, the rank for $X_{k}$ is not fixed and so it is an adaptive algorithm that will terminate when a stopping condition is achieved. It constructs a set of orthogonal basis vectors $V_{k}$ based on rational Krylov subspaces and let $W_{k}$ be the $m \times m$ identity matrix and then imposes a Galerkin condition on the residual. Since the coefficient matrix of the Kronecker form (1.38) is symmetric and positive definite in [64], the Galerkin condition ensures that the energy norm of the residual at every iteration is monotonically decreasing. When the coefficient matrix of (1.38) is not symmetric and positive definite, one can impose the Petrov-Galerkin condition and the 2-norm of the residual is minimised at every iteration. In [60], a Petrov-Galerkin condition is proposed to minimise the 2-norm of the residual for the linear system associated with a small scale multi-term
matrix equation. The convergence of some Krylov subspaces methods for solving linear systems such as the minimal residual method (MINRES) and the generalised minimal residual method (GMRES) is based on this property, see [59, 70].

There are also other methods for solving multi-term matrix equations of the form (1.40). The approach in $[46,52]$ is based on an alternating energy minimisation framework, which seeks an approximation in the form of a product of two matrices and solves a pair of energy minimisation problems iteratively. These methods increase the approximation rank by one per iteration. In [40], a projection method is developed for a generalised Sylvester equation of the form

$$
K X+X G+\sum_{r=1}^{M} K_{r} X G_{r}=H
$$

This method is based on the extended Krylov subspace method for the Sylvester equation (see [12, 77]) and it increases the approximation rank by two per iteration. For problems whose true solution rank is not low enough, these methods may require a large number of iterations to satisfy the desired stopping condition.

### 1.4 Outline

In Chapter 2, we will discuss in more detail the three-field linear elasticity problem with parameter-dependent Young modulus, and apply a SGMFEM to the problem. We then consider the associated symmetric and indefinite linear system with Kronecker product structure, and use the preconditioned MINRES algorithm to solve it. For large problems, we show that memory is quickly exhausted, which motivates the need for more memory-efficient solution approaches. In Chapter 3, we rewrite the linear system as a linear multi-term matrix equation (LMTME). We investigate the decay of the singular values of the solution matrix and then consider how to extend the multi-term reduced basis method (Multi-RB) from [64] to this new class of problems. We introduce two distinct preconditioning strategies and investigate two projection techniques. In Chapters $4-5$, we perform some numerical experiments using the two preconditioning strategies on test problems. To compare how the solver behaves on a different LMTME with the same structure, we also apply the Multi-RB method to the parameter-dependent groundwater flow problem in Chapter 6. In Chapter 7, we
draw some conclusions about the performance of the reduced basis solver based on the numerical results obtained.

Below, we summarise the key contributions of this thesis.

1. Key contributions in Chapter 2:
(a) Proposition 2.20, which shows that the first block of the coefficient matrix associated with the parameter-dependent linear elasticity problem is symmetric and positive definite.
(b) Lemma 2.22 and Theorem 2.23 provide new theoretical eigenvalue bounds for preconditioned system matrices with a particular preconditioner $\mathcal{M}_{1}$.
(c) A detailed investigation of the performance of preconditioned MINRES using two symmetric and positive definite preconditioners.
2. Key contributions in Chapter 3:
(a) We introduce two new preconditioning strategies for the matrix equation (3.1). A symmetric preconditioning strategy is applied with a symmetric and positive definite preconditioner. A left preconditioning strategy is also applied with a symmetric and indefinite preconditioner.
(b) Lemmas 3.1-3.2 provide new theoretical eigenvalue bounds for the left matrices associated with the symmetric preconditioning strategy.
(c) Lemma 3.3 provides new theoretical eigenvalue bounds for the left matrices associated with the left preconditioning strategy.
(d) We incorporate a Petrov-Galerkin condition into the Multi-RB method. The original scheme used a Galerkin condition.
3. Key contributions in Chapter 4: exhaustive numerical experiments and convergence study of three different errors for the parameter-dependent linear elasticity problem using our solver with the chosen symmetric preconditioning strategy.
4. Key contributions in Chapter 5: exhaustive numerical experiments and convergence study of three different errors for the parameter-dependent linear elasticity problem using our solver with the chosen left preconditioning strategy.
5. Key contributions in Chapter 6: exhaustive numerical experiments and convergence study of three different errors for the parameter-dependent groundwater flow problem using our solver with two preconditioning strategies.

## Chapter 2

## Parameter-dependent Linear Elasticity Problem

In this chapter, we give more details about the three-field parameter-dependent linear elasticity problem (1.24)-(1.26) for nearly incompressible materials with uncertain Young's modulus that was introduced in Section 1.2. First, we discuss the chosen model for the uncertain Young modulus. We then introduce a stochastic Galerkin mixed finite element method and apply it to the parameter-dependent linear elasticity problem. Finally, we apply a standard Krylov subspace method to the linear systems associated with two test problems and explain why large-scale problems cannot be solved. This motivates the need for a more memory-efficient solution approach.

Recall that we want to find the displacement $\boldsymbol{u}: D \times \Gamma \rightarrow \mathbb{R}^{2}$, the Herrmann pressure $p: D \times \Gamma \rightarrow \mathbb{R}$ and the auxiliary pressure $\tilde{p}: D \times \Gamma \rightarrow \mathbb{R}$ such that

$$
\begin{align*}
-\nabla \cdot \boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{y}) & =\boldsymbol{f}(\boldsymbol{x}), & \boldsymbol{x} \in D, \boldsymbol{y} \in \Gamma  \tag{2.1}\\
\nabla \cdot \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{y})+\tilde{\lambda}^{-1} \tilde{p}(\boldsymbol{x}, \boldsymbol{y}) & =0, & \boldsymbol{x} \in D, \boldsymbol{y} \in \Gamma  \tag{2.2}\\
\tilde{\lambda}^{-1} p(\boldsymbol{x}, \boldsymbol{y})-\tilde{\lambda}^{-1} E(\boldsymbol{x}, \boldsymbol{y}) \tilde{p}(\boldsymbol{x}, \boldsymbol{y}) & =0, & \boldsymbol{x} \in D, \boldsymbol{y} \in \Gamma \tag{2.3}
\end{align*}
$$

with the homogenous mixed boundary conditions

$$
\begin{align*}
\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{y})=\mathbf{0}, & \boldsymbol{x} \in \partial D_{D}, \boldsymbol{y} \in \Gamma  \tag{2.4}\\
\boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{y}) \cdot \boldsymbol{n}=\mathbf{0}, & \boldsymbol{x} \in \partial D_{N}, \boldsymbol{y} \in \Gamma \tag{2.5}
\end{align*}
$$

Recall that $\boldsymbol{\sigma}: D \rightarrow \mathbb{R}^{2 \times 2}$ is defined in (1.6) for the deterministic problem (1.7)-(1.8). The stress tensor $\boldsymbol{\sigma}: D \times \Gamma \rightarrow \mathbb{R}^{2 \times 2}$ for the parameter-dependent problem (2.1)-(2.5)
is defined by

$$
\boldsymbol{\sigma}(\boldsymbol{x}, \boldsymbol{y}):=\mu(\boldsymbol{x}, \boldsymbol{y})\left(\nabla \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{y})+(\nabla \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{y}))^{\top}\right)+p(\boldsymbol{x}, \boldsymbol{y}) \boldsymbol{I} .
$$

In this model, we assume the body force $\boldsymbol{f}: D \rightarrow \mathbb{R}^{2}$ is deterministic, the Poisson ratio $\nu \in\left(0, \frac{1}{2}\right)$ is a known constant, and hence so is $\tilde{\lambda}$ which was defined in (1.23). The solution fields $\boldsymbol{u}, p, \tilde{p}$ and the Young modulus $E$ are all parameter-dependent. We model $E$ as a linear parameter-dependent function of the form

$$
\begin{equation*}
E(\boldsymbol{x}, \boldsymbol{y}):=e_{0}(\boldsymbol{x})+\sum_{r=1}^{M} e_{r}(\boldsymbol{x}) y_{r}, \quad \boldsymbol{x} \in D, \boldsymbol{y} \in \Gamma, \tag{2.6}
\end{equation*}
$$

where each $y_{r} \in \Gamma_{r}$ is the image of a random variable $\xi_{r}: \Omega \rightarrow \Gamma_{r} \subset \mathbb{R}$ with known probability density function $\rho_{r}\left(y_{r}\right): \Gamma_{r} \rightarrow \mathbb{R}$. We will assume these random variables are independent so that the joint probability density function is given by

$$
\begin{equation*}
\rho(\boldsymbol{y})=\prod_{r=1}^{M} \rho_{r}\left(y_{r}\right) . \tag{2.7}
\end{equation*}
$$

Parameter-dependent expressions of the form (2.6) arise when we model spatially varying uncertain inputs as random fields and use the so-called Karhunen-Loève expansion. We will give more details about this next.

### 2.1 Random fields and the Karhunen-Loève expansion

In order to explain the chosen model for $E$ in (2.6), we introduce some technical definitions relating to real-valued random variables and random fields and then introduce the Karhunen-Loève expansion. The material in this section mainly follows [55].

Definition 2.1 ( $\sigma$-algebra). Given a set $X$, a collection of subsets $\mathcal{F}$ is a $\sigma$-algebra if
(i) $\varnothing \in \mathcal{F}$,
(ii) the complement $A^{c}:=\{x \in X ; x \notin A\} \in X$ for all $A \in X$,
(iii) the union $\bigcup_{i \in \mathbb{N}} A_{i} \in \mathcal{F}$ for $A_{i} \in \mathcal{F}$.

Definition 2.2 (measure). For a set $X$ and a $\sigma$-algebra $\mathcal{F}$ of $X$, the pair $(X, \mathcal{F})$ is known as a measurable space. A measure $\mu$ on $(X, \mathcal{F})$ is defined as a function $\mu: \mathcal{F} \rightarrow \mathbb{R}^{+} \cup \infty$ that satisfies
(i) $\mu(\varnothing)=0$,
(ii) $\mu\left(\bigcup_{i \in \mathbb{N}} A_{i}\right)=\sum_{i \in \mathbb{N}} \mu\left(A_{i}\right)$ if $A_{i} \in \mathcal{F}$ are disjoint (i.e., $A_{i} \cap A_{j}=\varnothing$ for $i \neq j$ ).

Definition 2.3 (probability space). A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is composed of a sample space $\Omega$ containing all possible outcomes of a probabilistic event, a $\sigma$-algebra $\mathcal{F}$ of $\Omega$, and a probability measure $\mathbb{P}$. The probability measure $\mathbb{P}$ is a mapping from $\mathcal{F}$ to $[0,1]$ and satisfies $\mathbb{P}(\Omega)=1$.

Definition 2.4 (random variable). Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a measurable space $(\Gamma, \mathcal{G})$ with $\Gamma \subset \mathbb{R}^{M}$, $X$ is a $\Gamma$-valued random variable if it is a measurable function from $(\Omega, \mathcal{F})$ to $(\Gamma, \mathcal{G})$.

Definition 2.5 (probability density function). If $X$ is a $\Gamma$-valued random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ with $\Gamma \subset \mathbb{R}$, then it has an associated probability distribution $\mathbb{P}_{X}$ and the image space is $\left(\Gamma, \mathcal{B}(\Gamma), \mathbb{P}_{X}\right)$ where $\mathcal{B}(\Gamma)$ denotes the Borel $\sigma$-algebra. For any possible outcome $A \in \mathcal{B}(\Gamma)$, the probability that $X(\omega) \in A$ can be computed by integrating the probability density function $\rho: \Gamma \rightarrow[0, \infty)$ over $A$

$$
\mathbb{P}_{X}(X(\omega) \in A)=\int_{A} 1 \mathrm{~d} \mathbb{P}_{X}=\int_{A} \rho(y) \mathrm{d} y .
$$

Definition 2.6 (expectation). Let $X$ be a $\Gamma$-valued random variable with $\Gamma \subset \mathbb{R}$ from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to $\left(\Gamma, \mathcal{G}, \mathbb{P}_{X}\right)$ with probability density function $\rho: \Gamma \rightarrow \mathbb{R}$. The expectation of $X$ is defined by,

$$
\mathbb{E}(X):=\int_{\Omega} X(\omega) \mathrm{d} \mathbb{P}(\omega)=\int_{\Gamma} y \mathbb{P}_{X}(y)=\int_{\Gamma} y \rho(y) \mathrm{d} y .
$$

Definition 2.7 (variance and standard deviation). Let $\mu$ be the expectation of a real-valued random variable $X$, then the variance of $X$ is defined by

$$
\operatorname{Var}(X):=\mathbb{E}\left[(X-\mu)^{2}\right]=\mathbb{E}\left(X^{2}\right)-\mu^{2} .
$$

The standard deviation of $X$ is defined by,

$$
\sigma:=\sqrt{\operatorname{Var}(X)}
$$

Definition 2.8 (covariance). Given two jointly distributed real-valued random variables $X, Y$ on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the covariance of $X, Y$ is defined as,

$$
\operatorname{Cov}(X, Y):=\mathbb{E}\left[\left(X-\mu_{X}\right)\left(Y-\mu_{Y}\right)\right]=\mathbb{E}(X Y)-\mu_{X} \mu_{Y}
$$

where $\mu_{X}=\mathbb{E}(X)$ and $\mu_{Y}=\mathbb{E}(Y)$. Note that $\operatorname{Cov}(X, X)=\operatorname{Var}(X)$.

In the following definitions we assume $D \subset \mathbb{R}^{2}$ but these generalise to $\mathbb{R}^{n}$ with $n>2$.

Definition 2.9 (random field). Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a spatial domain $D \subset \mathbb{R}^{2}$, a random field $u(\boldsymbol{x}, \omega)$ is a random variable for each $\boldsymbol{x} \in D$. A random field can also be considered as
(i) a single function $u: D \times \Omega \rightarrow \mathbb{R}$,
(ii) a collection of realisations $u(\cdot, \omega)$ (spatial functions on $D$ ) for each $\omega \in \Omega$.

Definition 2.10 ( $L^{2}$ spaces). For a domain $D, L^{2}(D)$ is the set of square-integrable functions on $D$ defined by

$$
\begin{equation*}
L^{2}(D):=\left\{f: D \rightarrow \mathbb{R} ; \int_{D} f^{2}(\boldsymbol{x}) \mathrm{d} \boldsymbol{x}<\infty\right\} . \tag{2.8}
\end{equation*}
$$

This is a Hilbert space with the inner product,

$$
\langle f, g\rangle_{L^{2}(D)}:=\int_{D} f(\boldsymbol{x}) g(\boldsymbol{x}) \mathrm{d} \boldsymbol{x}
$$

and the norm,

$$
\|f\|_{L^{2}(D)}:=\langle f, f\rangle_{L^{2}(D)}^{1 / 2} .
$$

For a probability space $(\Omega, \mathcal{F}, \mathbb{P}), L^{2}(\Omega)$ is defined by

$$
L^{2}(\Omega):=\left\{X: \Omega \rightarrow \mathbb{R} ; \int_{\Omega}|X(\omega)|^{2} \mathrm{dP}(\omega)<\infty\right\}
$$

This is the set of random variables with finite second moment. We also define the following set of functions (random fields) on $D \times \Omega$

$$
L^{2}\left(\Omega, L^{2}(D)\right):=\left\{u: D \times \Omega \rightarrow \mathbb{R} ; \int_{\Omega}\|u(\boldsymbol{x}, \omega)\|_{L^{2}(D)}^{2} \mathrm{~d} \mathbb{P}(\omega)<\infty\right\}
$$

Definition 2.11 (second-order random field). A random field $u(\boldsymbol{x}, \omega)$ is second-order if $u(\boldsymbol{x}, \cdot) \in L^{2}(\Omega)$ for every $\boldsymbol{x} \in D$. The covariance function $C: D \times D \rightarrow \mathbb{R}$ of a second-order random field is defined by,

$$
C\left(\boldsymbol{x}_{1}, \boldsymbol{x}_{2}\right):=\operatorname{Cov}\left(u\left(\boldsymbol{x}_{1}, \omega\right), u\left(\boldsymbol{x}_{2}, \omega\right)\right)=\mathbb{E}\left(u\left(\boldsymbol{x}_{1}, \omega\right)-\mu\left(\boldsymbol{x}_{1}\right)\right) \mathbb{E}\left(u\left(\boldsymbol{x}_{2}, \omega\right)-\mu\left(\boldsymbol{x}_{2}\right)\right),
$$

where $\mu(\boldsymbol{x}):=\mathbb{E}(u(\boldsymbol{x}, \omega))$.

Definition 2.12 (integral operator). If $C \in L^{2}(D \times D)$ is a covariance function, we can define an associated integral operator $\mathcal{L}: L^{2}(D) \rightarrow L^{2}(D)$ by,

$$
\mathcal{L}(f(\boldsymbol{x})):=\int_{D} C\left(\boldsymbol{x}, \boldsymbol{x}^{*}\right) f\left(\boldsymbol{x}^{*}\right) \mathrm{d} \boldsymbol{x}^{*}
$$

Definition 2.13 (eigenpairs of an integral operator). The eigenvalues $\lambda_{i}$ and eigenfunctions $\eta_{i} \in L^{2}(D)$ of an integral operator $\mathcal{L}: L^{2}(D) \rightarrow L^{2}(D)$ are defined by,

$$
\lambda_{i} \eta_{i}(\boldsymbol{x})=\mathcal{L}\left(\eta_{i}(\boldsymbol{x})\right)
$$

Remark 2.14. Eigenvalues and eigenfunctions of integral operators associated with covariance functions satisfying Definition 2.12 have special properties by the HilbertSchmidt theorem. In particular, the eigenfunctions can be used to form an orthonormal basis for $L^{2}(D)$ and can be used to represent realisations of random fields in $L^{2}\left(\Omega, L^{2}(D)\right)$. This is called a Karhunen-Loève expansion.

Definition 2.15 (Karhunen-Loève expansion). Given a domain $D \subset \mathbb{R}^{2}$ and a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a second-order random field $u \in L^{2}\left(\Omega, L^{2}(D)\right)$ can be expressed as

$$
\begin{equation*}
u(\boldsymbol{x}, \omega)=\mu(\boldsymbol{x})+\sum_{i=1}^{\infty} \sqrt{\lambda_{i}} \gamma_{i}(\omega) \eta_{i}(\boldsymbol{x}) \tag{2.9}
\end{equation*}
$$

where $\mu(\boldsymbol{x})=\mathbb{E}(u(\boldsymbol{x}, \omega))$, the random variables $\gamma_{i}$ are defined as

$$
\begin{equation*}
\gamma_{i}(\omega):=\frac{1}{\sqrt{\lambda_{i}}}\left\langle u(\boldsymbol{x}, \omega)-\mu(\boldsymbol{x}), \eta_{i}(\boldsymbol{x})\right\rangle_{L^{2}(D)}, \tag{2.10}
\end{equation*}
$$

and have mean zero, unit variance and are pairwise uncorrelated, and $\left(\lambda_{i}, \eta_{i}(\boldsymbol{x})\right)$ are the eigenpairs of the integral operator associated with the covariance function of $u(\boldsymbol{x}, \omega)$ with the terms ordered so that $\lambda_{1} \geqslant \lambda_{2} \geqslant \cdots \geqslant 0$.

When we model the random inputs in a PDE problem, we do not have to strictly speaking use a KL expansion. We can also construct a random field $u \in L^{2}\left(\Omega, L^{2}(D)\right)$ with the same structure as a KL expansion with a chosen mean $\mu(\boldsymbol{x})$ and covariance function $C \in L^{2}(D \times D)$ as follows

$$
\begin{equation*}
u(\boldsymbol{x}, \omega)=\mu(\boldsymbol{x})+\sum_{r=1}^{\infty} \sqrt{\lambda_{r}} \eta_{r}(\boldsymbol{x}) \xi_{r}(\omega) \tag{2.11}
\end{equation*}
$$

where $\left(\lambda_{r}, \eta_{r}(\boldsymbol{x})\right)$ are eigenpairs of the covariance function $C\left(\boldsymbol{x}, \boldsymbol{x}^{*}\right)$ and we can choose the distribution of the random variables $\xi_{r}$ so that $u$ has desired properties. In computations, we can approximate it by the truncated expansion,

$$
\begin{equation*}
u_{M}(\boldsymbol{x}, \omega):=\mu(\boldsymbol{x})+\sum_{r=1}^{M} \sqrt{\lambda_{r}} \eta_{r}(\boldsymbol{x}) \xi_{r}(\omega) . \tag{2.12}
\end{equation*}
$$

One computable measure of truncation error is

$$
\begin{equation*}
\varepsilon_{M}:=\frac{\int_{D} \operatorname{Var}\left(u(\boldsymbol{x}, \omega)-u_{M}(\boldsymbol{x}, \omega)\right) \mathrm{d} \boldsymbol{x}}{\int_{D} \operatorname{Var}(u(\boldsymbol{x}, \omega)) \mathrm{d} \boldsymbol{x}} \tag{2.13}
\end{equation*}
$$

or alternatively the percentage

$$
\begin{equation*}
\beta_{M}:=\frac{\int_{D} \operatorname{Var}\left(u_{M}(\boldsymbol{x}, \omega)\right) \mathrm{d} \boldsymbol{x}}{\int_{D} \operatorname{Var}(u(\boldsymbol{x}, \omega)) \mathrm{d} \boldsymbol{x}} \times 100 \% . \tag{2.14}
\end{equation*}
$$

The parameter $\beta_{M}$ gives the proportion of the total variance retained after truncating (2.11) after $M$ terms.

Example 2.16. Let $D:=[-1,1] \times[-1,1]$ and define a random field $u(\boldsymbol{x}, \omega)$ as in (2.11) with $\mu(\boldsymbol{x})=1$, independent random variables $\xi_{r}(\omega) \sim U(-\sqrt{3}, \sqrt{3})$, and the separable exponential covariance function

$$
\begin{equation*}
C\left(\boldsymbol{x}, \boldsymbol{x}^{*}\right):=\sigma^{2} \exp \left(-\frac{\left|x_{1}-x_{1}^{*}\right|}{2}-\frac{\left|x_{2}-x_{2}^{*}\right|}{2}\right), \quad \forall \boldsymbol{x}, \boldsymbol{x}^{*} \in D, \tag{2.15}
\end{equation*}
$$

where $\sigma$ represents the standard deviation. We can write the covariance function (2.15) in the separable form,

$$
C\left(\boldsymbol{x}, \boldsymbol{x}^{*}\right)=\sigma^{2} C_{1}\left(x_{1}, x_{1}^{*}\right) C_{1}\left(x_{2}, x_{2}^{*}\right),
$$

where $C_{1}$ is defined as

$$
\begin{equation*}
C_{1}\left(x, x^{*}\right):=\exp \left(-\frac{\left|x-x^{*}\right|}{2}\right), \quad \forall x, x^{*} \in(-1,1) . \tag{2.16}
\end{equation*}
$$

The eigenpairs $\left(\theta_{i}, \chi_{i}(x)\right)$ of the integral operator associated with the covariance function $C_{1}$ (see [55, Example 7.55]) can be computed analytically. We have

$$
\theta_{i}:=\frac{1}{\alpha_{i}^{2}+1 / 4},
$$

where $\alpha_{i}$ is a root of $f_{\text {odd }}(\alpha):=\frac{1}{2}-\alpha \tan (\alpha)$ when $i$ is odd, or a root of $f_{\text {even }}(\alpha):=$ $\frac{1}{2} \tan (\alpha)+\alpha$ when $i$ is even. The eigenfunctions $\chi_{i}(x)$ are defined by,

$$
\chi_{i}(x)= \begin{cases}A_{i} \cos \left(\alpha_{i} x\right) & \text { for } i \text { odd } \\ B_{i} \sin \left(\alpha_{i} x\right) & \text { for } i \text { even }\end{cases}
$$

where $A_{i}, B_{i}$ are chosen as

$$
A_{i}=\frac{1}{\sqrt{1+\sin \left(2 \alpha_{i}\right) / 2 \alpha_{i}}}, \quad B_{i}=\frac{1}{\sqrt{1-\sin \left(2 \alpha_{i}\right) / 2 \alpha_{i}}}
$$

to have $\left\|\chi_{i}(x)\right\|_{L^{2}(-1,1)}=1$. The eigenpairs $(\lambda, \eta(\boldsymbol{x}))$ of the integral operator associated with the covariance function in (2.15) can then be expressed as

$$
\lambda=\sigma^{2} \theta_{i} \theta_{j}, \quad \eta(\boldsymbol{x})=\chi_{i}\left(x_{1}\right) \chi_{j}\left(x_{2}\right), \quad i, j \in \mathbb{N} .
$$

We can label these eigenvalues as $\lambda_{r}, r \in \mathbb{N}$ after ordering them from largest to smallest and the corresponding eigenfunctions as $\eta_{r}(\boldsymbol{x})$.

If we truncate $u(\boldsymbol{x}, \omega)$ after $M$ terms, we have

$$
\beta_{M}=\frac{\sum_{r=1}^{M} \lambda_{r}}{4 \sigma^{2}}, \quad \quad \varepsilon_{M}=1-\beta_{M}
$$

In Table 2.1 we record the smallest value of $M$ required to achieve $\varepsilon_{M} \leqslant \epsilon$ or $\beta_{M} \geqslant \vartheta$ for varying $\epsilon$ and $\vartheta$. A large number of terms are required to capture a sufficient amount of variance. For example, we need at least 270 terms to capture $99 \%$ of the variance. This is due to the slow decay of the eigenvalues associated with the covariance function in (2.15). In Figure 2.1, we plot one realisation of $u_{M}(\boldsymbol{x}, \omega)$ for $\sigma=0.085$ computed with $M=5,13,33,270$ which corresponds to retaining $80 \%, 90 \%, 95 \%$ and $99 \%$ of the variance, respectively. In Figure 2.2, we plot four different realisations of $u_{M}(\boldsymbol{x}, \omega)$ with the covariance (2.15) when we fix $M=13$.

| $\epsilon$ | 0.2 | 0.15 | 0.1 | 0.05 | 0.01 |
| :---: | :---: | :---: | :---: | :---: | :---: |
| $\vartheta$ | $80 \%$ | $85 \%$ | $90 \%$ | $95 \%$ | $99 \%$ |
| $M$ | 5 | 7 | 13 | 33 | 270 |

Table 2.1: The smallest number of terms $M$ needed to ensure $\varepsilon_{M} \leqslant \epsilon$ or $\beta_{M} \geqslant \vartheta$.


Figure 2.1: One realisation of $u_{M}(\boldsymbol{x}, \omega)$ with the covariance (2.15) when $\sigma=0.085$.


Figure 2.2: Four realisations of $u_{M}(\boldsymbol{x}, \omega)$ with the covariance (2.15) when $\sigma=0.085$ and $M=13$.

Next, we introduce a synthetic expansion of the form (2.11), which is studied in [22]. The terms in this expansion are carefully chosen so that the realisations of $u_{M}(\boldsymbol{x}, \boldsymbol{y})$ are always positive for any choice of $M$. Here, the functions $\eta_{r}(\boldsymbol{x})$ are chosen to be cosine modes.

Example 2.17. Let $D=[0,1] \times[0,1]$ and consider a random field $u(\boldsymbol{x}, \omega)$ defined as in (2.11) with $\mu(\boldsymbol{x})=1$ and independent random variables $\xi_{r}(\omega) \sim U(-1,1)$. We choose $\sqrt{\lambda_{r}}=\kappa r^{-m}$, where $\kappa=0.574$ and $m=2$ for slow decay or $\kappa=0.832$ and $m=4$ for fast decay, and

$$
\eta_{r}(\boldsymbol{x}):=\cos \left(2 \pi \varrho_{1}(r) x_{1}\right) \cos \left(2 \pi \varrho_{2}(r) x_{2}\right),
$$

with

$$
\varrho_{1}(r):=r-h(r)(h(r)+1) / 2, \quad \varrho_{2}(r):=h(r)-\varrho_{1}(r),
$$

and

$$
h(r):=\lfloor-1 / 2+\sqrt{1 / 4+2 r}\rfloor .
$$

As shown in [68], we have

$$
\int_{D} \operatorname{Var}\left(u_{M}(\boldsymbol{x}, \omega)\right) \mathrm{d} \boldsymbol{x}=\frac{\kappa^{2}}{12} \sum_{r=1}^{M} r^{-2 m},
$$

and similarly,

$$
\int_{D} \operatorname{Var}(u(\boldsymbol{x}, \omega)) \mathrm{d} \boldsymbol{x}=\frac{\kappa^{2}}{12} \sum_{r=1}^{\infty} r^{-2 m},
$$

where $\sum_{r=1}^{\infty} r^{-2 m}$ is a convergent series. The proportion $\beta_{M}$ defined in (2.14) is

$$
\beta_{M}=\frac{\sum_{r=1}^{M} r^{-2 m}}{\sum_{r=1}^{\infty} r^{-2 m}}
$$

In Table 2.2, we record the proportion $\beta_{M}$ and the relative error $\varepsilon_{M}$ for both the slow decay case and the fast decay case with $M=1, \ldots, 5$. For both cases, we need a very small number of terms to retain $99 \%$ of the variance. To achieve $\beta_{M} \geqslant 99 \%$, we need to keep at least 3 terms for the slow decay case and we need at least 1 term for the fast decay case. Note that there are many other ways to measure truncation error. For the fast decay case in Table $2.2, \beta_{M}$ is rounded up to $100 \%$ for $M=3,4,5$. In Figure 2.3, we plot the variance of $u_{M}(\boldsymbol{x}, \omega)$ with $M=3$ for both cases. We note that the maximum value of the variance of the truncated field in the fast decay case is greater than that in the slow decay case.

| $M$ | slow decay |  | fast decay |  |
| :---: | :---: | :---: | :---: | :---: |
|  | $\beta_{M}$ | $\varepsilon_{M}$ | $\beta_{M}$ | $\varepsilon_{M}$ |
| 1 | $92.39 \%$ | $7.61 \mathrm{e}-2$ | $99.59 \%$ | $4.06 \mathrm{e}-3$ |
| 2 | $98.17 \%$ | $1.83 \mathrm{e}-2$ | $99.98 \%$ | $1.70 \mathrm{e}-4$ |
| 3 | $99.31 \%$ | $6.91 \mathrm{e}-3$ | $100.00 \%$ | $1.86 \mathrm{e}-5$ |
| 4 | $99.67 \%$ | $3.30 \mathrm{e}-3$ | $100.00 \%$ | $3.42 \mathrm{e}-6$ |
| 5 | $99.82 \%$ | $1.82 \mathrm{e}-3$ | $100.00 \%$ | $8.68 \mathrm{e}-7$ |

Table 2.2: The proportion $\beta_{M}$ and the relative error $\varepsilon_{M}$ for $M=1, \ldots, 5$.


Figure 2.3: Variance of $u_{M}(\boldsymbol{x}, \omega)$ when $M=3$.


Figure 2.4: One realisation of $u_{M}(\boldsymbol{x}, \omega)$ for the slow decay case.

In Figures 2.4-2.5, we plot one realisation of $u_{M}(\boldsymbol{x}, \omega)$ with $M=1,2,3,4$. Compared with Example 2.16, we observe that the truncated field converges more quickly as $M$ increases. The KL expansion in Test problem 1 converges more slowly due to the slow decay of the eigenvalues.


Figure 2.5: One realisation of $u_{M}(\boldsymbol{x}, \omega)$ for the fast decay case.

In Figures 2.6-2.7, we plot four different realisations of $u_{M}(\boldsymbol{x}, \omega)$ with $M=5$ fixed for the slow decay case and the fast decay case.


Figure 2.6: Four realisations of $u_{M}(\boldsymbol{x}, \omega)$ for the slow decay case when $M=5$.


Figure 2.7: Four realisations of $u_{M}(\boldsymbol{x}, \omega)$ for the fast decay case when $M=5$.

### 2.2 Stochastic Galerkin approximation

In this section, we apply a stochastic Galerkin mixed finite element method (SGMFEM) to the parameter-dependent linear elasticity problem (2.1)-(2.5) and explain how the associated discrete problem can be written as a linear system.

### 2.2.1 Weak formulation

To set up the weak formulation of (2.1)-(2.5) we first define appropriate function spaces. First, we define the Hilbert space $\mathcal{H}_{0}^{1}(D)$ as

$$
\mathcal{H}_{0}^{1}(D):=\left\{\boldsymbol{v} \in\left(L^{2}(D)\right)^{2} ; \nabla \boldsymbol{v} \in\left(L^{2}(D)\right)^{2 \times 2}, \boldsymbol{v}=0 \text { on } \partial D_{D}\right\}
$$

where the spaces $\left(L^{2}(D)\right)^{2}$ and $\left(L^{2}(D)\right)^{2 \times 2}$ contain square-integrable vector-valued and matrix-valued functions, respectively. That is,

$$
\begin{aligned}
\left(L^{2}(D)\right)^{2} & :=\left\{\boldsymbol{v}: D \rightarrow \mathbb{R}^{2} ; \int_{D} \boldsymbol{v}(\boldsymbol{x}) \cdot \boldsymbol{v}(\boldsymbol{x}) \mathrm{d} \boldsymbol{x}<\infty\right\}, \\
\left(L^{2}(D)\right)^{2 \times 2} & :=\left\{A(\boldsymbol{x}): D \rightarrow \mathbb{R}^{2 \times 2} ; \int_{D} A(\boldsymbol{x}): A(\boldsymbol{x}) \mathrm{d} \boldsymbol{x}<\infty\right\} .
\end{aligned}
$$

The norm $\|\cdot\|_{1, D}$ on $\mathcal{H}^{1}(D)^{2}$ is defined by

$$
\|\boldsymbol{v}\|_{1, D}:=\left(\int_{D} \boldsymbol{v}(\boldsymbol{x}) \cdot \boldsymbol{v}(\boldsymbol{x})+\nabla \boldsymbol{v}(\boldsymbol{x}): \nabla \boldsymbol{v}(\boldsymbol{x}) \mathrm{d} \boldsymbol{x}\right)^{1 / 2} .
$$

Here, the operator : is the sum of components product. For two matrices $A:=\left[a_{i j}\right]$, $B:=\left[b_{i j}\right]$ in $\mathbb{R}^{m \times n}, A: B$ is defined by

$$
\begin{equation*}
A: B=\sum_{i=1}^{m} \sum_{j=1}^{n} a_{i j} b_{i j} . \tag{2.17}
\end{equation*}
$$

For the displacement $\boldsymbol{u}$, we choose the vector-valued solution space

$$
\mathcal{V}:=L_{\rho}^{2}\left(\Gamma, \mathcal{H}_{0}^{1}(D)\right),
$$

where

$$
L_{\rho}^{2}\left(\Gamma, \mathcal{H}_{0}^{1}(D)\right):=\left\{\boldsymbol{u}: D \times \Gamma \rightarrow \mathbb{R}^{2} ;\|\boldsymbol{u}\|_{L_{\rho}^{2}\left(\Gamma, \mathcal{H}_{0}^{1}(D)\right)}^{2}<\infty \text { and } \boldsymbol{u}=0 \text { on } \partial D_{D}\right\}
$$

and the norm $\|\cdot\|_{L_{\rho}^{2}\left(\Gamma, \mathcal{H}_{0}^{1}(D)\right)}$ is defined as

$$
\|\boldsymbol{u}\|_{L_{\rho}^{2}\left(\Gamma, \mathcal{H}_{0}^{1}(D)\right)}:=\left(\int_{\Gamma} \rho(\boldsymbol{y})\|\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{y})\|_{1, D}^{2} \mathrm{~d} \boldsymbol{y}\right)^{1 / 2}
$$

For the Hermann pressure $p$, and the auxiliary pressure $\tilde{p}$, we choose the space

$$
\mathcal{Q}:=L_{\rho}^{2}\left(\Gamma, L^{2}(D)\right),
$$

where

$$
L_{\rho}^{2}\left(\Gamma, L^{2}(D)\right):=\left\{p: D \times \Gamma \rightarrow \mathbb{R} ; \int_{\Gamma} \rho(\boldsymbol{y})\|p(\boldsymbol{x}, \boldsymbol{y})\|_{L^{2}(D)}^{2} \mathrm{~d} \boldsymbol{y}<\infty\right\}
$$

and $L^{2}(D)$ is the set of square-integrable functions on the spatial domain defined in (2.8).

The weak formulation of the parameter-dependent linear elasticity problem (2.1)(2.5) can now be written as: find $\boldsymbol{u} \in \mathcal{V}, p \in \mathcal{Q}$, and $\tilde{p} \in \mathcal{Q}$ such that

$$
\begin{align*}
a(\boldsymbol{u}, \boldsymbol{v})+b(\boldsymbol{v}, p) & =h(\boldsymbol{v}), & & \forall \boldsymbol{v} \in \mathcal{V},  \tag{2.18}\\
b(\boldsymbol{u}, q)-c(\tilde{p}, q) & =0, & & \forall q \in \mathcal{Q},  \tag{2.19}\\
-c(p, \tilde{q})+d(\tilde{p}, \tilde{q}) & =0, & & \forall \tilde{q} \in \mathcal{Q}, \tag{2.20}
\end{align*}
$$

where $a(\cdot, \cdot): \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}, b(\cdot, \cdot): \mathcal{V} \times \mathcal{Q} \rightarrow \mathbb{R}, c(\cdot, \cdot): \mathcal{Q} \times \mathcal{Q} \rightarrow \mathbb{R}$, and $d(\cdot, \cdot):$ $\mathcal{Q} \times \mathcal{Q} \rightarrow \mathbb{R}$ are bilinear forms defined by

$$
\begin{aligned}
a(\boldsymbol{u}, \boldsymbol{v}) & :=\beta \int_{\Gamma} \rho(\boldsymbol{y}) \int_{D} E(\boldsymbol{x}, \boldsymbol{y}) \boldsymbol{\epsilon}(\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{y})): \boldsymbol{\epsilon}(\boldsymbol{v}(\boldsymbol{x}, \boldsymbol{y})) \mathrm{d} \boldsymbol{x} \mathrm{~d} \boldsymbol{y} \\
b(\boldsymbol{v}, p) & :=-\int_{\Gamma} \rho(\boldsymbol{y}) \int_{D} p(\boldsymbol{x}, \boldsymbol{y}) \nabla \cdot \boldsymbol{v}(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d} \boldsymbol{x} \mathrm{~d} \boldsymbol{y} \\
c(\tilde{p}, q) & :=\tilde{\lambda}^{-1} \int_{\Gamma} \rho(\boldsymbol{y}) \int_{D} \tilde{p}(\boldsymbol{x}, \boldsymbol{y}) q(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d} \boldsymbol{x} \mathrm{~d} \boldsymbol{y} \\
d(\tilde{p}, \tilde{q}) & :=\tilde{\lambda}^{-1} \int_{\Gamma} \rho(\boldsymbol{y}) \int_{D} E(\boldsymbol{x}, \boldsymbol{y}) \tilde{p}(\boldsymbol{x}, \boldsymbol{y}) \tilde{q}(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d} \boldsymbol{x} \mathrm{~d} \boldsymbol{y} .
\end{aligned}
$$

Recall that the constants $\beta$ and $\tilde{\lambda}$ are defined in (1.32) and (1.23), and depend only on the Poisson ratio $\nu$. When $\nu \rightarrow \frac{1}{2}$, we have $\tilde{\lambda}^{-1} \rightarrow 0$. The function $h: \mathcal{V} \rightarrow \mathbb{R}$ on the right hand side in (2.18) is defined as

$$
\begin{equation*}
h(\boldsymbol{v}):=\int_{\Gamma} \rho(\boldsymbol{y}) \int_{D} \boldsymbol{f}(\boldsymbol{x}) \boldsymbol{v}(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d} \boldsymbol{x} \mathrm{~d} \boldsymbol{y} . \tag{2.21}
\end{equation*}
$$

To ensure the well-posedness of (2.18)-(2.20), we must make the following assumption concerning the Young modulus $E(\boldsymbol{x}, \boldsymbol{y})$ in (2.6).

Assumption 2.18 (Young's modulus). The Young modulus is bounded away from zero, and there exist two constants $E_{\min }, E_{\max } \in \mathbb{R}^{+}$such that

$$
\begin{equation*}
0<E_{\min } \leqslant E(\boldsymbol{x}, \boldsymbol{y}) \leqslant E_{\max }<\infty, \quad \text { a.e. in } D \times \Gamma . \tag{2.22}
\end{equation*}
$$

If Assumption 2.18 holds, it can be shown (see [44] for the proof) that there exists a unique solution $(\boldsymbol{u}, p, \tilde{p})$ to the weak formulation (2.18)-(2.20). To ensure that Assumption 2.18 holds, the parameters $y_{r}$ in (2.6) cannot take values in an unbounded set. We will choose the random variables $\xi_{r}$ to be independent and uniformly distributed with mean zero. Note that the variance of $\xi_{r} \sim U(-a, a)$ is

$$
\operatorname{Var}\left(\xi_{r}\right)=\frac{a^{2}}{3} .
$$

In Definition 2.15, the random variables in a KL expansion have mean zero and unit variance. If we model the Young modulus $E$ as a random field with the same structure as a truncated KL expansion, we can choose $\xi_{r} \sim U(-1,1)$, for $r=1,2, \ldots, M$, and $e_{r}(\boldsymbol{x})=\sqrt{3} \sqrt{\lambda_{r}} \eta_{r}(\boldsymbol{x})$ where $\left(\lambda_{r}, \eta_{r}(\boldsymbol{x})\right)$ are eigenpairs, because $\sqrt{3} \xi_{r} \sim U(-\sqrt{3}, \sqrt{3})$ has unit variance. We then have $y_{r} \in \Gamma_{r}=[-1,1]$. For this choice, the density function of $\xi_{r}$ is $\rho_{r}\left(y_{r}\right)=1 / 2$ and the joint density function is $\rho(\boldsymbol{y})=(1 / 2)^{M}$. The first term $e_{0}(\boldsymbol{x})$ in (2.6) is chosen to be the mean of $E(\boldsymbol{x}, \boldsymbol{y})$. To ensure the lower bound in (2.6) holds when $\boldsymbol{y} \in[-1,1]^{M}$, we make a further assumption.

Assumption 2.19. For $e_{0}$ appearing in (2.6), there exist two constants $e_{0}^{\min }, e_{0}^{\max } \in \mathbb{R}^{+}$ such that

$$
0<e_{0}^{\min } \leqslant e_{0}(\boldsymbol{x}) \leqslant e_{0}^{\max }<\infty, \quad \text { a.e. in } D .
$$

We model the Young modulus $E$ following Examples 2.16-2.17. In Table 2.3, we record numerical values of $E_{\min }$ and $E_{\max }$ when $M=7$ for Example 2.16 and $M=5$ for Example 2.17.

| Example | 2.16 |  | 2.17 |  |
| :---: | :---: | :---: | :---: | :---: |
| case | $\sigma=0.085$ | $\sigma=0.17$ | slow decay | fast decay |
| $E_{\min }$ | 0.6757 | 0.3514 | 0.1994 | 0.1011 |
| $E_{\max }$ | 1.3243 | 1.6486 | 1.8006 | 1.8989 |

Table 2.3: Numerical values of the minimum and maximum of the Young modulus.

### 2.2.2 Finite-dimensional problem

The solution to (2.18)-(2.20) can now be approximated by solving a discrete problem associated with finite-dimensional subspaces of $\mathcal{V}$ and $\mathcal{Q}$.

First, we define a finite-dimensional subspace $S_{d}$ of $L_{\rho}^{2}(\Gamma)$ for the parametric approximation. Specifically, we choose $S_{d}$ to be the set of multivariate polynomials on $\Gamma$ of total degree less than or equal to $d$ and construct an orthonormal basis as follows.
(i) Pick $d \in \mathbb{N}_{0}$ and define $\Lambda_{d}$ to be the set of multi-indices of length $M$ that sum to $d$ or less. That is,

$$
\Lambda_{d}:=\left\{\boldsymbol{\alpha} \in \mathbb{N}_{0}^{M} ;\|\boldsymbol{\alpha}\|_{1} \leqslant d\right\}
$$

(ii) For each multi-index $\boldsymbol{\alpha} \in \Lambda_{d}$, define the multivariate basis polynomial

$$
\psi_{\boldsymbol{\alpha}}(\boldsymbol{y}):=\prod_{r=1}^{M} L_{\alpha_{r}}\left(y_{r}\right)
$$

where $\alpha_{r}$ represents the $r$-th entry of $\boldsymbol{\alpha}$ and $L_{\alpha_{r}}:[-1,1] \rightarrow \mathbb{R}$ is a univariate Legendre polynomial of degree $\alpha_{r}$. Specifically, we select the set of Legendre polynomials that are generated by

$$
L_{n}\left(y_{r}\right):=\frac{\sqrt{2 n+1}}{2^{n} n!} \frac{\mathrm{d}^{n}}{\mathrm{~d} y_{r}^{n}}\left(y_{r}^{2}-1\right)^{n}, \quad n=0,1, \ldots
$$

The resulting Legendre polynomials

$$
L_{0}\left(y_{r}\right)=1, \quad L_{1}\left(y_{r}\right)=\sqrt{3} y_{r}, \quad L_{2}\left(y_{r}\right)=\frac{\sqrt{5}}{2}\left(3 y_{r}^{2}-1\right), \quad \ldots
$$

are othonormal with respect to the inner product $\langle\cdot, \cdot\rangle_{\rho_{r}}$ defined by

$$
\left\langle L_{i}\left(y_{r}\right), L_{j}\left(y_{r}\right)\right\rangle_{\rho_{r}}:=\int_{-1}^{1} \rho_{r} L_{i}\left(y_{r}\right) L_{j}\left(y_{r}\right) \mathrm{d} y_{r}
$$

where $\rho_{r}=1 / 2$. That is, we select polynomials that are orthonormal with respect to the probability measure associated with the distribution $U(-1,1)$.
(iii) Given the set of multivariate polynomials associated with the chosen multi-index set $\Lambda_{d}$, we can order them and then label them as $\psi_{i}(\boldsymbol{y}), i=1, \ldots, n_{y}$, where

$$
\begin{equation*}
n_{y}=\frac{(M+d)!}{M!d!} \tag{2.23}
\end{equation*}
$$

Note that when $\boldsymbol{\alpha}=\mathbf{0}$, we have $\psi_{\boldsymbol{\alpha}}(\boldsymbol{y})=1$ by definition. We will assume $\boldsymbol{\alpha}=\mathbf{0}$ is the first multi-index, so $\psi_{1}(\boldsymbol{y})=1$.

With the above construction, we have

$$
S_{d}:=\operatorname{span}\left\{\psi_{1}(\boldsymbol{y}), \psi_{2}(\boldsymbol{y}), \ldots, \psi_{n_{y}}(\boldsymbol{y})\right\} \subset L_{\rho}^{2}(\Gamma)
$$

and since we assume the joint density $\rho(\boldsymbol{y})$ is separable, the multivariate basis functions $\psi_{i}, i=1, \ldots, n_{y}$, are orthonormal with respect to the inner product $\langle\cdot, \cdot\rangle_{\rho}$ defined by

$$
\left\langle\psi_{i}(\boldsymbol{y}), \psi_{j}(\boldsymbol{y})\right\rangle_{\rho}=\int_{\Gamma} \rho(\boldsymbol{y}) \psi_{i}(\boldsymbol{y}) \psi_{j}(\boldsymbol{y}) \mathrm{d} \boldsymbol{y}
$$

For the spatial approximation, we choose $Q_{2}-Q_{1}$ finite element approximation. That is, for the displacement, we choose $\boldsymbol{V}_{h} \subset \mathcal{H}_{0}^{1}(D)$ to be the set of vector-valued functions whose components are continuous piecewise biquadratic functions defined with respect to a uniform mesh of square elements on $D$ and, for the pressures, we choose $Q_{h} \subset L^{2}(D)$ to be the set of continuous piecewise bilinear functions defined with respect to the same mesh. Other possible inf-sup stable approximation pairs are $Q_{2}-P_{-1}$ (continuous biquadratic approximation for the displacement and discontinuous linear approximation for the pressure) and $Q_{2}-P_{0}$ (continuous biquadratic approximation for the displacement and discontinuous constant approximation for the pressure) finite element approximations. In [44], $Q_{2}-P_{-1}$ approximation is used for the spatial approximation of the parameter-dependent linear elasticity problem (2.1)-(2.5). In our numerical experiments, for simplicity, we will choose $D$ to be a square so that it can be exactly partitioned into a uniform mesh of square elements of edge length $h$. We denote the grid level and partition the domain into $2^{l-1} \times 2^{l-1}$ elements, so that $h=c / 2^{l-1}$ where $c$ denotes the edge length of the square domain. However, this is not necessary and more general quadrilateral or triangular elements can also be used.

The vector-valued basis functions for $\boldsymbol{V}_{h}$ have the form

$$
\begin{aligned}
& \boldsymbol{\phi}_{i}(\boldsymbol{x})=\left[\begin{array}{c}
\phi_{i}(\boldsymbol{x}) \\
0
\end{array}\right], \quad i=1, \ldots, n_{u} \\
& \boldsymbol{\phi}_{i}(\boldsymbol{x})=\left[\begin{array}{c}
0 \\
\phi_{i-n_{u}}(\boldsymbol{x})
\end{array}\right], \quad i=n_{u}+1, \ldots, 2 n_{u}
\end{aligned}
$$

where $\left\{\phi_{i}: D \rightarrow \mathbb{R}\right\}_{i=1}^{n_{u}}$ is a set of scalar-valued continuous piecewise biquadratic basis functions defined with respect to the chosen mesh. Using the vector-valued basis functions $\boldsymbol{\phi}_{i}(\boldsymbol{x}), i=1, \ldots, 2 n_{u}$, we have

$$
\boldsymbol{V}_{h}:=\operatorname{span}\left\{\boldsymbol{\phi}_{1}(\boldsymbol{x}), \boldsymbol{\phi}_{2}(\boldsymbol{x}), \ldots, \boldsymbol{\phi}_{2 n_{u}}(\boldsymbol{x})\right\} \subset \mathcal{H}_{0}^{1}(D) .
$$

These basis functions $\phi_{i}(\boldsymbol{x})$ are associated with interior nodes only. We denote the scalar-valued continuous piecewise bilinear basis functions for $Q_{h}$ as $\left\{\varphi_{i}: D \rightarrow \mathbb{R}\right\}_{i=1}^{n_{p}}$ so that

$$
Q_{h}:=\operatorname{span}\left\{\varphi_{1}(\boldsymbol{x}), \varphi_{2}(\boldsymbol{x}), \ldots, \varphi_{n_{p}}(\boldsymbol{x})\right\} \subset L^{2}(D) .
$$

The chosen $Q_{2}-Q_{1}$ approximation is inf-sup stable, that is, there exists a constant $\gamma>0$ (the inf-sup constant) independent of the mesh parameter $h$ such that

$$
\begin{equation*}
\sup _{\boldsymbol{v} \neq \mathbf{0}, \boldsymbol{v} \in \boldsymbol{V}_{h}} \frac{\int_{D} q \nabla \cdot \boldsymbol{v} \mathrm{~d} \boldsymbol{x}}{\|\nabla \boldsymbol{v}\|_{\left(L^{2}(D)\right)^{2 \times 2}}^{2 \times 2}} \geqslant \gamma\|q\|_{L^{2}(D)}, \quad \forall q \in Q_{h} \tag{2.24}
\end{equation*}
$$

See $[2,13,14,26]$ for more details.
Using the above spaces, we now define a subspace for $L_{\rho}^{2}\left(\Gamma, \mathcal{H}_{0}^{1}(D)\right)$

$$
\begin{equation*}
\boldsymbol{V}_{h d}:=\boldsymbol{V}_{h} \otimes S_{d}=\operatorname{span}\left\{\boldsymbol{\phi}_{i}(\boldsymbol{x}) \psi_{j}(\boldsymbol{y}), i=1, \ldots, 2 n_{u}, j=1, \ldots, n_{y}\right\}, \tag{2.25}
\end{equation*}
$$

and a subspace for $L_{\rho}^{2}\left(\Gamma, L^{2}(D)\right)$

$$
\begin{equation*}
Q_{h d}:=Q_{h} \otimes S_{d}=\operatorname{span}\left\{\varphi_{i}(\boldsymbol{x}) \psi_{j}(\boldsymbol{y}), i=1, \ldots, n_{p}, j=1, \ldots, n_{y}\right\} . \tag{2.26}
\end{equation*}
$$

The finite-dimensional weak formulation then becomes: find $\boldsymbol{u}_{h d} \in V_{h d}, p_{h d} \in Q_{h d}$ and $\tilde{p}_{h d} \in Q_{h d}$ satisfying

$$
\begin{align*}
a\left(\boldsymbol{u}_{h d}, \boldsymbol{v}\right)+b\left(\boldsymbol{v}, p_{h d}\right) & =h(\boldsymbol{v}), & & \forall \boldsymbol{v} \in V_{h d},  \tag{2.27}\\
b\left(\boldsymbol{u}_{h d}, q\right)-c\left(\tilde{p}_{h d}, q\right) & =0, & & \forall q \in Q_{h d},  \tag{2.28}\\
-c\left(p_{h d}, \tilde{q}\right)+d\left(\tilde{p}_{h d}, \tilde{q}\right) & =0, & & \forall \tilde{q} \in Q_{h d} . \tag{2.29}
\end{align*}
$$

If the Young modulus $E$ satisfies Assumption 2.18, the finite-dimensional weak formulation (2.27)-(2.29) admits a unique solution. See [44] for the proof.

The solution to (2.27)-(2.29) can be expanded in the chosen bases as

$$
\begin{align*}
& \boldsymbol{u}_{h d}(\boldsymbol{x}, \boldsymbol{y})=\sum_{i=1}^{n_{y}} \sum_{j=1}^{2 n_{u}} u_{i j} \psi_{i}(\boldsymbol{y}) \phi_{j}(\boldsymbol{x}),  \tag{2.30}\\
& p_{h d}(\boldsymbol{x}, \boldsymbol{y})=\sum_{i=1}^{n_{y}} \sum_{k=1}^{n_{p}} p_{i k} \psi_{i}(\boldsymbol{y}) \varphi_{k}(\boldsymbol{x}),  \tag{2.31}\\
& \tilde{p}_{h d}(\boldsymbol{x}, \boldsymbol{y})=\sum_{i=1}^{n_{y}} \sum_{k=1}^{n_{p}} \tilde{p}_{i k} \psi_{i}(\boldsymbol{y}) \varphi_{k}(\boldsymbol{x}) . \tag{2.32}
\end{align*}
$$

To derive the associated linear system of equations, we substitute (2.30)-(2.32) into $(2.27)-(2.29)$ and then make specific choices of the test functions $(\boldsymbol{v}, q, \tilde{q})$. Since (2.27)
holds for all $\boldsymbol{v} \in V_{h d}$, we choose $\boldsymbol{v}=\psi_{l}(\boldsymbol{y}) \phi_{m}(\boldsymbol{x})$, where we first vary $m=1, \ldots, n_{u}$ for fixed $l=1, \ldots, n_{y}$ and then we vary $m=n_{u}+1, \ldots, 2 n_{u}$ for fixed $l=1, \ldots, n_{y}$. This gives $2 n_{u} n_{y}$ equations

$$
\begin{array}{r}
\sum_{i=1}^{n_{y}} \sum_{j=1}^{2 n_{u}} \beta u_{i j} \int_{\Gamma} \rho \psi_{i} \psi_{l} \int_{D} E \boldsymbol{\epsilon}\left(\boldsymbol{\phi}_{j}\right): \boldsymbol{\epsilon}\left(\boldsymbol{\phi}_{m}\right) \mathrm{d} \boldsymbol{x} \mathrm{~d} \boldsymbol{y}-\sum_{i=1}^{n_{y}} \sum_{k=1}^{n_{p}} p_{i k} \int_{\Gamma} \rho \psi_{i} \psi_{l} \int_{D} \varphi_{k} \nabla \cdot \boldsymbol{\phi}_{m} \mathrm{~d} \boldsymbol{x} \mathrm{~d} \boldsymbol{y} \\
=\int_{\Gamma} \rho \psi_{l} \int_{D} \boldsymbol{f} \boldsymbol{\phi}_{m} \mathrm{~d} \boldsymbol{x} \mathrm{~d} \boldsymbol{y} \tag{2.33}
\end{array}
$$

If we substitute (2.6) into (2.33) and separate the $\boldsymbol{x}$ and $\boldsymbol{y}$ integrals, we have

$$
\begin{array}{r}
\sum_{i=1}^{n_{y}} \sum_{j=1}^{2 n_{u}} \beta u_{i j} \int_{\Gamma} \rho \psi_{i} \psi_{l} \mathrm{~d} \boldsymbol{y} \int_{D} e_{0} \boldsymbol{\epsilon}\left(\boldsymbol{\phi}_{j}\right): \boldsymbol{\epsilon}\left(\boldsymbol{\phi}_{m}\right) \mathrm{d} \boldsymbol{x} \\
+\sum_{i=1}^{n_{y}} \sum_{j=1}^{2 n_{u}} \sum_{r=1}^{M} \beta u_{i j} \int_{\Gamma} \rho \psi_{i} \psi_{l} y_{r} \mathrm{~d} \boldsymbol{y} \int_{D} e_{r} \boldsymbol{\epsilon}\left(\boldsymbol{\phi}_{j}\right): \boldsymbol{\epsilon}\left(\boldsymbol{\phi}_{m}\right) \mathrm{d} \boldsymbol{x} \\
-\sum_{i=1}^{n_{y}} \sum_{k=1}^{n_{p}} p_{i k} \int_{\Gamma} \rho \psi_{i} \psi_{l} \mathrm{~d} \boldsymbol{y} \int_{D} \varphi_{k} \nabla \cdot \boldsymbol{\phi}_{m} \mathrm{~d} \boldsymbol{x}=\int_{\Gamma} \rho \psi_{l} \mathrm{~d} \boldsymbol{y} \int_{D} \boldsymbol{f} \boldsymbol{\phi}_{m} \mathrm{~d} \boldsymbol{x} .
\end{array}
$$

Similarly in (2.28), we choose $q=\psi_{l}(\boldsymbol{y}) \varphi_{s}(\boldsymbol{x})$, where we vary $s=1, \ldots, n_{p}$ for fixed $l=1, \ldots, n_{y}$, which gives $n_{y} n_{p}$ equations

$$
\begin{equation*}
-\sum_{i=1}^{n_{y}} \sum_{j=1}^{2 n_{u}} u_{i j} \int_{\Gamma} \rho \psi_{i} \psi_{l} \int_{D} \varphi_{s} \nabla \cdot \phi_{j} \mathrm{~d} \boldsymbol{x} \mathrm{~d} \boldsymbol{y}-\sum_{i=1}^{n_{y}} \sum_{k=1}^{n_{p}} \tilde{\lambda}^{-1} \tilde{p}_{i k} \int_{\Gamma} \rho \psi_{i} \psi_{l} \int_{D} \varphi_{k} \varphi_{s} \mathrm{~d} \boldsymbol{x} \mathrm{~d} \boldsymbol{y}=0 \tag{2.34}
\end{equation*}
$$

and in (2.29) choose $\tilde{q}=\psi_{l}(\boldsymbol{y}) \varphi_{s}(\boldsymbol{x})$ in the same order, which gives $n_{y} n_{p}$ equations

$$
\begin{equation*}
-\sum_{i=1}^{n_{y}} \sum_{k=1}^{n_{p}} \tilde{\lambda}^{-1} p_{i k} \int_{\Gamma} \rho \psi_{i} \psi_{l} \int_{D} \varphi_{k} \varphi_{s} \mathrm{~d} \boldsymbol{x} \mathrm{~d} \boldsymbol{y}+\sum_{i=1}^{n_{y}} \sum_{k=1}^{n_{p}} \tilde{\lambda}^{-1} \tilde{p}_{i k} \int_{\Gamma} \rho \psi_{i} \psi_{l} \int_{D} E \varphi_{k} \varphi_{s} \mathrm{~d} \boldsymbol{x} \mathrm{~d} \boldsymbol{y}=0 \tag{2.35}
\end{equation*}
$$

Substituting (2.6) into (2.35) gives

$$
\begin{array}{r}
-\sum_{i=1}^{n_{y}} \sum_{k=1}^{n_{p}} \tilde{\lambda}^{-1} p_{i k} \int_{\Gamma} \rho \psi_{i} \psi_{l} \mathrm{~d} \boldsymbol{y} \int_{D} \varphi_{k} \varphi_{s} \mathrm{~d} \boldsymbol{x}+\sum_{i=1}^{n_{y}} \sum_{k=1}^{n_{p}} \tilde{\lambda}^{-1} \tilde{p}_{i k} \int_{\Gamma} \rho \psi_{i} \psi_{l} \mathrm{~d} \boldsymbol{y} \int_{D} e_{0} \varphi_{k} \varphi_{s} \mathrm{~d} \boldsymbol{x} \\
+\sum_{i=1}^{n_{y}} \sum_{k=1}^{n_{p}} \sum_{r=1}^{M} \tilde{\lambda}^{-1} \tilde{p}_{i k} \int_{\Gamma} \rho \psi_{i} \psi_{l} \mathrm{~d} \boldsymbol{y} \int_{D} e_{r} \varphi_{k} \varphi_{s} \mathrm{~d} \boldsymbol{x}=0
\end{array}
$$

If we swap the order of the equations (2.34) and (2.35), and store the coefficients
in the expansions (2.30)-(2.32) in a vector in the following order

$$
\mathbf{x}:=\left[\begin{array}{c}
\mathbf{u}_{1} \\
\mathbf{u}_{2} \\
\widetilde{\mathbf{p}} \\
\mathbf{p}
\end{array}\right] \in \mathbb{R}^{n_{x} n_{y}},
$$

where $\mathbf{u}_{1}, \mathbf{u}_{2} \in \mathbb{R}^{n_{u} n_{y}}$ and $\widetilde{\mathbf{p}}, \mathbf{p} \in \mathbb{R}^{n_{p} n_{y}}$ are defined by

$$
\mathbf{u}_{1}:=\left[\begin{array}{c}
\mathbf{u}_{1,1} \\
\mathbf{u}_{1,2} \\
\vdots \\
\mathbf{u}_{1, n_{y}}
\end{array}\right], \quad \mathbf{u}_{2}:=\left[\begin{array}{c}
\mathbf{u}_{2,1} \\
\mathbf{u}_{2,2} \\
\vdots \\
\mathbf{u}_{2, n_{y}}
\end{array}\right], \quad \widetilde{\mathbf{p}}:=\left[\begin{array}{c}
\tilde{\mathbf{p}}_{1} \\
\widetilde{\mathbf{p}}_{2} \\
\vdots \\
\widetilde{\mathbf{p}}_{n_{y}}
\end{array}\right], \quad \mathbf{p}:=\left[\begin{array}{c}
\mathbf{p}_{1} \\
\mathbf{p}_{2} \\
\vdots \\
\mathbf{p}_{n_{y}}
\end{array}\right],
$$

with the ordering

$$
\begin{aligned}
\mathbf{u}_{1, i} & :=\left(u_{i 1}, u_{i 2}, \ldots, u_{i n_{u}}\right)^{\top} \in \mathbb{R}^{n_{u}}, & i & =1,2, \ldots, n_{y} \\
\mathbf{u}_{2, i} & :=\left(u_{i, n_{u}+1}, u_{i, n_{u}+2}, \ldots, u_{i, 2 n_{u}}\right)^{\top} \in \mathbb{R}^{n_{u}}, & & i=1,2, \ldots, n_{y} \\
\widetilde{\mathbf{p}}_{i} & :=\left(\tilde{p}_{i 1}, \tilde{p}_{i 2}, \ldots, \tilde{p}_{i n_{p}}\right)^{\top} \in \mathbb{R}^{n_{p}}, & i & =1,2, \ldots, n_{y} \\
\mathbf{p}_{i} & :=\left(p_{i 1}, \ldots, p_{i n_{p}}\right)^{\top} \in \mathbb{R}^{n_{p}}, & i & =1,2, \ldots, n_{y}
\end{aligned}
$$

then the discrete problem can be written as a saddle point system of the form

$$
\left[\begin{array}{cc}
\mathcal{A} & \mathcal{B}^{\top}  \tag{2.36}\\
\mathcal{B} & 0
\end{array}\right]\left[\begin{array}{l}
\mathbf{v} \\
\mathbf{p}
\end{array}\right]=\left[\begin{array}{l}
\mathbf{b} \\
\mathbf{0}
\end{array}\right]
$$

where

$$
\mathbf{v}:=\left[\begin{array}{c}
\mathbf{u}_{1} \\
\mathbf{u}_{2} \\
\widetilde{\mathbf{p}}
\end{array}\right] .
$$

The matrix $\mathcal{A}$ has the block structure

$$
\mathcal{A}=\left[\begin{array}{cc}
\mathbb{A} & 0  \tag{2.37}\\
0 & \mathbb{D}
\end{array}\right] \in \mathbb{R}^{\left(2 n_{u}+n_{p}\right) n_{y} \times\left(2 n_{u}+n_{p}\right) n_{y}},
$$

where $\mathbb{A} \in \mathbb{R}^{2 n_{u} n_{y} \times 2 n_{u} n_{y}}$ and $\mathbb{D} \in \mathbb{R}^{n_{p} n_{y} \times n_{p} n_{y}}$ have the Kronecker product structure

$$
\begin{aligned}
& \mathbb{A}:=\left[\begin{array}{cc}
\beta \sum_{r=0}^{M} G_{r} \otimes A_{r}^{(11)} & \beta \sum_{r=0}^{M} G_{r} \otimes A_{r}^{(21)} \\
\beta \sum_{r=0}^{M} G_{r} \otimes A_{r}^{(12)} & \beta \sum_{r=0}^{M} G_{r} \otimes A_{r}^{(22)}
\end{array}\right], \\
& \mathbb{D}:=\tilde{\lambda}^{-1} \sum_{r=0}^{M} G_{r} \otimes D_{r} .
\end{aligned}
$$

The matrix $\mathcal{B}$ has the structure

$$
\mathcal{B}=\left[\begin{array}{ll|l}
G_{0} \otimes B_{1} & G_{0} \otimes B_{2} & -\tilde{\lambda}^{-1} G_{0} \otimes C \tag{2.38}
\end{array}\right] \in \mathbb{R}^{n_{p} n_{y} \times\left(2 n_{u}+n_{p}\right) n_{y}}
$$

The matrices $G_{r} \in \mathbb{R}^{n_{y} \times n_{y}}$, for $r=0,1, \ldots, M$, are associated with the parametric polynomial approximation. The entries of $G_{0}$ are

$$
\begin{equation*}
G_{0}(i, l)=\left\langle\psi_{i}, \psi_{l}\right\rangle_{\rho}, \quad i, l=1, \ldots, n_{y} . \tag{2.39}
\end{equation*}
$$

Since $\left\{\psi_{i}\right\}$ is chosen to be an orthonormal basis, all of the diagonal entries of $G_{0}$ are 1 and the rest are zero. That is, $G_{0}$ is the $n_{y} \times n_{y}$ identity matrix. For $r=1, \ldots, M$, the entries of $G_{r}$ are

$$
\begin{equation*}
G_{r}(i, l)=\left\langle y_{r} \psi_{i}, \psi_{l}\right\rangle_{\rho}, \quad i, l=1, \ldots, n_{y} \tag{2.40}
\end{equation*}
$$

Clearly, each $G_{r}$ is symmetric and it can be shown (see [63] and [28]) that each has at most two nonzero entries per row. That is, they are highly sparse.

The matrices $\mathbb{A}, B_{1}, B_{2}, C$, and $D_{r}$, for $r=0,1, \ldots, M$, are associated with the spatial approximation. If we define $\boldsymbol{x}:=\left(x_{1}, x_{2}\right)^{\top}$, the entries of $A_{r}^{(11)}, A_{r}^{(21)}, A_{r}^{(12)}$ and $A_{r}^{(22)}$, for $r=0,1, \ldots, M$, are defined as

$$
\begin{aligned}
A_{r}^{(11)}(j, m) & :=\int_{D} e_{r}(\boldsymbol{x})\left(\frac{\partial \phi_{j}}{\partial x_{1}} \frac{\partial \phi_{m}}{\partial x_{1}}+\frac{1}{2} \frac{\partial \phi_{j}}{\partial x_{2}} \frac{\partial \phi_{m}}{\partial x_{2}}\right) \mathrm{d} \boldsymbol{x}, \quad j, m=1, \ldots, n_{u}, \\
A_{r}^{(21)}(j, m) & :=\frac{1}{2} \int_{D} e_{r}(\boldsymbol{x}) \frac{\partial \phi_{j}}{\partial x_{2}} \frac{\partial \phi_{m}}{\partial x_{1}} \mathrm{~d} \boldsymbol{x}, \quad j, m=1, \ldots, n_{u}, \\
A_{r}^{(12)}(j, m) & :=\frac{1}{2} \int_{D} e_{r}(\boldsymbol{x}) \frac{\partial \phi_{j}}{\partial x_{1}} \frac{\partial \phi_{m}}{\partial x_{2}} \mathrm{~d} \boldsymbol{x}, \quad j, m=1, \ldots, n_{u}, \\
A_{r}^{(22)}(j, m) & :=\int_{D} e_{r}(\boldsymbol{x})\left(\frac{1}{2} \frac{\partial \phi_{j}}{\partial x_{1}} \frac{\partial \phi_{m}}{\partial x_{1}}+\frac{\partial \phi_{j}}{\partial x_{2}} \frac{\partial \phi_{m}}{\partial x_{2}}\right) \mathrm{d} \boldsymbol{x}, \quad j, m=1, \ldots, n_{u} .
\end{aligned}
$$

The entries of $B_{1}$ and $B_{2}$ are

$$
\begin{array}{ll}
B_{1}(k, m):=-\int_{D} \varphi_{k} \frac{\partial \phi_{m}}{\partial x_{1}} \mathrm{~d} \boldsymbol{x}, & k=1, \ldots, n_{p}, m=1, \ldots, n_{u}, \\
B_{2}(k, m):=-\int_{D} \varphi_{k} \frac{\partial \phi_{m}}{\partial x_{2}} \mathrm{~d} \boldsymbol{x}, & k=1, \ldots, n_{p}, m=1, \ldots, n_{u} .
\end{array}
$$

The entries of $C$ are

$$
\begin{equation*}
C(k, s)=\int_{D} \varphi_{k}(\boldsymbol{x}) \varphi_{s}(\boldsymbol{x}) \mathrm{d} \boldsymbol{x}, \quad k, s=1, \ldots, n_{p} \tag{2.41}
\end{equation*}
$$

The entries of $D_{r}, r=0,1, \ldots, M$, are

$$
\begin{equation*}
D_{r}(k, s)=\int_{D} e_{r}(\boldsymbol{x}) \varphi_{k}(\boldsymbol{x}) \varphi_{s}(\boldsymbol{x}) \mathrm{d} \boldsymbol{x}, \quad k, s=1, \ldots, n_{p} \tag{2.42}
\end{equation*}
$$

Note that $C$ and $D_{r}, r=0,1, \ldots, M$ are symmetric matrices. $C$ is a standard finite element mass matrix and the $D_{r}$ are weighted mass matrices. Since they are finite element matrices, $A_{r}^{(11)}, A_{r}^{(21)}, A_{r}^{(12)}, A_{r}^{(22)}, B_{1}, B_{2}, C$ and $D_{r}$ are all sparse.

Proposition 2.20. If Assumption 2.18 holds, the matrix $\mathcal{A}$ defined in (2.37) is symmetric and positive definite.

Proof. Since the matrices $G_{r}, A_{r}^{(11)}, A_{r}^{(22)}$, and $D_{r}$ for $r=0,1, \ldots, M$, are symmetric, and $A_{r}^{(12)}=\left(A_{r}^{(21)}\right)^{\top}$, the matrix $\mathcal{A}$ is clearly symmetric. For a vector $\mathbf{w} \in \mathbb{R}^{2 n_{u} n_{y}} \backslash\{\mathbf{0}\}$ with entries $w_{i j}$ for $i=1,2, \ldots, n_{y}, j=1,2, \ldots, 2 n_{u}$, there is a corresponding function $\boldsymbol{w}_{h d} \in \boldsymbol{V}_{h d}$ defined by

$$
\boldsymbol{w}_{h d}(\boldsymbol{x}, \boldsymbol{y})=\sum_{i=1}^{n_{y}} \sum_{j=1}^{2 n_{u}} w_{i j} \psi_{i}(\boldsymbol{y}) \phi_{j}(\boldsymbol{x}) .
$$

If Assumption 2.18 holds, $E(\boldsymbol{x}, \boldsymbol{y}) \geqslant E_{\min }>0$ and we have

$$
\begin{aligned}
\mathbf{w}^{\top} \mathbb{A} \mathbf{w}=a(\boldsymbol{w}, \boldsymbol{w}) & =\beta \int_{\Gamma} \rho(\boldsymbol{y}) \int_{D} E(\boldsymbol{x}, \boldsymbol{y}) \boldsymbol{\epsilon}\left(\boldsymbol{w}_{h d}(\boldsymbol{x}, \boldsymbol{y})\right): \boldsymbol{\epsilon}\left(\boldsymbol{w}_{h d}(\boldsymbol{x}, \boldsymbol{y})\right) \mathrm{d} \boldsymbol{x} \mathrm{~d} \boldsymbol{y} \\
& \geqslant \beta E_{\min }\left\|\boldsymbol{\epsilon}\left(\boldsymbol{w}_{h d}(\boldsymbol{x}, \boldsymbol{y})\right)\right\|_{L_{\rho}\left(\Gamma,\left(L^{2}(D)\right)^{2 \times 2}\right)}^{2}>0 .
\end{aligned}
$$

Similarly, for a vector $\mathbf{q} \in \mathbb{R}^{n_{p} n_{y}} \backslash\{\mathbf{0}\}$ with entries $q_{i j}$ for $i=1,2, \ldots, n_{y}, j=1,2, \ldots, n_{p}$, there is a function $\boldsymbol{q}_{h d} \in Q_{h d}$ defined by

$$
\boldsymbol{q}_{h d}(\boldsymbol{x}, \boldsymbol{y})=\sum_{i=1}^{n_{y}} \sum_{j=1}^{n_{p}} q_{i j} \psi_{i}(\boldsymbol{y}) \varphi_{j}(\boldsymbol{x})
$$

and we have

$$
\begin{aligned}
\mathbf{q}^{\top} \mathbb{D} \mathbf{q} & =\tilde{\lambda}^{-1} \int_{\Gamma} \rho(\boldsymbol{y}) \int_{D} E(\boldsymbol{x}, \boldsymbol{y}) q_{h d}(\boldsymbol{x}, \boldsymbol{y}) q_{h d}(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d} \boldsymbol{x} \mathrm{~d} \boldsymbol{y} \\
& \geqslant \tilde{\lambda}^{-1} E_{\min }\left\|q_{h d}\right\|_{L_{\rho}\left(\Gamma, L^{2}(D)\right)}^{2}>0 .
\end{aligned}
$$

The matrices $\mathbb{A}$ and $\mathbb{D}$ are positive definite, and hence so is $\mathcal{A}$.

The following result from [67] can be used to establish eigenvalue bounds for the coefficient matrix of (2.36). Note that $\mathcal{B}$ is of full row rank if the inf-sup condition in (2.24) is satisfied and $\boldsymbol{V}_{h d}$ and $Q_{h d}$ are constructed as in (2.25)-(2.26).

Theorem 2.21. Let $0<\mu_{\min }, \mu_{\max }$ be the minimum and maximum eigenvalues of the symmetric and positive definite matrix $\mathcal{A}$, and let $\sigma_{\min }, \sigma_{\max }$ be the minimum and
maximum singular values of the matrix $\mathcal{B}$ with full row rank. The eigenvalues of the coefficient matrix of (2.36) lie in the union of the intervals

$$
\begin{array}{r}
{\left[\frac{1}{2}\left(\mu_{\min }-\sqrt{\mu_{\min }^{2}+4 \sigma_{\max }^{2}}\right), \frac{1}{2}\left(\mu_{\max }-\sqrt{\mu_{\max }^{2}+4 \sigma_{\min }^{2}}\right)\right] \cup} \\
{\left[\mu_{\min }, \frac{1}{2}\left(\mu_{\max }+\sqrt{\mu_{\max }^{2}+4 \sigma_{\max }^{2}}\right)\right] .}
\end{array}
$$

The nonzero part of the right hand side of the linear system (2.36) is associated with the body force $\boldsymbol{f}(\boldsymbol{x})$, where $\boldsymbol{f}(\boldsymbol{x}):=\left(f_{1}(\boldsymbol{x}), f_{2}(\boldsymbol{x})\right)^{\top}$. The vector $\mathbf{b} \in \mathbb{R}^{\left(2 n_{u}+n_{p}\right) n_{y}}$ can be written as

$$
\mathbf{b}:=\left[\begin{array}{c}
\mathbf{b}_{1} \\
\mathbf{b}_{2} \\
\mathbf{0}
\end{array}\right]
$$

where $\mathbf{b}_{1}=\mathbf{g}_{0} \otimes \mathbf{f}_{1}$ and $\mathbf{b}_{2}=\mathbf{g}_{0} \otimes \mathbf{f}_{2}$ where

$$
\begin{array}{ll}
\mathbf{f}_{1}(m)=\int_{D} f_{1}(\boldsymbol{x}) \phi_{m}(\boldsymbol{x}) \mathrm{d} \boldsymbol{x}, & m=1, \ldots, n_{u} \\
\mathbf{f}_{2}(m)=\int_{D} f_{2}(\boldsymbol{x}) \phi_{m}(\boldsymbol{x}) \mathrm{d} \boldsymbol{x}, & m=1, \ldots, n_{u}
\end{array}
$$

If we assume that $\psi_{1}=1$, then the entries of $\mathbf{g}_{0}$ are

$$
\mathbf{g}_{0}(l)=\left\langle 1, \psi_{l}\right\rangle_{\rho}= \begin{cases}1, & l=1  \tag{2.43}\\ 0, & l=2, \ldots, n_{y}\end{cases}
$$

which is the first column of the $n_{y} \times n_{y}$ identity matrix.
As mentioned in Section 1.1, the saddle point system (2.36) can also be rewritten as

$$
\begin{equation*}
\left(\sum_{r=0}^{M} G_{r} \otimes K_{r}\right) \mathbf{x}=\mathbf{h} \tag{2.44}
\end{equation*}
$$

where the matrices $K_{r}$, for $r=0,1, \ldots, M$ are associated with the spatial approximation. In this formulation, $K_{0} \in \mathbb{R}^{n_{x} \times n_{x}}$ where $n_{x}:=2\left(n_{u}+n_{p}\right)$, is defined as

$$
K_{0}:=\left[\begin{array}{ccc}
\beta A_{0} & 0 & B^{\top}  \tag{2.45}\\
0 & \tilde{\lambda}^{-1} D_{0} & -\tilde{\lambda}^{-1} C \\
B & -\tilde{\lambda}^{-1} C & 0
\end{array}\right]
$$

and for $r=1, \ldots, M$, the matrices $K_{r} \in \mathbb{R}^{n_{x} \times n_{x}}$ are defined as

$$
K_{r}:=\left[\begin{array}{ccc}
\beta A_{r} & 0 & 0  \tag{2.46}\\
0 & \tilde{\lambda}^{-1} D_{r} & 0 \\
0 & 0 & 0
\end{array}\right]
$$

where the matrices $A_{r}$, for $r=0,1, \ldots, M$, and $B$ are defined by

$$
A_{r}:=\left[\begin{array}{ll}
A_{r}^{(11)} & A_{r}^{(21)} \\
A_{r}^{(12)} & A_{r}^{(22)}
\end{array}\right], \quad B:=\left[\begin{array}{ll}
B_{1} & B_{2}
\end{array}\right] .
$$

The vector $\mathbf{x}$ in (2.44) is defined as

$$
\mathbf{x}:=\left[\begin{array}{c}
\mathbf{x}_{1} \\
\mathbf{x}_{2} \\
\vdots \\
\mathbf{x}_{n_{y}}
\end{array}\right], \quad \mathbf{x}_{i}:=\left[\begin{array}{c}
\mathbf{u}_{1, i} \\
\mathbf{u}_{2, i} \\
\widetilde{\mathbf{p}}_{i} \\
\mathbf{p}_{i}
\end{array}\right] \in \mathbb{R}^{n_{x}} .
$$

In addition, the vector $\mathbf{h}$ is defined as

$$
\mathbf{h}:=\mathbf{g}_{0} \otimes \mathbf{f},
$$

where $\mathbf{f}$ is defined as

$$
\mathbf{f}:=\left[\begin{array}{c}
\mathbf{f}_{1}  \tag{2.47}\\
\mathbf{f}_{2} \\
0 \\
0
\end{array}\right] .
$$

For simplicity, we denote the coefficient matrix of (2.44) as $\mathcal{M}$. Then, the linear system can simply be written as

$$
\begin{equation*}
\mathcal{M} \mathrm{x}=\mathrm{h} . \tag{2.48}
\end{equation*}
$$

Since $\mathcal{M}$ is symmetric and indefinite, we can approximate the solution to this linear system using the minimal residual method (MINRES, see [59]). We discuss this next.

### 2.3 Preconditioned MINRES

MINRES is a commonly used Krylov subspace method for solving Hermitian linear systems. Krylov subspace methods are a family of iterative methods that generate
approximations $\mathbf{x}_{k}$ in affine subspaces of the form $\mathbf{x}_{0}+\mathcal{K}_{k}$, where $\mathbf{x}_{0}$ is the chosen initial guess and $\mathcal{K}_{k}$ is a so-called Krylov subspace. See [34, 71] for more details.

For the linear system (2.48) with an initial guess $\mathbf{x}_{0}$, the associated residual is defined by

$$
\mathbf{r}_{0}:=\mathbf{h}-\mathcal{M} \mathbf{x}_{0} .
$$

The Krylov subspace $\mathcal{K}_{k}\left(\mathcal{M}, \mathbf{r}_{0}\right)$ of order $k$ associated with the matrix $\mathcal{M}$ and the residual $\mathbf{r}_{0}$ is defined as

$$
\mathcal{K}_{k}\left(\mathcal{M}, \mathbf{r}_{0}\right):=\operatorname{span}\left\{\mathbf{r}_{0}, \mathcal{M} \mathbf{r}_{0}, \ldots, \mathcal{M}^{k-1} \mathbf{r}_{0}\right\}
$$

At the $k$ th iteration, a Krylov subspace method generates an approximation $\mathbf{x}_{k}$ in the space $\mathbf{x}_{0}+\mathcal{K}_{k}\left(\mathcal{M}, \mathbf{r}_{0}\right)$. The associated residual

$$
\mathbf{r}_{k}:=\mathbf{h}-\mathcal{M} \mathbf{x}_{k},
$$

belongs to the space $\mathcal{R}_{k}:=\mathbf{r}_{0}+\mathcal{M} \mathcal{K}_{k}\left(\mathcal{M}, \mathbf{r}_{0}\right)$. MINRES constructs an approximation $\mathbf{x}_{k}$ in such a way that the 2 -norm of the residual $\mathbf{r}_{k}$ is minimised over the space $\mathcal{R}_{k}$. Pseudocode for MINRES can be found in many standard texts on iterative methods, for example in [71]. The usual stopping criterion chosen for MINRES is

$$
\begin{equation*}
\frac{\left\|\mathbf{r}_{k}\right\|_{2}}{\left\|\mathbf{r}_{0}\right\|_{2}}<\epsilon \tag{2.49}
\end{equation*}
$$

where $\epsilon$ is a tolerance. It is well known (see $[26,34]$ ) that a theoretical upper bound for the relative residual error at the $k$-th step is given by

$$
\frac{\left\|\mathbf{r}_{k}\right\|_{2}}{\left\|\mathbf{r}_{0}\right\|_{2}} \leqslant \min _{p_{k} \in \Pi_{k}, p_{k}(0)=1} \max _{j}\left|p_{k}\left(\lambda_{j}\right)\right|,
$$

where $\Pi_{k}$ is the set of polynomials of degree less than or equal to $k$ and $\lambda_{j}$ is an eigenvalue of $\mathcal{M}$. If the eigenvalues of $\mathcal{M}$ are spread out, MINRES will require a large number of iterations to terminate. To fix this, we can use preconditioned MINRES.

Preconditioned MINRES requires a symmetric and positive definite preconditioner $\mathcal{P}$ to preserve the symmetry of the coefficient matrix. Suppose a symmetric and positive definite preconditioner $P$ is available. Then, we can write its Cholesky factorization as

$$
\begin{equation*}
\mathcal{P}=L L^{\top}, \tag{2.50}
\end{equation*}
$$

where $L \in \mathbb{R}^{n_{x} n_{y} \times n_{x} n_{y}}$ for our problem. If we apply $L^{-1}$ to the left on both sides of (2.48) and replace the solution vector $\mathbf{x}$ with $\mathbf{y}=L^{\top} \mathbf{x}$, then we obtain the symmetrically preconditioned linear system

$$
\begin{equation*}
L^{-1} \mathcal{M} L^{-\top} \mathbf{y}=L^{-1} \mathbf{h} \tag{2.51}
\end{equation*}
$$

The residual of the symmetrically preconditioned linear system at the $k$ th iteration is

$$
\mathbf{r}_{k}^{p}:=L^{-1}\left(\mathbf{h}-\mathcal{M} \mathbf{x}_{k}\right)=L^{-1} \mathbf{r}_{k}
$$

where $\mathbf{r}_{k}$ is the $k$ th residual of the unpreconditioned linear system (2.48). Given a tolerance $\epsilon$, the usual stopping criterion chosen for preconditioned MINRES is

$$
\begin{equation*}
\frac{\left\|\mathbf{r}_{k}\right\|_{\mathcal{P}^{-1}}}{\left\|\mathbf{r}_{0}\right\|_{\mathcal{P}^{-1}}}<\epsilon \tag{2.52}
\end{equation*}
$$

The theoretical upper bound for the preconditioned relative residual error at the $k$-th step is given by

$$
\frac{\left\|\mathbf{r}_{k}\right\|_{\mathcal{P}-1}}{\left\|\mathbf{r}_{0}\right\|_{\mathcal{P}-1}} \leqslant \min _{p_{k} \in \Pi_{k}, p_{k}(0)=1} \max _{j}\left|p_{k}\left(\lambda_{j}\right)\right|
$$

where $\Pi_{k}$ is the set of polynomials of degree less than or equal to $k$ as before and $\lambda_{j}$ now represents an eigenvalue of $\mathcal{P}^{-1} \mathcal{M}$. A good preconditioner $\mathcal{P}$ should make the eigenvalues of $\mathcal{P}^{-1} \mathcal{M}$ clustered so that the number of required iterations is reduced.

### 2.3.1 Preconditioning

We will discuss two symmetric and positive definite preconditioners for the linear system (2.48) for use with MINRES. A good preconditioner $\mathcal{P}$ generally satisfies two conditions:
(i) the action of the inverse of $\mathcal{P}$ on a vector is cheap to compute compared to that of $\mathcal{M}$;
(ii) the eigenvalues of $\mathcal{P}^{-1} \mathcal{M}$ are clustered.

For stochastic Galerkin finite element systems associated with scalar elliptic PDEs

$$
\begin{equation*}
-\nabla \cdot(a \nabla u)=f, \tag{2.53}
\end{equation*}
$$

with parameter-dependent diffusion coefficient $a$, the coefficient matrices are symmetric and positive definite. In that case, one can choose the block-diagonal preconditioner $\mathcal{P}=I_{n_{y}} \otimes K_{0}$ since $K_{0}$ is also symmetric and positive definite. This approach has been studied in [30,63]. For the parameter-dependent linear elasticity problem considered here, $K_{0}$ is symmetric but indefinite, and so $I_{n_{y}} \otimes K_{0}$ is also indefinite. We consider two symmetric and positive definite approximations to $I_{n_{y}} \otimes K_{0}$ as preconditioners.
(i) The matrix $K_{0}$ defined in (2.45) has a saddle point structure, and a general approximation to it can be written as

$$
K_{S}:=\left[\begin{array}{ccc}
\beta A_{0} & 0 & 0  \tag{2.54}\\
0 & \tilde{\lambda}^{-1} D_{0} & 0 \\
0 & 0 & S_{0}
\end{array}\right]
$$

where

$$
\begin{equation*}
S_{0}:=\beta^{-1} B A_{0}^{-1} B^{\top}+\tilde{\lambda}^{-1} C D_{0}^{-1} C \tag{2.55}
\end{equation*}
$$

is the Schur complement of $K_{0}$. This leads to the first preconditioner $\mathcal{P}_{1}:=I \otimes K_{S}$.
(ii) To make the action of the inverse of $\mathcal{P}$ on a vector cheaper to compute, we can find an approximation to $K_{S}$. In [44], a block diagonal matrix is used to approximate $A_{0}$, and an approximation of the Schur complement is used instead of the dense matrix $S_{0}$. The approximation to $A_{0}$ is chosen to be

$$
A_{0, \text { approx }}:=\left[\begin{array}{cc}
\frac{2}{3}\left(A_{0}^{(11)}+A_{0}^{(22)}\right) & 0 \\
0 & \frac{2}{3}\left(A_{0}^{(11)}+A_{0}^{(22)}\right)
\end{array}\right],
$$

and the approximation to the Schur complement $S_{0}$ is chosen as

$$
\begin{equation*}
S_{0, \text { approx }}:=\left(\beta^{-1}+\tilde{\lambda}^{-1}\right) C . \tag{2.56}
\end{equation*}
$$

When the expectation $e_{0}$ of the Young modulus is a constant, we can instead choose

$$
\begin{equation*}
S_{0, \text { approx }}:=\left(\beta^{-1}+\tilde{\lambda}^{-1}\right) e_{0}^{-1} C \tag{2.57}
\end{equation*}
$$

The approximation to $K_{S}$ can be written as

$$
K_{S, \text { approx }}:=\left[\begin{array}{ccc}
\beta A_{0, \text { approx }} & 0 & 0  \tag{2.58}\\
0 & \tilde{\lambda}^{-1} D_{0} & 0 \\
0 & 0 & S_{0, \text { approx }}
\end{array}\right]
$$

and the second preconditioner we choose is $\mathcal{P}_{2}:=I \otimes K_{S, \text { approx }}$.

Next, we will discuss eigenvalue bounds for the preconditioned matrices $\mathcal{P}_{1}^{-1} \mathcal{M}$ and $\mathcal{P}_{2}^{-1} \mathcal{M}$. Since $\mathcal{M}$ can be obtained by permuting the rows and columns of the coefficient matrix in (2.36), they are similar and have the same eigenvalues (see [37]). Let $\mathcal{M}_{p}$ denote the coefficient matrix

$$
\mathcal{M}_{p}:=\left[\begin{array}{cc}
\mathcal{A} & \mathcal{B}^{\top} \\
\mathcal{B} & 0
\end{array}\right]
$$

where $\mathcal{A}$ and $\mathcal{B}$ are defined in (2.37) and (2.38). The preconditioner $\mathcal{P}_{1}$ is similar to the following matrix

$$
\mathcal{Q}_{1}:=\left[\begin{array}{cc}
\mathcal{A}_{0} & 0 \\
0 & \mathcal{S}_{0}
\end{array}\right]
$$

where $\mathcal{A}_{0}$ is defined as

$$
\mathcal{A}_{0}:=\left[\begin{array}{cc}
\mathbb{A}_{0} & 0  \tag{2.59}\\
0 & \mathbb{D}_{0}
\end{array}\right]
$$

with $\mathbb{A}_{0}$ and $\mathbb{D}_{0}$ defined by

$$
\mathbb{A}_{0}:=\left[\begin{array}{ll}
\beta I_{n_{y}} \otimes A_{0}^{(11)} & \beta I_{n_{y}} \otimes A_{0}^{(21)} \\
\beta I_{n_{y}} \otimes A_{0}^{(12)} & \beta I_{n_{y}} \otimes A_{0}^{(22)}
\end{array}\right], \quad \mathbb{D}_{0}:=\tilde{\lambda}^{-1} I_{n_{y}} \otimes D_{0}
$$

and $\mathcal{S}_{0}$ is defined as

$$
\mathcal{S}_{0}:=\mathcal{B}_{p} \mathcal{A}_{0}^{-1} \mathcal{B}_{p}^{\top}=I_{n_{y}} \otimes S_{0},
$$

where $S_{0}$ is defined in (2.55). If we factorize $\mathcal{Q}_{1}=L_{\mathcal{Q}_{1}} L_{\mathcal{Q}_{1}}^{\top}$ with

$$
L_{\mathcal{Q}_{1}}:=\left[\begin{array}{cc}
L_{\mathcal{A}_{0}} & 0 \\
0 & L_{\mathcal{S}_{0}}
\end{array}\right]
$$

where $L_{\mathcal{A}_{0}} L_{\mathcal{A}_{0}}^{\top}=\mathcal{A}_{0}, L_{\mathcal{S}_{0}} L_{\mathcal{S}_{0}}^{\top}=\mathcal{S}_{0}$, and precondition the saddle point matrix $\mathcal{M}_{p}$ symmetrically, we obtain the following symmetric matrix with a saddle point structure

$$
\mathcal{M}_{\mathcal{Q}_{1}}:=\left[\begin{array}{cc}
L_{\mathcal{A}_{0}}^{-1} \mathcal{A} L_{\mathcal{A}_{0}}^{-\top} & L_{\mathcal{A}_{0}}^{-1} \mathcal{B}^{\top} L_{\mathcal{S}_{0}}^{-\top}  \tag{2.60}\\
L_{\mathcal{S}_{0}}^{-1} \mathcal{B} L_{\mathcal{A}_{0}}^{-\top} & 0
\end{array}\right]
$$

Eigenvalue bounds for $\mathcal{M}_{\mathcal{Q}_{1}}$ can be obtained using Theorem 2.21 again. The $(1,1)$ block $L_{\mathcal{A}_{0}}^{-1} \mathcal{A} L_{\mathcal{A}_{0}}^{-\top}$ is symmetric and positive definite and its eigenvalues are the same as those of $\mathcal{A}_{0}^{-1} \mathcal{A}$. The eigenvalue bounds are given in the following lemma.

Lemma 2.22. Let $\mathcal{A}$ and $\mathcal{A}_{0}$ be defined in (2.37) and (2.59). If Assumption 2.18 holds, the eigenvalues of $\mathcal{A}_{0}^{-1} \mathcal{A}$ lie in the following interval

$$
\begin{equation*}
\left[\frac{E_{\min }}{e_{0}^{\max }}, \frac{E_{\max }}{e_{0}^{\min }}\right], \tag{2.61}
\end{equation*}
$$

where, recall, $E_{\min }, E_{\max }$ are defined in Assumption 2.18 and $e_{0}^{\min }, e_{0}^{\max }$ are defined in Assumption 2.19.

Proof. The matrix $\mathcal{A}_{0}^{-1} \mathcal{A}$ is block diagonal and is defined by

$$
\mathcal{A}_{0}^{-1} \mathcal{A}=\left[\begin{array}{cc}
\mathbb{A}_{0}^{-1} \mathbb{A} & 0 \\
0 & \mathbb{D}_{0}^{-1} \mathbb{D}
\end{array}\right] .
$$

The eigenvalues of $\mathcal{A}_{0}^{-1} \mathcal{A}$ are the union of eigenvalues of the first diagonal block and the second diagonal block. For a vector $\mathbf{w} \in \mathbb{R}^{2 n_{u} n_{y}}$, there is a corresponding function $\boldsymbol{w} \in \boldsymbol{V}_{h d}$. We can obtain eigenvalue bounds for $\mathbb{A}_{0}^{-1} \mathbb{A}$ by investigating the bounds for the Rayleigh quotient

$$
\frac{\mathbf{w}^{\top} \mathbb{A} \mathbf{w}}{\mathbf{w}^{\top} \mathbb{A}_{0} \mathbf{w}}
$$

Since $\mathbb{A}$ is symmetric and positive definite, we have

$$
\begin{aligned}
\mathbf{w}^{\top} \mathbb{A} \mathbf{w}=a(\boldsymbol{w}, \boldsymbol{w}) & =\beta \int_{\Gamma} \rho(\boldsymbol{y}) \int_{D} E(\boldsymbol{x}, \boldsymbol{y}) \boldsymbol{\epsilon}(\boldsymbol{w}): \boldsymbol{\epsilon}(\boldsymbol{w}) \mathrm{d} \boldsymbol{x} \mathrm{~d} \boldsymbol{y} \\
& \leqslant \frac{E_{\max }}{e_{0}^{\min }} \int_{\Gamma} \rho(\boldsymbol{y}) \int_{D} e_{0}(\boldsymbol{x}) \boldsymbol{\epsilon}(\boldsymbol{w}): \boldsymbol{\epsilon}(\boldsymbol{w}) \mathrm{d} \boldsymbol{x} \mathrm{~d} \boldsymbol{y} \\
& =\frac{E_{\max }}{e_{0}^{\min }} \mathbf{w}^{\top} \mathbb{A}_{0} \mathbf{w} .
\end{aligned}
$$

Similarly,

$$
\mathbf{w}^{\top} \mathbb{A} \mathbf{w} \geqslant \frac{E_{\min }}{e_{0}^{\max }} \mathbf{w}^{\top} \mathbb{A}_{0} \mathbf{w}
$$

Therefore, the eigenvalues of $\mathbb{A}_{0}^{-1} \mathbb{A}$ lie in the interval (2.61). Analogously, we can obtain eigenvalue bounds for $\mathbb{D}_{0}^{-1} \mathbb{D}$ by investigating bounds for the Rayleigh quotient

$$
\frac{\mathbf{q}^{\top} \mathbb{D} \mathbf{q}}{\mathbf{q}^{\top} \mathbb{D}_{0} \mathbf{q}}
$$

where $\mathbf{q} \in \mathbb{R}^{n_{p} n_{y}}$. It is straightforward to prove that the eigenvalue of $\mathbb{D}_{0}^{-1} \mathbb{D}$ lie in the same interval as $\mathbb{A}_{0}^{-1} \mathbb{A}$. Hence, the eigenvalues of $\mathcal{A}_{0}^{-1} \mathcal{A}$ also lie in the interval (2.61).

The above result tells us that the minimum and maximum eigenvalues $\tilde{\mu}_{\min }, \tilde{\mu}_{\max }$ of $L_{\mathcal{A}_{0}}^{-1} \mathcal{A} L_{\mathcal{A}_{0}}^{-\top}$ satisfy

$$
\tilde{\mu}_{\min } \geqslant \frac{E_{\min }}{e_{0}^{\max }}, \quad \tilde{\mu}_{\max } \leqslant \frac{E_{\max }}{e_{0}^{\min }} .
$$

Multiplying the $(2,1)$ block and the $(1,2)$ block of $\mathcal{M}_{\mathcal{Q}_{1}}$ gives

$$
L_{\mathcal{S}_{0}}^{-1} \mathcal{B} L_{\mathcal{A}_{0}}^{-\top} L_{\mathcal{A}_{0}}^{-1} \mathcal{B}^{\top} L_{\mathcal{S}_{0}}^{-\top}=L_{\mathcal{S}_{0}}^{-1} \mathcal{S}_{0} L_{\mathcal{S}_{0}}^{-\top}=I .
$$

Therefore, all singular values of $L_{\mathcal{S}_{0}}^{-1} \mathcal{B} L_{\mathcal{A}_{0}}^{-\top}$ are 1. Eigenvalue bounds for $\mathcal{M}_{\mathcal{Q}_{1}}$ are given in the following theorem.

Theorem 2.23. Let $\tilde{\mu}_{\min }, \tilde{\mu}_{\max }$ be the minimum and maximum eigenvalues of $L_{\mathcal{A}_{0}}^{-1} \mathcal{A} L_{\mathcal{A}_{0}}^{-\top}$, where $\mathcal{A}$ is defined in (2.37) and $L_{\mathcal{A}_{0}} L_{\mathcal{A}_{0}}^{\top}=\mathcal{A}_{0}$ with $\mathcal{A}_{0}$ defined in (2.59). Then the eigenvalues of $\mathcal{M}_{\mathcal{Q}_{1}}$ defined in (2.60) lie in the union of the following intervals

$$
\begin{array}{r}
{\left[\frac{1}{2}\left(\tilde{\mu}_{\min }-\sqrt{\tilde{\mu}_{\min }^{2}+4}\right), \frac{1}{2}\left(\tilde{\mu}_{\max }-\sqrt{\tilde{\mu}_{\max }^{2}+4}\right)\right] \cup} \\
{\left[\tilde{\mu}_{\min }, \frac{1}{2}\left(\tilde{\mu}_{\max }+\sqrt{\tilde{\mu}_{\max }^{2}+4}\right)\right] .} \tag{2.62}
\end{array}
$$

Since $\mathcal{M}$ and $\mathcal{P}_{1}$ are similar to $\mathcal{M}_{p}$ and $\mathcal{Q}_{1}$, respectively, $\mathcal{P}_{1}^{-1} \mathcal{M}$ is similar to $\mathcal{M}_{\mathcal{Q}_{1}}$. The eigenvalues of $\mathcal{P}_{1}^{-1} \mathcal{M}$ also lie in the intervals (2.62). We note that the theoretical eigenvalue bounds are independent of the finite element mesh size $h$, the polynomial degree $d$ and crucially, when working in the nearly incompressible case, the Poisson ratio $\nu$. However, they do depend on the maximum and minimum values of $E$ and its mean.

Eigenvalue bounds for $\mathcal{P}_{2}^{-1} \mathcal{M}$ are established in [44]. The preconditioner $\mathcal{P}_{2}$ is similar to the following matrix

$$
\mathcal{Q}_{2}:=\left[\begin{array}{cc}
\mathcal{A}_{0, \text { approx }} & 0 \\
0 & \mathcal{S}_{0, \text { approx }}
\end{array}\right],
$$

where $\mathcal{A}_{0, \text { approx }}$ is defined as

$$
\mathcal{A}_{0, \text { approx }}:=\left[\begin{array}{ccc}
\beta I_{n_{y}} \otimes \frac{2}{3}\left(A_{0}^{(11)}+A_{0}^{(22)}\right) & 0 & 0  \tag{2.63}\\
0 & \beta I_{n_{y}} \otimes \frac{2}{3}\left(A_{0}^{(11)}+A_{0}^{(22)}\right) & 0 \\
0 & 0 & \mathbb{D}_{0}
\end{array}\right],
$$

and $\mathcal{S}_{0, \text { approx }}$ is defined as

$$
\mathcal{S}_{0, \text { approx }}:=\mathcal{B} \mathcal{A}_{0, \text { approx }}^{-1} \mathcal{B}^{\top}=I \otimes S_{0, \text { approx }},
$$

where $S_{0, \text { approx }}$ is defined in (2.56) when $e_{0}(\boldsymbol{x})$ is a spatial function, or in (2.57) when $e_{0}(\boldsymbol{x})$ is a constant. The eigenvalues of $\mathcal{A}_{0, \text { approx }}^{-1} \mathcal{A}$ lie in the following interval

$$
\left[C_{K} \frac{E_{\min }}{e_{0}^{\max }}, \frac{E_{\max }}{e_{0}^{\min }}\right],
$$

where $0<C_{K}<1$ is the Korn constant, see [16]. The minimum and maximum eigenvalues $\hat{\mu}_{\text {min }}, \hat{\mu}_{\text {max }}$ of $\mathcal{A}_{0, \text { approx }}^{-1} \mathcal{A}$ therefore satisfy

$$
\hat{\mu}_{\min } \geqslant C_{K} \frac{E_{\min }}{e_{0}^{\max }}, \quad \hat{\mu}_{\max } \leqslant \frac{E_{\max }}{e_{0}^{\min }} .
$$

The matrix $\mathcal{P}_{2}^{-1} \mathcal{M}$ is similar to

$$
\begin{equation*}
\mathcal{M}_{\mathcal{Q}_{2}}:=L_{\mathcal{Q}_{2}}^{-1} \mathcal{M}_{p} L_{\mathcal{Q}_{2}}^{-\top} \tag{2.64}
\end{equation*}
$$

where $L_{\mathcal{Q}_{2}} L_{\mathcal{Q}_{2}}^{\top}=\mathcal{Q}_{2}$. Eigenvalue bounds for $\mathcal{M}_{\mathcal{Q}_{2}}$ are derived in [44], which are given in the following theorem.

Theorem 2.24. Let $\hat{\mu}_{\min }, \hat{\mu}_{\max }$ be the minimum and maximum eigenvalues of the matrix $L_{\mathcal{A}_{0, \text { approx }}}^{-1} \mathcal{A} L_{\mathcal{A}_{0}, \text { approx }}^{-\top}$, where $\mathcal{A}$ is defined in (2.37) and $L_{\mathcal{A}_{0, \text { approx }}} L_{\mathcal{A}_{0, \text { approx }}}^{\top}=\mathcal{A}_{0, \text { approx }}$ with $\mathcal{A}_{0, \text { approx }}$ defined in (2.63). Then the eigenvalues of $\mathcal{M}_{\mathcal{Q}_{2}}$ defined in (2.64) lie in the union of the following intervals

$$
\begin{array}{r}
{\left[\frac{1}{2}\left(\hat{\mu}_{\min }-\sqrt{\hat{\mu}_{\min }^{2}+4 \Theta^{2}}\right), \frac{1}{2}\left(\hat{\mu}_{\max }-\sqrt{\hat{\mu}_{\max }^{2}+4 \theta^{2}}\right)\right] \cup} \\
{\left[\hat{\mu}_{\min }, \frac{1}{2}\left(\hat{\mu}_{\max }+\sqrt{\hat{\mu}_{\max }+4 \Theta^{2}}\right)\right],} \tag{2.65}
\end{array}
$$

where

$$
\theta^{2}:=\frac{\gamma^{2}}{e_{0}^{\max }}, \quad \Theta^{2}:=\frac{2}{e_{0}^{\min }},
$$

and $\gamma$ is the inf-sup constant defined in (2.24). Alternatively, if $e_{0}(\boldsymbol{x})$ is a constant and $S_{0, a p p r o x}$ is defined in (2.57), then we obtain a slightly improved bound with

$$
\theta^{2}=\gamma^{2}, \quad \Theta^{2}=2
$$

Again, we can see that the theoretical eigenvalue bounds for $\mathcal{P}_{2}^{-1} \mathcal{M}$ are independent of the finite element mesh size $h$, the polynomial degree $d$ and the Poisson ratio $\nu$ but depend on the maximum and minimum values of $E$ and its mean.

### 2.3.2 Numerical experiments

In this section, we introduce two test problems and look at the eigenvalues of $\mathcal{P}_{1}^{-1} \mathcal{M}$ and $\mathcal{P}_{2}^{-1} \mathcal{M}$ for the two test problems. We then apply preconditioned MINRES to the linear systems obtained when we discretise them with the stochastic Galerkin mixed finite element method discussed in Section 2.2.

Test problem 1. First, we choose the spatial domain to be $D:=[-1,1]^{2}$ with Neumann boundary $\partial D_{N}:=\{1\} \times(-1,1)$ and Dirichlet boundary $\partial D_{D}=\partial D \backslash \partial D_{N}$ and we choose the body force

$$
\boldsymbol{f}(\boldsymbol{x})=\left[\begin{array}{l}
10^{-2} \\
10^{-2}
\end{array}\right]
$$

Following Example 2.16, we choose the parameter-dependent Young modulus to be

$$
\begin{equation*}
E(\boldsymbol{x}, \boldsymbol{y}):=e_{0}(\boldsymbol{x})+\sum_{r=1}^{M} e_{r}(\boldsymbol{x}) y_{r}, \quad \boldsymbol{x} \in D, \boldsymbol{y} \in \Gamma, \tag{2.66}
\end{equation*}
$$

where $\Gamma=[-1,1]^{M}, e_{0}=1$ and

$$
\begin{equation*}
e_{r}(\boldsymbol{x})=\sqrt{3} \sqrt{\lambda_{r}} \eta_{r}(\boldsymbol{x}), \quad r=1,2, \ldots, M, \tag{2.67}
\end{equation*}
$$

where $\left(\lambda_{r}, \eta_{r}(\boldsymbol{x})\right)$ are the eigenpairs associated with the separable exponential covariance function (2.15).

Test problem 2. In this problem, we choose the spatial domain $D=[0,1]^{2}$ with Neumann boundary $\partial D_{N}=\{1\} \times(0,1)$ and Dirichlet boundary $\partial D_{D}=\partial D \backslash \partial D_{N}$. Again, we define the body force as

$$
\boldsymbol{f}(\boldsymbol{x})=\left[\begin{array}{l}
10^{-2} \\
10^{-2}
\end{array}\right] .
$$

The Young modulus is defined as in (2.66), but following Example 2.17, we now choose

$$
e_{0}(\boldsymbol{x})=1, \quad e_{r}(\boldsymbol{x})=\gamma_{r} \cos \left(2 \pi \varrho_{1}(r) x_{1}\right) \cos \left(2 \pi \varrho_{2}(r) x_{2}\right),
$$

where $\gamma_{r}=0.832 r^{-4}$ for fast decay, or $\gamma_{r}=0.547 r^{-2}$ for slow decay and $\varrho_{1}(r)$ and $\varrho_{2}(r)$ are defined in Example 2.17.


Figure 2.8: Expectations and variances of components of the displacement $\boldsymbol{u}$ and Herrmann pressure $p$ when $l=6, \nu=0.4999, M=13, d=4$ and $\sigma=0.17$ for Test problem 1.


Figure 2.9: Expectations and variances of components of the displacement $\boldsymbol{u}$ and Herrmann pressure $p$ when $l=6, \nu=0.4999, M=13$ and $d=4$ for the fast decay case in Test problem 2.

In Figures 2.8-2.9, we plot the expectation and variance of the solution obtained using preconditioned MINRES for both test problems with fixed Poisson ratio $\nu=$ $0.4999, M=13$, finite element grid level $l=6$ and polynomial degree $d=4$. The stopping condition is when the preconditioned relative residual in (2.52) is less than $10^{-6}$. Note that these problems have spatial singularities due to the choice of boundary condition so a fine spatial mesh is needed.

Next, we will apply preconditioned MINRES to problems of small and moderate size with the two preconditioners $\mathcal{P}_{1}$ and $\mathcal{P}_{2}$. All numerical experiments are performed in MATLAB R2022a on a laptop with a 2.4 GHz 8-Core Intel Core i9 processor with 64 GB of RAM. The MATLAB code we used for the parameter-dependent linear elasticity problem is a combination of the code for deterministic linear elasticity problem and the code for stochastic diffusion problem in S-IFISS toolbox [75]. As discussed in Section 2.3.1, the theoretical eigenvalue bounds for the preconditioned coefficient matrix with these two preconditioners are independent of the mesh size $h$, the total polynomial degree $d$ and the Poisson ratio $\nu$ but depend on the maximum and minimum values of $E$ and its mean. This means that the convergence of preconditioned MINRES should only depend on the maximum and minimum values of $E$ and its mean. For Test problem 1, we need to choose a small value of the standard deviation $\sigma$ to ensure that Assumption 2.18 holds and the variance of $E(\boldsymbol{x}, \boldsymbol{y})$ increases slowly if we increase $M$. In theory, the convergence of MINRES will be affected by keeping more parameters $y_{r}$ and increasing $\sigma$ in Test problem 1 since these values affect $E_{\min }$ and $E_{\text {max }}$. We now perform some numerical experiments to verify the above. For all numerical experiments, we terminate MINRES when the preconditioned relative residual is less than $10^{-6}$ and we choose the initial guess $\mathbf{x}_{0}=\mathbf{0}$. The timings we record are in seconds. Note that the iteration counts are not comparable for $\mathcal{P}_{1}$ and $\mathcal{P}_{2}$, since their stopping conditions are different

Preconditioned MINRES with varying grid level: The iteration counts and timings using preconditioned MINRES with varying mesh level l, fixed Poisson ratio $\nu=0.4999$ and total polynomial degree $d=4$ are recorded for the two test problems in Tables 2.4-2.6. For Test problem 1, we choose $M=7$ since it keeps $85 \%$ of the total variance of $E$. For the two cases of Test problem 2, we choose $M=5$ since it keeps nearly $100 \%$ of the total variance of $E$. We can see the iteration counts are independent of mesh size using the preconditioner $\mathcal{P}_{1}$ and nearly independent of mesh size using the preconditioner $\mathcal{P}_{2}$.

| $l$ | $\mathcal{P}_{1}$ | $\mathcal{P}_{2}$ |
| :---: | :---: | :---: |
| 5 | $27(2.21 \mathrm{e} 1)$ | $109(1.28 \mathrm{e} 1)$ |
| 6 | $27(3.48 \mathrm{e} 2)$ | $116(7.75 \mathrm{e} 1)$ |
| 7 | $27(5.27 \mathrm{e} 3)$ | $119(8.24 \mathrm{e} 2)$ |

Table 2.4: Iteration counts and timings (in brackets) using preconditioned MINRES for Test problem 1 when $\sigma=0.17, \nu=0.4999, M=7$ and $d=4$.

| $l$ | $\mathcal{P}_{1}$ | $\mathcal{P}_{2}$ |
| :---: | :---: | :---: |
| 5 | $40(1.22 \mathrm{e} 1)$ | $134(5.42 \mathrm{e} 0)$ |
| 6 | $43(2.04 \mathrm{e} 2)$ | $145(3.67 \mathrm{e} 1)$ |
| 7 | $43(3.17 \mathrm{e} 3)$ | $155(4.06 \mathrm{e} 2)$ |

Table 2.5: Iteration counts and timings (in brackets) using preconditioned MINRES for the slow decay case in Test problem 2 when $\nu=0.4999, M=5$ and $d=4$.

| $l$ | $\mathcal{P}_{1}$ | $\mathcal{P}_{2}$ |
| :---: | :---: | :---: |
| 5 | $59(1.81 \mathrm{e} 1)$ | $184(8.13 \mathrm{e} 0)$ |
| 6 | $64(3.02 \mathrm{e} 2)$ | $205(5.50 \mathrm{e} 1)$ |
| 7 | $65(4.82 \mathrm{e} 3)$ | $213(5.65 \mathrm{e} 2)$ |

Table 2.6: Iteration counts and timings (in brackets) using preconditioned MINRES for the fast decay case in Test problem 2 when $\nu=0.4999, M=5$ and $d=4$.

Preconditioned MINRES with varying Poisson ratio: In Tables 2.7-2.9, we record numerical results for the two test problems with varying Poisson ratio $\nu$. We can see the iteration counts remain bounded as $\nu \rightarrow 0.5$ for both preconditioners.

| $\nu$ | $\mathcal{P}_{1}$ | $\mathcal{P}_{2}$ |
| :---: | :---: | :---: |
| 0.4 | $27(3.53 \mathrm{e} 2)$ | $109(6.22 \mathrm{e} 1)$ |
| 0.49 | $27(3.49 \mathrm{e} 2)$ | $114(7.78 \mathrm{e} 1)$ |
| 0.499 | $27(3.51 \mathrm{e} 2)$ | $116(9.11 \mathrm{e} 1)$ |
| 0.4999 | $27(3.49 \mathrm{e} 2)$ | $116(9.03 \mathrm{e} 1)$ |

Table 2.7: Iteration counts and timings (in brackets) using preconditioned MINRES for Test problem 1 with varying $\nu$ when $\sigma=0.17, l=6, M=7$ and $d=4$.

| $\nu$ | $\mathcal{P}_{1}$ | $\mathcal{P}_{2}$ |
| :---: | :---: | :---: |
| 0.4 | $43(2.05 \mathrm{e} 2)$ | $107(2.74 \mathrm{e} 1)$ |
| 0.49 | $43(2.06 \mathrm{e} 2)$ | $138(3.52 \mathrm{e} 1)$ |
| 0.499 | $43(2.14 \mathrm{e} 2)$ | $144(3.60 \mathrm{e} 1)$ |
| 0.4999 | $43(2.09 \mathrm{e} 2)$ | $145(3.62 \mathrm{e} 1)$ |

Table 2.8: Iteration counts and timings (in brackets) using preconditioned MINRES for the slow decay case in Test problem 2 with varying $\nu$ when $l=6, M=5$ and $d=4$.

| $\nu$ | $\mathcal{P}_{1}$ | $\mathcal{P}_{2}$ |
| :---: | :---: | :---: |
| 0.4 | $68(3.20 \mathrm{e} 2)$ | $153(4.05 \mathrm{e} 1)$ |
| 0.49 | $65(3.06 \mathrm{e} 2)$ | $194(4.90 \mathrm{e} 1)$ |
| 0.499 | $65(3.07 \mathrm{e} 2)$ | $205(5.17 \mathrm{e} 1)$ |
| 0.4999 | $64(3.02 \mathrm{e} 2)$ | $205(5.14 \mathrm{e} 1)$ |

Table 2.9: Iteration counts and timings (in brackets) using preconditioned MINRES for the fast decay case in Test problem 2 with varying $\nu$ when $l=6, M=5$ and $d=4$.

Preconditioned MINRES with varying polynomial degree: In Tables 2.102.12, we record results for the two test problems with varying total polynomial degree $d$. The iteration counts barely depend on the polynomial degree $d$ for both preconditioners.

| $d$ | $\mathcal{P}_{1}$ | $\mathcal{P}_{2}$ |
| :---: | :---: | :---: |
| 3 | $24(1.13 \mathrm{e} 2)$ | $111(2.80 \mathrm{e} 1)$ |
| 4 | $27(3.43 \mathrm{e} 2)$ | $116(8.22 \mathrm{e} 1)$ |
| 5 | $29(8.80 \mathrm{e} 2)$ | $120(1.97 \mathrm{e} 2)$ |
| 6 | $30(1.99 \mathrm{e} 3)$ | $122(4.19 \mathrm{e} 2)$ |

Table 2.10: Iteration counts and timings (in brackets) using preconditioned MINRES for Test problem 1 with varying $d$ when $\sigma=0.17, l=6, \nu=0.4999$ and $M=7$.

| $d$ | $\mathcal{P}_{1}$ | $\mathcal{P}_{2}$ |
| :---: | :---: | :---: |
| 3 | $39(8.17 \mathrm{e} 1)$ | $134(1.90 \mathrm{e} 1)$ |
| 4 | $43(2.01 \mathrm{e} 2)$ | $145(4.55 \mathrm{e} 1)$ |
| 5 | $45(4.20 \mathrm{e} 2)$ | $153(8.39 \mathrm{e} 1)$ |
| 6 | $46(8.13 \mathrm{e} 2)$ | $158(1.50 \mathrm{e} 2)$ |

Table 2.11: Iteration counts and timings (in brackets) using preconditioned MINRES for the slow decay case in Test problem 2 with varying $d$ when $l=6, \nu=0.4999$ and $M=5$.

| $d$ | $\mathcal{P}_{1}$ | $\mathcal{P}_{2}$ |
| :---: | :---: | :---: |
| 3 | $58(1.20 \mathrm{e} 2)$ | $184(2.36 \mathrm{e} 1)$ |
| 4 | $64(3.10 \mathrm{e} 2)$ | $205(6.09 \mathrm{e} 1)$ |
| 5 | $68(6.55 \mathrm{e} 2)$ | $221(1.24 \mathrm{e} 2)$ |
| 6 | $73(1.28 \mathrm{e} 3)$ | $232(2.20 \mathrm{e} 2)$ |

Table 2.12: Iteration counts and timings (in brackets) using preconditioned MINRES for the fast decay case in Test problem 2 with varying $d$ when $l=6, \nu=0.4999$ and $M=5$.

Preconditioned MINRES with varying number of parameters: In Tables 2.13-2.15, we record the iteration counts and timings for the two test problems with varying $M$. For both preconditioners, the iteration counts are nearly independent of $M$.

| $M$ | $\mathcal{P}_{1}$ | $\mathcal{P}_{2}$ |
| :---: | :---: | :---: |
| 5 | $26(1.26 \mathrm{e} 2)$ | $112(2.81 \mathrm{e} 1)$ |
| 7 | $27(3.44 \mathrm{e} 2)$ | $116(8.04 \mathrm{e} 1)$ |
| 13 | $28(2.59 \mathrm{e} 3)$ | $117(6.38 \mathrm{e} 2)$ |

Table 2.13: Iteration counts and timings (in brackets) using preconditioned MINRES for Test problem 1 with varying $M$ when $\sigma=0.17, l=6, \nu=0.4999$ and $d=4$.

| $M$ | $\mathcal{P}_{1}$ | $\mathcal{P}_{2}$ |
| :---: | :---: | :---: |
| 1 | $36(6.95 \mathrm{e} 0)$ | $126(1.89 \mathrm{e} 0)$ |
| 3 | $42(5.57 \mathrm{e} 1)$ | $142(1.03 \mathrm{e} 1)$ |
| 5 | $43(2.05 \mathrm{e} 2)$ | $145(3.64 \mathrm{e} 1)$ |

Table 2.14: Iteration counts and timings (in brackets) using preconditioned MINRES for the slow decay case in Test problem 2 with varying $M$ when $l=6, \nu=0.4999$ and $d=4$.

| $M$ | $\mathcal{P}_{1}$ | $\mathcal{P}_{2}$ |
| :---: | :---: | :---: |
| 1 | $62(1.18 \mathrm{e} 1)$ | $195(2.86 \mathrm{e} 0)$ |
| 3 | $64(8.35 \mathrm{e} 1)$ | $205(1.47 \mathrm{e} 1)$ |
| 5 | $64(3.14 \mathrm{e} 2)$ | $205(5.35 \mathrm{e} 1)$ |

Table 2.15: Iteration counts and timings (in brackets) using preconditioned MINRES for the fast decay case in Test problem 2 with varying $M$ when $l=6, \nu=0.4999$ and $d=4$.

## Preconditioned MINRES with varying standard deviation for Test prob-

 lem 1: For Test problem 1, we record the numerical results obtained with varying standard deviation $\sigma$ in Table 2.16. As expected, the iteration counts increase as $\sigma$ increases.| $\sigma$ | $\mathcal{P}_{1}$ | $\mathcal{P}_{2}$ |
| :---: | :---: | :---: |
| 0.085 | $19(2.44 \mathrm{e} 2)$ | $95(6.68 \mathrm{e} 1)$ |
| 0.170 | $27(3.56 \mathrm{e} 2)$ | $116(8.05 \mathrm{e} 1)$ |
| 0.255 | $37(4.72 \mathrm{e} 2)$ | $147(1.06 \mathrm{e} 2)$ |

Table 2.16: Iteration counts and timings (in brackets) in seconds using preconditioned MINRES for Test problem 1 with varying $\sigma$ when $l=6, \nu=0.4999, M=7$ and $d=4$.

Preconditioned MINRES for large problems: Finally, we implement preconditioned MINRES on some large problems using the chosen preconditioners. We consider Test problem 1 when $l=8,9, M=15, d=3,4, \nu=0.4$ and $\sigma=0.085$. The sizes of these high-dimensional problems are recorded in Table 2.17.

| $l$ | $n_{x}$ | $M$ | $d$ | $n_{y}$ | $n_{x} \cdot n_{y}$ |
| :---: | :---: | :---: | :---: | :---: | :---: |
| 8 | 163842 | 15 | 3 | 816 | 1.34 e 8 |
|  |  |  | 4876 | 6.35 e 8 |  |
| 9 | 655362 | 15 | 3 | 816 | 5.35 e 8 |
| 4 | 3876 | 2.54 e 9 |  |  |  |

Table 2.17: Sizes of some large problems.

We apply preconditioned MINRES with the preconditioners $\mathcal{P}_{1}$ and $\mathcal{P}_{2}$ to these large problems. The numerical results are recorded in Table 2.18. When we use preconditioner $\mathcal{P}_{1}$ to solve these problems, the machine that we used for the numerical experiments runs out of memory. This is mainly because there are dense blocks of size $n_{p} \times n_{p}$ in the preconditioner $\mathcal{P}_{1}$. If we use the preconditioner $\mathcal{P}_{2}$, we can solve linear systems with up to $5.35 \times 10^{8}$ equations but the machine runs out of memory if the number of equations is greater than $6 \times 10^{8}$. Since preconditioned MINRES requires assembling vectors of length $n_{x} \cdot n_{y}$, it cannot be used to solve high-dimensional problems when there is insufficient memory to store such vectors. In this case, a more memory-efficient algorithm is required to overcome this issue.

| $l$ | $M$ | $d$ | $\mathcal{P}_{1}$ | $\mathcal{P}_{2}$ |
| :---: | :---: | :---: | :---: | :---: |
| 8 | 15 | 3 | out of memory | $70(9.46 \mathrm{e} 3)$ |
| 9 | 15 | 3 | out of memory | out of memory |
|  |  | out of memory | $73(7.45 \mathrm{e} 4)$ |  |
| out of memory |  |  |  |  |

Table 2.18: Iteration counts and timings (in brackets) using preconditioned MINRES for Test problem 1 when $\sigma=0.085, \nu=0.4$; the stopping tolerance is $10^{-6}$.

### 2.3.3 Conclusions

The results obtained in Section 2.3.2 demonstrate that preconditioned MINRES can be used to solve the linear system (2.44) when the problem size is not too large. However, when fine finite element meshes are required and the number of input parameters is large, $n_{x} n_{y}$ is large. The dimensions shown in Table 2.17 are not uncommon. We considered two preconditioners: $\mathcal{P}_{1}$, which contains dense blocks of size $n_{p} \times n_{p}$ on the diagonal, and $\mathcal{P}_{2}$, which is a sparse approximation of $\mathcal{P}_{1}$.

When we apply preconditioned MINRES with the two preconditioners $\mathcal{P}_{1}$ and $\mathcal{P}_{2}$ to both test problems, we observe that iteration counts are nearly independent of the Poisson ratio $\nu$, the number of terms $M$, the mesh size $h$, and the polynomial degree $d$, but do depend on the standard deviation $\sigma$ in Test problem 1. The iteration counts are also higher for the fast decay case than the slow decay case in Test problem 2 because $E$ has a higher variance in the fast decay case. Timings are independent of the Poisson ratio $\nu$, but depend on the problem size and standard deviation in Test problem 1.

If we use preconditioned MINRES to solve high-dimensional problems, the machine that we used runs out of memory because it requires the storage of long vectors of length $n_{x} \cdot n_{y}$. To overcome this issue, we will introduce a reduced basis solver in the next chapter.

## Chapter 3

## A Reduced Basis Method

In [64], a reduced basis method called Multi-RB was developed for a multi-term matrix equation of the form (1.40) associated with the scalar elliptic PDE in (2.53) with parameter-dependent diffusion coefficient. We will now attempt to extend the method to more complex problems with saddle point structure. First, we consider the linear system (2.44) that arises when we apply a stochastic Galerkin mixed finite element method to (2.1)-(2.5). The solution vector $\mathbf{x} \in \mathbb{R}^{n_{x} n_{y}}$ can be reshaped into a matrix of the following form

$$
X:=\left[\begin{array}{c}
U_{1} \\
U_{2} \\
\widetilde{P} \\
P
\end{array}\right] \in \mathbb{R}^{n_{x} \times n_{y}},
$$

where the matrices $U_{1}, U_{2}, \widetilde{P}$ and $P$ are defined by

$$
\begin{aligned}
U_{1} & :=\left[\mathbf{u}_{1,1}, \mathbf{u}_{1,2}, \ldots, \mathbf{u}_{1, n_{y}}\right] \in \mathbb{R}^{n_{u} \times n_{y}}, \\
U_{2} & :=\left[\mathbf{u}_{2,1}, \mathbf{u}_{2,2}, \ldots, \mathbf{u}_{2, n_{y}}\right] \in \mathbb{R}^{n_{u} \times n_{y}}, \\
\widetilde{P} & :=\left[\widetilde{\mathbf{p}}_{1}, \widetilde{\mathbf{p}}_{2}, \ldots, \widetilde{\mathbf{p}}_{n_{y}}\right] \in \mathbb{R}^{n_{p} \times n_{y}}, \\
P & :=\left[\mathbf{p}_{1}, \mathbf{p}_{2}, \ldots, \mathbf{p}_{n_{y}}\right] \in \mathbb{R}^{n_{p} \times n_{y}} .
\end{aligned}
$$

If we stack the columns of the solution matrix $X$ on top of one another, we obtain the solution vector $\mathbf{x} \in \mathbb{R}^{n_{x} n_{y}}$ of the linear system (2.44). The number of rows $n_{x}=2\left(n_{u}+n_{p}\right)$ depends on the chosen finite element mesh size and the number of columns $n_{y}$ depends on the number $M$ of parameters in the expression for $E(\boldsymbol{x}, \boldsymbol{y})$ as
well as the total polynomial degree $d$ chosen for the parametric approximation space $S_{d}$. Using the property of the matrix Kronecker product introduced in (1.39), the linear system (2.44) can be reformulated as the multi-term matrix equation

$$
\begin{equation*}
\sum_{r=0}^{M} K_{r} X G_{r}=H \tag{3.1}
\end{equation*}
$$

If the singular values of the solution matrix $X$ decay rapidly enough, we expect to be able to approximate it well by a low-rank matrix.

In Section 3.1, we explain why a matrix with rapidly decaying singular values can be well approximated by a low rank matrix and investigate how the singular values of $X$ behave for our parameter-dependent linear elasticity problem. In Section 3.2, the multi-term reduced basis method from [64] is extended so that it can be applied to the matrix equations associated with the parameter-dependent linear elasticity problem. The resulting algorithm is provided at the end of this chapter.

### 3.1 Low rank analysis

The Frobenius norm of the solution matrix $X$ is defined as

$$
\|X\|_{F}=\sqrt{X: X}=\sqrt{\operatorname{trace}\left(X^{\top} X\right)}
$$

where : is the component product defined in (2.17). The Frobenius norm of $X$ is equivalent to the 2 norm of $\mathbf{x}$, where $\mathbf{x}:=\operatorname{vec}(X)$. Let $m$ denote the actual rank of $X$ and suppose our aim is to approximate $X$ by another matrix $\tilde{X} \in \mathbb{R}^{n_{x} \times n_{y}}$ of lower rank than $m$. Specifically, suppose we try to find a matrix $\widetilde{X}_{p}$ of rank $p$, with $p<m$, which solves the following problem

$$
\begin{align*}
& \min _{\tilde{X}_{p} \in \mathbb{R}^{n_{x} \times n_{y}}}\left\|X-\tilde{X}_{p}\right\|_{F}^{2}  \tag{3.2}\\
& \text { subject to } \operatorname{rank}\left(\tilde{X}_{p}\right)=p .
\end{align*}
$$

The solution to this minimisation problem can be found by decomposing $X$ using the singular value decomposition (SVD)

$$
X=U \Sigma V^{\top}, \quad \Sigma:=\operatorname{diag}\left(\sigma_{1}, \sigma_{2}, \ldots, \sigma_{m}, 0, \ldots, 0\right) \in \mathbb{R}^{n_{x} \times n_{y}}
$$

where $U \in \mathbb{R}^{n_{x} \times n_{x}}, V \in \mathbb{R}^{n_{y} \times n_{y}}$ are unitary matrices and $\sigma_{1} \geqslant \sigma_{2} \geqslant \ldots \geqslant \sigma_{m}>0$ are the nonzero singular values of $X$. Alternatively, the SVD of the matrix $X$ can be expressed as

$$
\begin{equation*}
X=\sum_{i=1}^{m} \sigma_{i} \mathbf{u}_{i} \mathbf{v}_{i}^{\top} \tag{3.3}
\end{equation*}
$$

where $\mathbf{u}_{i} \in \mathbb{R}^{n_{x}}, \mathbf{v}_{i} \in \mathbb{R}^{n_{y}}$ denote the $i$ th columns of $U, V$, respectively. The best approximation $\tilde{X}_{p}$ of rank $p$ that solves the problem (3.2) is

$$
\begin{equation*}
\tilde{X}_{p}=\sum_{i=1}^{p} \sigma_{i} \mathbf{u}_{i} \mathbf{v}_{i}^{\top} . \tag{3.4}
\end{equation*}
$$

See [21] for more details. We then have

$$
\left\|X-\tilde{X}_{p}\right\|_{F}^{2}=\sum_{i=p+1}^{m} \sigma_{i}^{2}
$$

and define the relative error $\varepsilon_{p}$ as

$$
\varepsilon_{p}:=\frac{\left\|X-\tilde{X}_{p}\right\|_{F}}{\|X\|_{F}}=\sqrt{\frac{\sum_{i=p+1}^{m} \sigma_{i}^{2}}{\sum_{i=1}^{m} \sigma_{i}^{2}}}
$$

Recall that if we solve the linear system (2.44) using preconditioned MINRES, we have to store the approximation $\mathbf{x}_{k}$ of length $n_{x} \cdot n_{y}$ and some other vectors of the same length at the $k$ th iteration. For a small tolerance tol, if there exists a matrix $\tilde{X}_{p}$ of rank $p \ll m$ solving (3.2) with $\varepsilon_{p}<t o l$, it will be possible to approximate $X$ by a low-rank matrix. The storage cost of the approximation $\widetilde{X}_{p}$ of rank $p$ in (3.4) is $O\left(p\left(n_{x}+n_{y}\right)\right)$, which is clearly less than the cost of storing vectors of length $n_{x} \cdot n_{y}$, or equivalently, matrices of size $n_{x} \times n_{y}$ if $p$ is much smaller than $n_{x}$ and $n_{y}$. However, the singular values and singular vectors of the exact solution are usually not available and so finding the approximation $\widetilde{X}_{p}$ solving (3.2) is impossible in practice.

In Tables 3.1-3.4, we record the smallest value of $p$ needed to achieve $\varepsilon_{p}<10^{-6}$ for numerical solutions $X$ obtained using preconditioned MINRES with the preconditioner $\mathcal{P}_{2}$ defined in Section 2.3.1 and stopping tolerance $10^{-6}$ for Test problems $1-2$ defined in Section 2.3.2. The value of $p$ is computed using numerical singular values of $X$. The actual rank is computed using the function rank in MATLAB with the default tolerance.

For Test problem 1, we record the smallest value of $p$ required for two different finite element meshes in Tables 3.1-3.2. The value of $p$ is independent of the dimension $n_{x}$
associated with the spatial discretisation and is lower for $\nu=0.4999$ than for $\nu=0.4$, but grows as the number of terms $M$ of the KL expansion of $E$ and the standard deviation $\sigma$ increase. For small $\sigma$, the rank $p$ required is insensitive to the polynomial degree $d$, but for large $\sigma$, there is some sensitivity to $d$. We observe that the actual rank of $X$ is close to the value of $n_{y}$, so the solution matrix is nearly of full column rank. The value of $p$ required for the chosen tolerance is much less than the actual rank when $n_{y}$ is not too small. This indicates that for large $n_{y}$ it is possible to approximate the solution matrix to (3.1) associated with Test problem 1 by a low rank matrix.

| $M$ |  |  | $\nu=0.4$ |  | $\nu=0.4999$ |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | $\sigma=0.085$ | $\sigma=0.17$ | $\sigma=0.085$ | $\sigma=0.17$ |
| 7 | 3 | 120 | $82(120)$ | $107(120)$ | $69(120)$ | $93(120)$ |
|  | 4 | 330 | $84(330)$ | $140(330)$ | $73(330)$ | $122(330)$ |
|  | 5 | 792 | $85(792)$ | $150(792)$ | $73(792)$ | $132(792)$ |
| 13 | 3 | 560 | $195(560)$ | $271(560)$ | $170(560)$ | $231(560)$ |
|  | 4 | 2380 | $200(2176)$ | $323(2359)$ | $176(2201)$ | $283(2366)$ |

Table 3.1: The smallest $p$ needed to achieve $\varepsilon_{p}<10^{-6}$ for Test problem 1 with $n_{x}=2562$. The actual rank of $X$ is given in brackets.

| $M$ | $d$ | $n_{y}$ | $\nu=0.4$ |  | $\nu=0.4999$ |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | $\sigma=0.085$ | $\sigma=0.17$ | $\sigma=0.085$ | $\sigma=0.17$ |
| 7 | 3 | 120 | $80(120)$ | $106(120)$ | $68(120)$ | $92(120)$ |
|  | 4 | 330 | $83(330)$ | $137(330)$ | $71(330)$ | $120(330)$ |
|  | 5 | 792 | $83(792)$ | $146(792)$ | $72(792)$ | $129(792)$ |
| 13 | 3 | 560 | $190(560)$ | $268(560)$ | $166(560)$ | $229(560)$ |
|  | 4 | 2380 | $194(2333)$ | $319(2380)$ | $171(2342)$ | $282(2380)$ |

Table 3.2: The smallest $p$ needed to achieve $\varepsilon_{p}<10^{-6}$ for Test problem 1 with $n_{x}=10242$. The actual rank of $X$ is given in brackets.

The smallest values of $p$ needed to achieve $\varepsilon_{p}<10^{-6}$ for Test problem 2 using two different finite element meshes are recorded in Tables 3.3-3.4. Again, the value of $p$ required is independent of $n_{x}$ and decreases when the Poisson ratio $\nu$ changes from 0.4
to 0.4999 , but grows as the number of terms $M$ and the polynomial degree $d$ increase. For the slow decay case, the actual rank is close to the value of $n_{y}$ and so the solution matrix is nearly of full rank. We also observe that the value of $p$ is much less than the actual rank when $n_{y}$ is not too small. This means it is possible to approximate $X$ by a low rank matrix when $n_{y}$ is large for the slow decay case. For the fast decay case, the actual rank is observed to be much less than $n_{y}$ when $M$ and $d$ (and hence $n_{y}$ ) are large enough and the value of $p$ is much less than the actual rank for all values of $n_{y}$. This means it should be possible to approximate $X$ by a low rank matrix for all choices of $n_{y}$ for the fast decay case.

| $M$ | $d$ | $n_{y}$ | $\nu=0.4$ |  | $\nu=0.4999$ |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | slow decay | fast decay | slow decay | fast decay |
| 7 | 3 | 120 | $94(120)$ | $40(120)$ | $82(120)$ | $37(120)$ |
|  | 4 | 330 | $138(330)$ | $56(313)$ | $106(330)$ | $48(302)$ |
|  | 5 | 792 | $164(791)$ | $67(548)$ | $118(787)$ | $56(504)$ |
| 13 | 3 | 560 | $205(560)$ | $57(498)$ | $142(560)$ | $48(452)$ |
|  | 4 | 2380 | $257(2076)$ | $75(840)$ | $165(1858)$ | $59(732)$ |

Table 3.3: The smallest $p$ needed to achieve $\varepsilon_{p}<10^{-6}$ for Test problem 2 with $n_{x}=2562$. The actual rank of $X$ is given in brackets.

| $M$ |  | $n_{y}$ | $\nu=0.4$ |  | $\nu=0.4999$ |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  |  | slow decay | fast decay | slow decay | fast decay |
| 7 | 3 | 120 | $92(120)$ | $39(120)$ | $81(120)$ | $36(120)$ |
|  | 4 | 330 | $136(330)$ | $55(318)$ | $105(330)$ | $47(311)$ |
|  | 5 | 792 | $160(792)$ | $66(585)$ | $118(792)$ | $55(563)$ |
| 13 | 3 | 560 | $199(560)$ | $56(518)$ | $139(560)$ | $47(481)$ |
|  | 4 | 2380 | $251(2273)$ | $73(960)$ | $165(2184)$ | $58(859)$ |

Table 3.4: The smallest $p$ needed to achieve $\varepsilon_{p}<10^{-6}$ for Test problem 2 with $n_{x}=10242$. The actual rank of $X$ is given in brackets.

For both test problems with mesh level $l=6$, (giving $n_{x}=10242$ ), $M=13$, and $d=4$, (giving $n_{y}=2380$ ), we plot the normalised singular values $\sigma_{i} / \sigma_{1}$ of $X$ and the
relative error $\varepsilon_{p}$ of the best rank $p$ approximation in Figures 3.1-3.2.


Figure 3.1: The normalised singular values of $X$ (left) and relative error $\varepsilon_{p}$ (right) of the best rank $p$ approximation for Test problem 1.


Figure 3.2: The normalised singular values of $X$ (left) and relative error $\varepsilon_{p}$ (right) of the best rank $p$ approximation for Test problem 2.

In Figure 3.1a, we can see that the normalised singular values are greater than machine epsilon when $\sigma=0.17$ and a small number of normalised singular values are less than machine epsilon when $\sigma=0.085$. The matrix $X$ is nearly of full rank for Test problem 1. For a fixed value of the normalised singular value $\sigma_{i} / \sigma_{1}$, the value of $i$ is smaller if the standard deviation is smaller. In Figure 3.1b, we can see for a fixed value of $\varepsilon_{p}$, the value of $p$ required is smaller if the standard deviation is smaller. In Figure 3.2a, we can see the decay rate of the normalised singular values depends on
the decay of the coefficients $\gamma_{r}$ of $e_{r}(\boldsymbol{x})$ for Test problem 2. The normalised singular values of the solution matrix decay faster for the fast decay case than for the slow decay case. More than half of the normalised singular values for the fast decay case are below machine epsilon. In this case, $X$ is not of full rank. In Figure 3.2b, we can see the decay of $\varepsilon_{p}$ is much faster in the fast decay case than in the slow decay case.

If it is possible to approximate the solution matrix $X$ by a low rank matrix, our aim is then to find an approximation with the following factored form

$$
\begin{equation*}
X_{k}=V_{k} Y_{k} \tag{3.5}
\end{equation*}
$$

where $V_{k} \in \mathbb{R}^{n_{x} \times n_{k}}, Y_{k} \in \mathbb{R}^{n_{k} \times n_{x}}$ and we aim to have $n_{k} \ll n_{x}$. To reduce the storage cost, we will work with the factors $V_{k}$ and $Y_{k}$ and avoid explicitly assembling the large matrix $X_{k}$. When we solve the linear system (1.38) using preconditioned MINRES, it requires storage of the approximation $\mathbf{x}_{k}$, residuals and search directions of length $n_{x} n_{y}$ at each iteration. The storage cost of $\mathbf{x}_{k}$ and vectors of the same length is $O\left(n_{x} n_{y}\right)$. If we store $\mathbf{x}_{k}$ in a factored form as in (3.5), the storage cost then becomes $O\left(n_{k}\left(n_{x}+n_{y}\right)\right)$. The storage cost is clearly reduced if $n_{k} \ll n_{x}$ and $n_{k} \ll n_{y}$.

### 3.2 Multi-term reduced basis method

As we discussed in Section 2.3.2, applying standard Krylov subspace methods (e.g. preconditioned MINRES) to the linear system (2.44) is infeasible when there is insufficient memory to store vectors of length $n_{x} \cdot n_{y}$. The machine we used to perform the numerical experiments ran out of memory for some high-dimensional problems. This deficiency motivates our need to develop a more memory-efficient solver.

We now extend the multi-term reduced basis method (Multi-RB) from [64] to the matrix equation (3.1) associated with the parameter-dependent linear elasticity problem in (2.1)-(2.5). The Multi-RB method was initially developed for simpler multi-term matrix equations associated with stochastic Galerkin finite element approximations of the scalar elliptic PDE in (2.53) with parameter-dependent diffusion coefficient. For such problems, the matrix equation has the same structure as (3.1). However, there is only one solution field and the coefficient matrix of the corresponding linear system is symmetric and positive definite. Consider a matrix equation (3.1) and
suppose $K_{r} \in \mathbb{R}^{n \times n}$, and $G_{r} \in \mathbb{R}^{m \times m}$. The method in [64] was designed for problems with $n>m$ and has two phases:

1. The preprocessing phase involves preconditioning and choosing shifts for the left matrices $K_{r}$ so that the spectra of the modified left matrices all lie in a small (ideally) positive interval independent of the spatial discretisation parameter $h$.
2. The iterative solution phase consists of an iterative orthogonal projection method applied to the modified matrix equation. We aim to find an approximation of the form $X_{k}=V_{k} Y_{k}$. At every iteration, we construct basis vectors and store them in the columns of a matrix $V_{k} \in \mathbb{R}^{n \times n_{k}}$, and then we apply a Galerkin condition to the residual and solve a reduced problem for $Y_{k}$. Note that the matrix $X_{k}$ is never explicitly formed.

The linear system associated with stochastic Galerkin approximation of the parameterdependent linear elasticity problem is indefinite. If we apply a symmetric preconditioning strategy with a positive definite preconditioner, the preconditioned linear system is still indefinite and applying a Galerkin condition does not guarantee the convergence. We now outline how the Multi-RB method can be modified to tackle the matrix equation (3.1) for the parameter-dependent linear elasticity problem.

### 3.2.1 Pre-processing phase

The pre-processing phase includes preconditioning and choosing shifts for all the left preconditioned matrices. First, we introduce two preconditioning strategies: a symmetric preconditioning strategy and a left preconditioning strategy.

Symmetric preconditioning: The symmetric preconditioning strategy can be applied with symmetric and positive definite preconditioners. For example, the preconditioner $K_{S}$ defined in (2.54) or $K_{S, \text { approx }}$ in (2.58). Suppose we choose $K_{S}$ as the preconditioner. Recall $K_{S}$ is a block diagonal matrix with the structure

$$
K_{S}:=\left[\begin{array}{ccc}
\beta A_{0} & 0 & 0 \\
0 & \tilde{\lambda}^{-1} D_{0} & 0 \\
0 & 0 & S_{0}
\end{array}\right]
$$

Since each diagonal block of $K_{S}$ is symmetric and positive definite, we may factorise them exploiting Cholesky factorisation

$$
A_{0}=L_{A_{0}} L_{A_{0}}^{\top}, \quad D_{0}=L_{D_{0}} L_{D_{0}}^{\top}, \quad S_{0}=L_{S_{0}} L_{S_{0}}^{\top}
$$

so that $K_{S}$ can also be factorised as $L_{K_{S}} L_{K_{S}}^{\top}$, where

$$
L_{K_{S}}:=\left[\begin{array}{ccc}
\beta^{1 / 2} L_{A_{0}} & 0 & 0 \\
0 & \tilde{\lambda}^{-1 / 2} L_{D_{0}} & 0 \\
0 & 0 & L_{S_{0}}
\end{array}\right] \in \mathbb{R}^{n_{x} \times n_{x}} .
$$

The symmetrically-preconditioned matrix equation can be obtained by multiplying (3.1) by $L_{K_{S}}^{-1}$ on the left. If we replace $X$ with $\hat{X}:=L_{K_{S}}^{\top} X$ then the preconditioned matrix equation becomes

$$
\begin{equation*}
\widehat{K}_{0} \widehat{X}+\sum_{r=1}^{M} \widehat{K}_{r} \widehat{X} G_{r}=\widehat{H} \tag{3.6}
\end{equation*}
$$

where the left matrices $\hat{K}_{r}$, for $r=0,1, \ldots, M$, and $\hat{H}$ are defined by

$$
\widehat{K}_{r}:=L_{K_{S}}^{-1} K_{r} L_{K_{S}}^{-\top}, \quad \hat{H}:=L_{K_{S}}^{-1} H=L_{K_{S}}^{-1} \mathbf{f g}_{0}^{\top}
$$

Note that since $\mathbf{g}_{0}$ has only one nonzero entry, $\hat{H}$ has only one nonzero column.
The first preconditioned left matrix $\widehat{K}_{0}$ in (3.6) is similar to $K_{S}^{-1} K_{0}$, and both have three distinct eigenvalues (see [57]), which are contained in the following set

$$
\begin{equation*}
\mathcal{S}_{0}:=\left\{\frac{1-\sqrt{5}}{2}, 1, \frac{1+\sqrt{5}}{2}\right\}, \tag{3.7}
\end{equation*}
$$

and independent of the finite element mesh size $h$. Hence, $\hat{K}_{0}$ is indefinite. For $r=1,2, \ldots, M$, eigenvalue bounds for $\widehat{K}_{r}$ are given in the following lemma,

Lemma 3.1. When Assumption 2.19 is satisfied, the eigenvalues of $\hat{K}_{r}$, for $r=$ $1,2, \ldots, M$, lie in the interval

$$
\begin{equation*}
\mathcal{S}_{r}:=\left[-\tau_{r}, \tau_{r}\right], \tag{3.8}
\end{equation*}
$$

where $\tau_{r} \rightarrow 0$ as $r \rightarrow \infty$, and we define

$$
\begin{equation*}
\tau_{r}:=\frac{\left\|e_{r}(\boldsymbol{x})\right\|_{\infty}}{e_{0}^{\min }} \tag{3.9}
\end{equation*}
$$

where $e_{r}(\boldsymbol{x})$ is defined for two test problems in Section 2.3.2.

Proof. For $r=1,2, \ldots, M$, the matrices $\widehat{K}_{r}$ and $K_{S}^{-1} K_{r}$ are similar and they have the same eigenvalues. The matrix $K_{S}^{-1} K_{r}$ is a block diagonal matrix

$$
K_{S}^{-1} K_{r}=\left[\begin{array}{ccc}
A_{0}^{-1} A_{r} & 0 & 0  \tag{3.10}\\
0 & D_{0}^{-1} D_{r} & 0 \\
0 & 0 & 0
\end{array}\right]
$$

where $A_{r}$ and $D_{r}$ are symmetric but indefinite. For a vector $\mathbf{v} \in \mathbb{R}^{2 n_{u}}$, there is a corresponding function $\boldsymbol{v} \in \boldsymbol{V}_{h}$. We can obtain eigenvalue bounds for $A_{0}^{-1} A_{r}$ by investigating bounds for the Rayleigh quotients

$$
\frac{\mathbf{v}^{\top} A_{r} \mathbf{v}}{\mathbf{v}^{\top} A_{0} \mathbf{v}}
$$

If Assumption 2.19 holds, the matrix $A_{0}$ is symmetric and positive definite. Since $A_{r}$, for $r=1,2, \ldots, M$, are indefinite, we have

$$
\begin{aligned}
\left|\mathbf{v}^{\top} A_{r} \mathbf{v}\right| & =\left|\int_{D} e_{r}(\boldsymbol{x}) \boldsymbol{\epsilon}(\boldsymbol{v}(\boldsymbol{x})): \boldsymbol{\epsilon}(\boldsymbol{v}(\boldsymbol{x})) \mathrm{d} \boldsymbol{x}\right| \\
& \leqslant \int_{D}\left|e_{r}(\boldsymbol{x})\right| \boldsymbol{\epsilon}(\boldsymbol{v}(\boldsymbol{x})): \boldsymbol{\epsilon}(\boldsymbol{v}(\boldsymbol{x})) \mathrm{d} \boldsymbol{x} \\
& \leqslant \frac{\left\|e_{r}(\boldsymbol{x})\right\|_{\infty}}{e_{0}^{\min }} \int_{D} e_{0}(\boldsymbol{x}) \boldsymbol{\epsilon}(\boldsymbol{v}(\boldsymbol{x})): \boldsymbol{\epsilon}(\boldsymbol{v}(\boldsymbol{x})) \mathrm{d} \boldsymbol{x} \\
& =\frac{\left\|e_{r}(\boldsymbol{x})\right\|_{\infty}}{e_{0}^{\min }} \mathbf{v}^{\top} A_{0} \mathbf{v} .
\end{aligned}
$$

Therefore, the eigenvalues of $A_{0}^{-1} A_{r}$ lie in the interval

$$
\mathcal{S}_{r}=\left[-\frac{\left\|e_{r}(\boldsymbol{x})\right\|_{\infty}}{e_{0}^{\min }}, \frac{\left\|e_{r}(\boldsymbol{x})\right\|_{\infty}}{e_{0}^{\min }}\right]
$$

Similarly, it is straightforward to prove that the eigenvalues of $D_{0}^{-1} D_{r}$ lie in the same interval as $A_{0}^{-1} A_{r}$. There is also a zero block with zero eigenvalues. Hence, the eigenvalues of $K_{S}^{-1} K_{r}$ also lie in $\mathcal{S}_{r}$.

If $e_{r}(\boldsymbol{x})$ is strictly positive on the spatial domain $D$ when $r \neq 0$, the eigenvalue bounds for $\hat{K}_{r}$ are given in the following lemma.

Lemma 3.2. When $e_{r}(\boldsymbol{x})$ is a strictly positive function, the eigenvalues of $\hat{K}_{r}$ lie in the union of two sets

$$
\begin{equation*}
\mathcal{S}_{r}:=\{0\} \cup\left[\frac{e_{r}^{\min }}{e_{0}^{\max }}, \tau_{r}\right], \tag{3.11}
\end{equation*}
$$

where $e_{r}^{\min }$ is defined as

$$
e_{r}^{\min }:=\inf _{\boldsymbol{x} \in D} e_{r}(\boldsymbol{x})
$$

Proof. The matrices $\hat{K}_{r}$ and $K_{S}^{-1} K_{r}$ are similar and they have the same eigenvalues. The matrix $K_{S}^{-1} K_{r}$ is defined as in (3.10). If $e_{r}(\boldsymbol{x})$ is strictly positive, then all of $A_{0}$, $D_{0}, A_{r}$ and $D_{r}$ are symmetric and positive definite. For a vector $\mathbf{v} \in \mathbb{R}^{2 n_{u}}$, there is a corresponding function $\boldsymbol{v} \in \boldsymbol{V}_{h}$. Again, we can obtain eigenvalue bounds for $A_{0}^{-1} A_{r}$ by investigating bounds for the Rayleigh quotients

$$
\frac{\mathbf{v}^{\top} A_{r} \mathbf{v}}{\mathbf{v}^{\top} A_{0} \mathbf{v}}
$$

Since $A_{0}$ and $A_{r}$ are symmetric and positive definite, we have

$$
\begin{aligned}
\mathbf{v}^{\top} A_{r} \mathbf{v} & =\int_{D} e_{r}(\boldsymbol{x}) \boldsymbol{\epsilon}(\boldsymbol{v}(\boldsymbol{x})): \boldsymbol{\epsilon}(\boldsymbol{v}(\boldsymbol{x})) \mathrm{d} \boldsymbol{x} \\
& \leqslant \frac{\left\|e_{r}(\boldsymbol{x})\right\|_{\infty}}{e_{0}^{\min }} \int_{D} e_{0}(\boldsymbol{x}) \boldsymbol{\epsilon}(\boldsymbol{v}(\boldsymbol{x})): \boldsymbol{\epsilon}(\boldsymbol{v}(\boldsymbol{x})) \mathrm{d} \boldsymbol{x} \\
& =\frac{\left\|e_{r}(\boldsymbol{x})\right\|_{\infty}}{e_{0}^{\min }} \mathbf{v}^{\top} A_{0} \mathbf{v}
\end{aligned}
$$

and

$$
\begin{aligned}
\mathbf{v}^{\top} A_{r} \mathbf{v} & =\int_{D} e_{r}(\boldsymbol{x}) \boldsymbol{\epsilon}(\boldsymbol{v}(\boldsymbol{x})): \boldsymbol{\epsilon}(\boldsymbol{v}(\boldsymbol{x})) \mathrm{d} \boldsymbol{x} \\
& \geqslant \frac{e_{r}^{\min }}{e_{0}^{\max }} \int_{D} e_{0}(\boldsymbol{x}) \boldsymbol{\epsilon}(\boldsymbol{v}(\boldsymbol{x})): \boldsymbol{\epsilon}(\boldsymbol{v}(\boldsymbol{x})) \mathrm{d} \boldsymbol{x} \\
& =\frac{e_{r}^{\min }}{e_{0}^{\max }} \mathbf{v}^{\top} A_{0} \mathbf{v}
\end{aligned}
$$

Therefore, the eigenvalues of $A_{0}^{-1} A_{r}$ lie in the interval

$$
\begin{equation*}
\left[\frac{e_{r}^{\min }}{e_{0}^{\max }}, \frac{\left\|e_{r}(\boldsymbol{x})\right\|_{\infty}}{e_{0}^{\min }}\right] . \tag{3.12}
\end{equation*}
$$

Similarly, we can prove that the eigenvalues of $D_{0}^{-1} D_{r}$ also lie in the interval (3.12). The last diagonal block in $K_{S}^{-1} K_{r}$ is a zero matrix with zero eigenvalues. Hence, the eigenvalues of $\hat{K}_{r}$ lie in the set in (3.11).

In particular, for the separable exponential covariance in Test problem $1, e_{1}(\boldsymbol{x})$ is strictly positive. Therefore, the eigenvalues of $\widehat{K}_{1}$ for Test problem 1 lie in (3.11).

The basis vectors in the iterative solution phase are selected based on the left matrices in (3.6). First, we need to make sure all the left matrices are non-singular. To achieve this, we choose $M+1$ shifts $\alpha_{r}$, one for each $\widehat{K}_{r}$ in (3.6), to make $\widehat{K}_{r}+\alpha_{r} I$ positive definite. Ideally, we would like all the spectra for each $r$ to lie in an interval
bounded away from zero. Adding and subtracting terms in (3.6) and rearranging gives

$$
\begin{equation*}
\widehat{X}\left(-\alpha_{0} I-\sum_{r=1}^{M} \alpha_{r} G_{r}\right)+\sum_{r=0}^{M}\left(\widehat{K}_{r}+\alpha_{r} I\right) \widehat{X} G_{r}=\widehat{H} . \tag{3.13}
\end{equation*}
$$

Note that the new matrix equation (3.13) has $M+2$ terms. The first left matrix is the identity matrix and the solutions of (3.13) and (3.6) are identical. Now, with appropriate shifts all the left matrices of the modified matrix equation are symmetric and positive definite. The second left matrix $\widehat{K}_{0}+\alpha_{0} I$ has three distinct real eigenvalues which lie in the following set

$$
\mathcal{T}_{0}:=\left\{\frac{1-\sqrt{5}}{2}+\alpha_{0}, 1+\alpha_{0}, \frac{1+\sqrt{5}}{2}+\alpha_{0}\right\} .
$$

For $r=1,2, \ldots, M$, the eigenvalues of the left matrices $\hat{K}_{r}+\alpha_{r} I$ are real and lie in the interval $\mathcal{T}_{r}:=\left[\alpha_{r}-\tau_{r}, \alpha_{r}+\tau_{r}\right]$, or else lie in

$$
\begin{equation*}
\mathcal{T}_{r}:=\left\{\alpha_{r}\right\} \cup\left[\frac{e_{r}^{\min }}{e_{0}^{\max }}+\alpha_{r}, \tau_{r}+\alpha_{r}\right] \tag{3.14}
\end{equation*}
$$

if $e_{r}(\boldsymbol{x})$ is strictly positive on $D$. We will discuss strategies for choosing the shifts $\alpha_{r}$ in Chapter 4.

Left Preconditioning: Left preconditioning can be used when the chosen preconditioner is a symmetric and indefinite saddle point matrix. Such a strategy was implemented for a matrix equation with a similar structure in the PhD thesis [58]. Choosing the non-singular matrix $K_{0}$ as the preconditioner and applying $K_{0}^{-1}$ to (3.1) on the left gives

$$
\begin{equation*}
X+\sum_{r=1}^{M} \widetilde{K}_{r} X G_{r}=\widetilde{H} \tag{3.15}
\end{equation*}
$$

where $\widetilde{K}_{r}$ and $\widetilde{H}$ are defined by

$$
\widetilde{K}_{r}:=K_{0}^{-1} K_{r}, \quad \widetilde{H}:=K_{0}^{-1} H=K_{0}^{-1} \mathbf{f g}_{0}^{\top}
$$

Again, note that $\widetilde{H}$ has one nonzero column because $\mathbf{g}_{0}$ has one nonzero entry. The modified matrix equation (3.15) has the advantage that the first left matrix is the identity matrix, but the other left matrices $\widetilde{K}_{r}$ are now non-symmetric. Preconditioners with a saddle-point structure such as $K_{0}$ have been studied in [42].

Recall that the matrices $K_{0}$ and $K_{r}$ are defined in (2.45)-(2.46). If we define

$$
N_{r}=\left[\begin{array}{cc}
\beta A_{r} & 0 \\
0 & \tilde{\lambda}^{-1} D_{r}
\end{array}\right] \in \mathbb{R}^{\left(2 n_{u}+n_{p}\right) \times\left(2 n_{u}+n_{p}\right)}, \quad r=0,1, \ldots, M,
$$

and

$$
W=\left[\begin{array}{ll}
B & -\tilde{\lambda}^{-1} C
\end{array}\right] \in \mathbb{R}^{n_{p} \times\left(2 n_{u}+n_{p}\right)},
$$

then, $K_{0}$ and $K_{r}$ for $r=1,2, \ldots, M$ can also be written as 2-by-2 block matrices of the form

$$
K_{0}=\left[\begin{array}{cc}
N_{0} & W^{\top} \\
W & 0
\end{array}\right], \quad K_{r}=\left[\begin{array}{cc}
N_{r} & 0 \\
0 & 0
\end{array}\right]
$$

Note that $N_{0}$ is symmetric and positive definite when $e_{0}(\boldsymbol{x})$ is positive, which we always assume, and $W$ has full row rank if the inf-sup condition (2.24) is satisfied.

Eigenvalue bounds for the left preconditioned matrices $\widetilde{K}_{r}$ in (3.15) are given in the following lemma. This result uses analysis from [42].

Lemma 3.3. When Assumption 2.18 is astisfied, the eigenvalues for $\widetilde{K}_{r}$, for $r=$ $1,2, \ldots, M$, are real and lie in the interval $\mathcal{S}_{r}$ in (3.8). If $e_{r}(\boldsymbol{x})$ is strictly positive, then the eigenvalues for $\widetilde{K}_{r}$ lie in the set defined in (3.11).

Proof. Any eigenvalue $\kappa \in \mathbb{C}$ and corresponding eigenvector $\left(\mathbf{a}^{\top}, \mathbf{b}^{\top}\right)^{\top} \in \mathbb{C}^{n_{x}}$ of $\widetilde{K}_{r}$, for $r=1, \ldots, M$, satisfy

$$
\left[\begin{array}{cc}
N_{r} & 0  \tag{3.16}\\
0 & 0
\end{array}\right]\left[\begin{array}{l}
\mathbf{a} \\
\mathbf{b}
\end{array}\right]=\kappa\left[\begin{array}{cc}
N_{0} & W^{\top} \\
W & 0
\end{array}\right]\left[\begin{array}{l}
\mathbf{a} \\
\mathbf{b}
\end{array}\right]
$$

where $\mathbf{a} \in \mathbb{C}^{2 n_{u}+n_{p}}$, and $\mathbf{b} \in \mathbb{C}^{n_{p}}$. The matrix $W^{\top}$ can be decomposed as

$$
W^{\top}=\left[\begin{array}{ll}
Y & Z
\end{array}\right]\left[\begin{array}{l}
R \\
0
\end{array}\right]
$$

where

$$
\left[\begin{array}{ll}
Y & Z
\end{array}\right] \in \mathbb{R}^{\left(2 n_{u}+n_{p}\right) \times\left(2 n_{u}+n_{p}\right)}
$$

is an unitary matrix, $Y \in \mathbb{R}^{\left(2 n_{u}+n_{p}\right) \times n_{p}}, Z \in \mathbb{R}^{\left(2 n_{u}+n_{p}\right) \times 2 n_{u}}$ is a basis for the null space of $W^{\top}$ and $R \in \mathbb{R}^{n_{p} \times n_{p}}$ is an upper triangular matrix. The vector a can be expressed as

$$
\mathbf{a}=Y \mathbf{a}_{y}+Z \mathbf{a}_{z},
$$

where $\mathbf{a}_{y} \in \mathbb{C}^{n_{p}}$, and $\mathbf{a}_{z} \in \mathbb{C}^{2 n_{u}}$. We can then rewrite the eigenvector $\left(\mathbf{a}^{\top}, \mathbf{b}^{\top}\right)^{\top}$ as

$$
\left[\begin{array}{ccc}
Z & Y & 0  \tag{3.17}\\
0 & 0 & I_{n_{p}}
\end{array}\right]\left[\begin{array}{c}
\mathbf{a}_{z} \\
\mathbf{a}_{y} \\
\mathbf{b}
\end{array}\right]
$$

where the matrix on the left is a unitary matrix of size $n_{x} \times n_{x}$ where $n_{x}=2\left(n_{u}+n_{p}\right)$. If we premultiply (3.16) by its transpose

$$
\left[\begin{array}{cc}
Z^{\top} & 0 \\
Y^{\top} & 0 \\
0 & I_{n_{p}}
\end{array}\right]
$$

and substitute (3.17) into (3.16), then we obtain the following eigenvalue problem

$$
\left[\begin{array}{ccc}
Z^{\top} N_{r} Z & Z^{\top} N_{r} Y & 0  \tag{3.18}\\
Y^{\top} N_{r} Z & Y^{\top} N_{r} Y & 0 \\
0 & 0 & 0
\end{array}\right]\left[\begin{array}{c}
\mathbf{a}_{z} \\
\mathbf{a}_{y} \\
\mathbf{b}
\end{array}\right]=\kappa\left[\begin{array}{ccc}
Z^{\top} N_{0} Z & Z^{\top} N_{0} Y & 0 \\
Y^{\top} N_{0} Z & Y^{\top} N_{0} Y & R \\
0 & R^{\top} & 0
\end{array}\right]\left[\begin{array}{c}
\mathbf{a}_{z} \\
\mathbf{a}_{y} \\
\mathbf{b}
\end{array}\right]
$$

Reordering the rows and columns of both matrices in (3.18) simultaneously gives two block lower triangular matrices

$$
K_{r}^{*}:=\left[\begin{array}{ccc}
0 & 0 & 0 \\
Z^{\top} N_{r} Y & Z^{\top} N_{r} Z & 0 \\
Y^{\top} N_{r} Y & Y^{\top} N_{r} Z & 0
\end{array}\right], \quad K_{0}^{*}:=\left[\begin{array}{ccc}
R^{\top} & 0 & 0 \\
Z^{\top} N_{0} Y & Z^{\top} N_{0} Z & 0 \\
Y^{\top} N_{0} Y & Y^{\top} N_{0} Z & R
\end{array}\right],
$$

for $r=1,2, \ldots, M$. The matrix $\widetilde{K}_{r}$ is similar to the lower block triangular matrix $K_{0}^{*-1} K_{r}^{*}$

$$
K_{0}^{*-1} K_{r}^{*}:=\left[\begin{array}{ccc}
0 & 0 & 0  \tag{3.19}\\
\Omega & \left(Z^{\top} N_{0} Z\right)^{-1}\left(Z^{\top} N_{r} Z\right) & 0 \\
\Phi & \Theta & 0
\end{array}\right]
$$

where $\Omega \in \mathbb{R}^{2 n_{u} \times n_{p}}, \Phi \in \mathbb{R}^{n_{p} \times n_{p}}$, and $\Theta \in \mathbb{R}^{n_{p} \times 2 n_{u}}$ are irrelevant for investigating the eigenvalues of $K_{0}^{*-1} K_{r}^{*}$. Therefore, the matrix $K_{0}^{*-1} K_{r}^{*}$ has,
(i) zero eigenvalue with multiplicity $2 n_{p}$,
(ii) $2 n_{u}$ eigenvalues defined by the eigenvalue problem $\left(Z^{\top} N_{r} Z\right) \mathbf{w}=\kappa\left(Z^{\top} N_{0} Z\right) \mathbf{w}$.

Since $N_{0}$ is symmetric and positive definite, $N_{r}$ is symmetric, and $Z$ is of full column rank, we can easily prove that $Z^{\top} N_{0} Z$ is symmetric and positive definite, and $Z^{\top} N_{r} Z$ is symmetric. In this case, the eigenvalues of $\left(Z^{\top} N_{0} Z\right)^{-1}\left(Z^{\top} N_{r} Z\right)$ are real. We can obtain eigenvalue bounds for $\left(Z^{\top} N_{0} Z\right)^{-1}\left(Z^{\top} N_{r} Z\right)$ by investigating bounds for the Rayleigh quotient

$$
\frac{\mathbf{d}^{\top} N_{r} \mathbf{d}}{\mathbf{d}^{\top} N_{0} \mathbf{d}}
$$

for any non-zero $\mathbf{d} \in \operatorname{Null}(W)$. This indicates that the eigenvalues of $\left(Z^{\top} N_{0} Z\right)^{-1}\left(Z^{\top} N_{r} Z\right)$ lie in a subset of the set that contains eigenvalues of $N_{0}^{-1} N_{r}$, which is a block-diagonal matrix

$$
N_{0}^{-1} N_{r}=\left[\begin{array}{cc}
A_{0}^{-1} A_{r} & 0 \\
0 & D_{0}^{-1} D_{r}
\end{array}\right]
$$

Since we know that the eigenvalues of $A_{0}^{-1} A_{r}$ and $D_{0}^{-1} D_{r}$ lie in the interval (3.8), the eigenvalues of $N_{0}^{-1} N_{r}$ must also lie in the same interval (3.8). Therefore, the eigenvalues of $\left(Z^{\top} N_{0} Z\right)^{-1}\left(Z^{\top} N_{r} Z\right)$ also lie in the interval (3.8). The eigenvalues of $\widetilde{K}_{r}$ are identical to those of $K_{0}^{*-1} K_{r}^{*}$, and so the eigenvalues of $\widetilde{K}_{r}$ are real and lie in (3.8). If $e_{r}(\boldsymbol{x})$ is strictly positive, then the eigenvalues of $\widetilde{K}_{r}$ lie in the set defined in (3.11).

Note that the eigenvalues of $\widetilde{K}_{r}$ lie in the same interval as $\widehat{K}_{r}$ for the symmetric preconditioning strategy, for $r=1,2, \ldots, M$. The first left matrix $\widehat{K}_{0}$ in the symmetrically preconditioned matrix equation (3.6) has three distinct nonzero eigenvalues but the first matrix in the left preconditioned matrix equation (3.15) is the $n_{x} \times n_{x}$ identity matrix whose eigenvalues are all one.

Similarly to (3.13), after applying left-preconditioning to obtain (3.15) we now choose $M$ shifts $\alpha_{r}$ to ensure that the eigenvalues of the matrices $\widetilde{K}_{r}+\alpha_{r} I$ are positive for each $r=1,2, \ldots, M$. This gives

$$
\begin{equation*}
X\left(I-\sum_{r=1}^{M} \alpha_{r} G_{r}\right)+\sum_{r=1}^{M}\left(\widetilde{K}_{r}+\alpha_{r} I\right) X G_{r}=\widetilde{H} \tag{3.20}
\end{equation*}
$$

Note, however, that this is an $M+1$ term matrix equation like the original one in (3.1) whereas (3.13) is an $M+2$ term matrix equation. The eigenvalues of $\widetilde{K}_{r}+\alpha_{r} I$ lie in the interval $\mathcal{T}_{r}=\left[\alpha_{r}-\tau_{r}, \alpha_{r}+\tau_{r}\right]$, or else lie in $\mathcal{T}_{r}$ defined in (3.14) if $e_{r}(\boldsymbol{x})$ is strictly positive. We will discuss strategies for choosing the shifts $\alpha_{r}$ in Chapter 5.

### 3.2.2 Iterative solution phase

After preconditioning and adding shifts, we obtain a matrix equation of the form

$$
\begin{equation*}
U Q_{0}+\sum_{r=1}^{J} P_{r} U Q_{r}=T \tag{3.21}
\end{equation*}
$$

where $J=M+1, U:=L_{K_{S}}^{\top} X$ and $T:=L_{K_{S}}^{-1} H$ for the matrix equation (3.13) associated with the symmetric preconditioning strategy, and $J=M, U:=X$ and $T:=K_{0}^{-1} H$ for the matrix equation (3.20) associated with the left preconditioning strategy. Note that $T$ has one nonzero column and $Q_{r}$ for $r=0,1, \ldots, J$ are symmetric in both cases. The iterative solution phase involves constructing basis vectors for the columns of the approximation $U_{k}$, applying a projection and solving a reduced problem.

Our aim is to solve (3.21) iteratively. At the $k$ th iteration, we define an approximation of the form

$$
\begin{equation*}
U_{k}=V_{k} Y_{k}, \tag{3.22}
\end{equation*}
$$

where $V_{k} \in \mathbb{R}^{n_{x} \times n_{k}}$ is orthogonal, $Y_{k} \in \mathbb{R}^{n_{k} \times n_{y}}$ and we aim to have $n_{k} \ll n_{x}$. Recall that $n_{x}=2\left(n_{u}+n_{p}\right)$ is the dimension of the spatial approximation associated with the three solution fields. If we define $U_{k}$ as in (3.22), each column of $U_{k}$ belongs to the space spanned by the columns of $V_{k}$. The matrix $Y_{k}$ is determined using a projection technique.

The strategy for constructing basis vectors is inspired by the rational Krylov subspace method for the two-term matrix equation

$$
\begin{equation*}
X G_{0}+K_{1} X G_{1}=H \tag{3.23}
\end{equation*}
$$

where $K_{1} \in \mathbb{R}^{n \times n}$ and $G_{0}, G_{1} \in \mathbb{R}^{m \times m}$ are some general matrices. We assume $G_{0}, G_{1}$ are nonsingular and apply $G_{0}^{-1}$ on the left of (3.23)

$$
\begin{equation*}
X+K_{1} X G_{1} G_{0}^{-1}=H G_{0}^{-1} \tag{3.24}
\end{equation*}
$$

and then write the eigenvalue decomposition for $G_{1} G_{0}^{-1}$

$$
G_{1} G_{0}^{-1}=Q \Lambda Q^{-1}
$$

We then apply $Q$ on the right of (3.24) and obtain

$$
\begin{equation*}
\tilde{X}+K_{1} \tilde{X} \Lambda=\widetilde{H} \tag{3.25}
\end{equation*}
$$

where $\widetilde{X}:=X Q$ and $\widetilde{H}:=H G_{0}^{-1} Q$. The exact solution $\tilde{X}$ to the matrix equation (3.25) can then be obtained by solving a set of linear systems

$$
\left(K_{1}+\frac{1}{\lambda_{i}} I\right) \tilde{\mathbf{x}}_{i}=\frac{1}{\lambda_{i}} \tilde{\mathbf{f}}_{i}, \quad i=1, \ldots, m,
$$

where $\widetilde{\mathbf{x}}_{i}$ is the $i$-th column of the exact solution $\widetilde{X}, \tilde{\mathbf{f}}_{i}$ is the $i$-th column of $\widetilde{F}$ and $\lambda_{i}$ for $i=1, \ldots, m$ are eigenvalues of $G_{1} G_{0}^{-1}$. We can approximate the solutions to these linear systems using rational Galerkin methods, which is applying the Galerkin projection onto rational Krylov subspaces, see [18, 20]. For the $i$ th linear system, we choose the rational Krylov subspace for approximating $\tilde{\mathbf{x}}_{i}$

$$
\mathcal{K}_{k}\left(K_{1}\right):=\left\{\left(K_{1}+s_{1} I\right)^{-1} \mathbf{v}_{0},\left(K_{1}+s_{2} I\right)^{-1}\left(K_{1}+s_{1} I\right)^{-1} \mathbf{v}_{0}, \ldots, \prod_{j=1}^{k}\left(K_{1}+s_{j} I\right)^{-1} \mathbf{v}_{0}\right\}
$$

where $\mathbf{v}_{0}$ are chosen to be normalised $\lambda_{i}^{-1} \widetilde{\mathbf{f}}_{i}$ and $s_{1}, \ldots, s_{k}$ is to be chosen. Suppose the columns of $V_{k}$ are a set of orthonormal basis vectors which span the rational Krylov subspace $\mathcal{K}_{k}\left(K_{1}\right)$. When we impose a Galerkin condition to determine the approximation for $\tilde{\mathbf{x}}_{i}$, the choice of $s_{1}, \ldots, s_{k}$ can be made by solving the Zolotarev minimax problem

$$
\min _{s_{1}, \ldots, s_{k}} \max _{\lambda \in \operatorname{spec}\left(K_{1}\right)}\left|\prod_{j=1}^{k} \frac{\lambda-s_{j}}{\lambda+s_{j}}\right|,
$$

where $\operatorname{spec}\left(K_{1}\right)$ is the spectrum of $K_{1}$.
For (3.21), we initialise $V_{0}=\mathbf{v}_{0}$, where $\mathbf{v}_{0}$ is usually chosen to be the first column of $T$. At the $k$ th iteration, we add at most $J$ new basis vectors to $V_{k-1}$ to construct $V_{k}$. First, we compute $J$ vectors of length $n_{x}$ and store them in a matrix $W_{k} \in \mathbb{R}^{n_{x} \times J}$ as follows

$$
\begin{equation*}
W_{k}:=\left[\left(P_{1}+s_{k, 1} I\right)^{-1} \mathbf{v}_{k-1}, \ldots,\left(P_{J}+s_{k, J} I\right)^{-1} \mathbf{v}_{k-1}\right] \tag{3.26}
\end{equation*}
$$

where $\mathbf{v}_{k-1}$ is the $(k-1)^{\text {st }}$ column of $V_{k-1}$ and $s_{k, r}$ for $r=1, \ldots, J$ are parameters to be chosen. Note that constructing $W_{k}$ requires the solution of $J$ shifted finite element systems with coefficient matrices of size $n_{x} \times n_{x}$. We discuss two strategies for selecting the parameters $s_{k, r}$ :
(i) Multi-parameter strategy - before the first iteration, we choose $n$ distinct parameters $s_{1, r}, s_{2, r}, \ldots, s_{n, r}$ for each matrix $P_{r}$ and cycle through these sets after
every $n$ iterations. These parameters could be obtained by solving the Zolotarev minimax problem, (see [72])

$$
\min _{s_{1, r}, s_{2, r}, \ldots, s_{n, r}} \max _{\lambda \in\left[\lambda_{\left.r^{\min }, \lambda_{r}^{\max }\right]}\right.}\left|\prod_{j=1}^{n} \frac{\lambda-s_{j, r}}{\lambda+s_{j, r}}\right|,
$$

where $\lambda_{r}^{\min }$ and $\lambda_{r}^{\max }$ are the minimum and maximum eigenvalues of $P_{r}$. Recall that we derived bounds for the eigenvalues of $P_{r}$ in Section 3.2.1.
(ii) Parameter-free strategy - choose one fixed parameter $s_{1, r}$ as the midpoint of the interval containing the eigenvalues of $P_{r}$, that is

$$
s_{1, r}=\frac{\lambda_{r}^{\min }+\lambda_{r}^{\max }}{2}
$$

This choice was suggested in [64].
Next, we orthonormalize the $J$ candidate basis vectors in $W_{k}$ with respect to the space spanned by the columns of $V_{k-1}$

$$
\widehat{W}_{k}=W_{k}-V_{k-1} V_{k-1}^{\top} W_{k}
$$

and truncate $\widehat{W}_{k}$ by keeping a fixed proportion $(\vartheta \%)$ of the most significant directions using a singular value decomposition. That is, we decompose $\widehat{W}_{k}$ as follows

$$
\widehat{W}_{k}=\sum_{i=1}^{J} \sigma_{i} \mathbf{u}_{i} \mathbf{v}_{i}^{\top}
$$

where $\sigma_{1} \geqslant \sigma_{2} \geqslant \ldots \geqslant \sigma_{J} \geqslant 0$ are the singular values of $\widehat{W}_{k}$, and $\mathbf{u}_{i} \in \mathbb{R}^{n_{x}}$ and $\mathbf{v}_{i} \in \mathbb{R}^{J}$ are singular vectors. We update $V_{k}$ by adding $\mathbf{u}_{1}, \mathbf{u}_{2}, \ldots, \mathbf{u}_{m_{k}}$ to $V_{k-1}$, where $m_{k}$ is the smallest number that satisfies

$$
\begin{equation*}
\frac{\sum_{i=1}^{m_{k}} \sigma_{i}}{\sum_{i=1}^{J} \sigma_{i}} \geqslant \vartheta \% . \tag{3.27}
\end{equation*}
$$

Typically, we choose $\vartheta=99$. Finally, we obtain the basis vectors at the $k$ th iteration

$$
V_{k}=\left[V_{k-1}, \mathbf{u}_{1}, \mathbf{u}_{2}, \ldots, \mathbf{u}_{m_{k}}\right]
$$

Next, we introduce two projection techniques to determine a good choice of $Y_{k}$ at each iteration. At the $k$ th iteration, the residual of the modified matrix equation (3.21) is

$$
\begin{equation*}
R_{k}:=U_{k} Q_{0}+\sum_{r=1}^{J} P_{r} U_{k} Q_{r}-T \tag{3.28}
\end{equation*}
$$

If we substitute the factored representation (3.22) into (3.28), we obtain

$$
\begin{equation*}
R_{k}:=V_{k} Y_{k} Q_{0}+\sum_{r=1}^{J} P_{r} V_{k} Y_{k} Q_{r}-T \tag{3.29}
\end{equation*}
$$

The first projection technique we consider is a Galerkin condition on the residual $R_{k}$, that is

$$
V_{k}^{\top} R_{k}=0,
$$

which gives the reduced problem

$$
\begin{equation*}
Y_{k} Q_{0}+\sum_{r=1}^{J}\left(V_{k}^{\top} P_{r} V_{k}\right) Y_{k} Q_{r}=V_{k}^{\top} T \tag{3.30}
\end{equation*}
$$

Notice that the size of the left matrices $V_{k}^{\top} P_{r} V_{k}$ is now $n_{k} \times n_{k}$, which is reduced compared with the original left matrices $P_{r}$. However, $V^{\top} P_{r} V_{k}$ is dense. The reduced basis problem (3.30) can also be rewritten as a linear system with Kronecker product structure

$$
\begin{equation*}
\left(Q_{0} \otimes I+\sum_{r=1}^{J} Q_{r} \otimes\left(V_{k}^{\top} P_{r} V_{k}\right)\right) \mathbf{y}_{k}=\operatorname{vec}\left(V_{k}^{\top} H\right) \tag{3.31}
\end{equation*}
$$

where $\mathbf{y}_{k}:=\operatorname{vec}\left(Y_{k}\right)$. The size of the coefficient matrix in (3.31) is $n_{k} n_{y} \times n_{k} n_{y}$, and $\mathbf{y}_{k} \in$ $\mathbb{R}^{n_{k} n_{y}}$. For the symmetric preconditioning strategy, the coefficient matrix is symmetric and so MINRES can be applied to solve (3.31). For the left preconditioning strategy, the coefficient matrix is non-symmetric and we can use Quasi-MINRES (QMR) to solve (3.31), see [29, 33].

In [64], the Galerkin condition is applied for the matrix equation considered there and in that case convergence is guaranteed because the associated linear system has a symmetric and positive definite coefficient matrix $\mathcal{M}$. Enforcing the Galerkin condition ensures that the $\mathcal{M}$-norm of the residual is minimised at each iteration. However, the coefficient matrix of the linear system for our problem is indefinite, and so does not induce a matrix norm. Applying a Galerkin condition does not guarantee convergence. Since the computational cost of applying a Galerkin condition is lower than that of a Petrov-Galerkin condition, we still consider applying the Galerkin condition for our problem.

The second projection technique we consider is a Petrov-Galerkin condition on the residual $R_{k}$. This was discussed in [60] and applied to much smaller matrix equations
than the ones considered here. To describe this, we first write the matrix equation (3.21) as the linear system

$$
F \mathbf{u}=\mathbf{t}
$$

where $\mathbf{u}=\operatorname{vec}(U), \mathbf{t}=\operatorname{vec}(T)$ and $F \in \mathbb{R}^{n_{x} n_{y} \times n_{x} n_{y}}$ is defined as

$$
F=\sum_{r=0}^{J} Q_{r} \otimes P_{r},
$$

with $P_{0}:=I_{n_{x}}$ and $Q_{r}$ symmetric for all $r$. We then define an operator $\mathcal{F}(\cdot)$ associated with $F$ as follows:

$$
\mathcal{F}(W):=W Q_{0}+\sum_{r=1}^{J} P_{r} W Q_{r}, \quad W \in \mathbb{R}^{n_{x} \times n_{y}}
$$

We then have $\mathcal{F}(W)=\operatorname{array}(F \mathbf{w})$ where $\mathbf{w}=\operatorname{vec}(W)$. The transpose of the operator $\mathcal{F}^{\top}(\cdot)$ is defined associated with $F^{\top}$, which is

$$
\mathcal{F}^{\top}(W):=W Q_{0}+\sum_{r=1}^{J} P_{r}^{\top} W Q_{r}, \quad W \in \mathbb{R}^{n_{x} \times n_{y}}
$$

The Petrov-Galerkin condition is

$$
V_{k}^{\top} \mathcal{F}^{\top}\left(R_{k}\right)=0,
$$

which can be expanded as follows

$$
\begin{equation*}
\sum_{i=0}^{J} \sum_{r=0}^{J}\left(V_{k}^{\top} P_{i}^{\top} P_{r} V_{k}\right) Y_{k} Q_{r} Q_{i}=\sum_{i=0}^{J} V_{k}^{\top} P_{i}^{\top} T Q_{i} \tag{3.32}
\end{equation*}
$$

where $P_{0}:=I_{n_{x}}$. This can be rewritten as the linear system

$$
\begin{equation*}
\sum_{i=0}^{J} \sum_{r=0}^{J}\left(Q_{r} Q_{i}\right) \otimes\left(V_{k}^{\top} P_{i}^{\top} P_{r} V_{k}\right) \mathbf{y}_{k}=\sum_{i=0}^{J}\left(Q_{i} \otimes V_{k}^{\top} P_{i}^{\top}\right) \operatorname{vec}(T) \tag{3.33}
\end{equation*}
$$

Since the coefficient matrix on the left hand side of (3.33) can be expressed as

$$
\left(I_{n_{y}} \otimes V_{k}\right)^{\top} F^{\top} F\left(I_{n_{y}} \otimes V_{k}\right)
$$

it is symmetric and positive definite and so we can use the conjugate gradient method to solve (3.32) for $\mathbf{y}_{k}$. Applying a Petrov-Galerkin condition at each iteration minimises the Frobenius norm of $R_{k}$. The problem size is also reduced compared to the matrix equation (3.21) but the computational cost increases compared to solving the one
associated with the Galerkin condition. There are $J+1$ terms in the reduced problem (3.30) when we apply the Galerkin condition, the computational cost and the storage cost depend on the value of $J$. For the Petrov-Galerkin condition, there are $(J+1)^{2}$ terms in the reduced problem (3.32) and the computational cost and the storage cost depend on the value of $J^{2}$. We will apply this projection technique to the matrix equation (3.21) associated with the symmetric preconditioning strategy in Chapter 4 for some small problems.

### 3.2.3 Stopping criteria

Usually, we want to control the iterative error in a norm that is compatible with the norm of the discretisation error. For example, see [76]. The natural stopping condition for iterative projection methods is when

$$
\frac{\left\|R_{k}\right\|_{F}}{\left\|R_{0}\right\|_{F}}<\epsilon
$$

where $\epsilon$ is a user-chosen tolerance and $R_{k}$ is the $k$ th preconditioned relative residual defined in (3.28). However, storing the dense and large matrix $R_{k} \in \mathbb{R}^{n_{x} \times n_{y}}$ is memoryconsuming. In some situations, a low-rank representation of $R_{k}$ can be considered, see [15, 47]. The residual at the $k$ th iteration can also be written as

$$
R_{k}=R_{k L} R_{k R}^{\top}
$$

where $R_{k L} \in \mathbb{R}^{n_{x} \times\left(J n_{k}+n_{k}+1\right)}$ and $R_{k R} \in \mathbb{R}^{n_{y} \times\left(J n_{k}+n_{k}+1\right)}$ are defined by

$$
\begin{aligned}
R_{k L} & :=\left[V_{k}, P_{1} V_{k}, \ldots, P_{J} V_{k},-\tilde{\mathbf{f}}\right] \\
R_{k R} & :=\left[Q_{0} Y_{k}^{\top}, Q_{1} Y_{k}^{\top}, \ldots, Q_{J} Y_{k}^{\top}, \mathbf{g}_{0}\right] .
\end{aligned}
$$

Here, $\widetilde{\mathbf{f}}:=L_{K_{S}}^{-1} \mathbf{f}$ for the symmetric preconditioning strategy or $\widetilde{\mathbf{f}}:=K_{0}^{-1} \mathbf{f}$ for the left preconditioning strategy and $\widetilde{\mathbf{f}}_{0}^{\top}=T$. If $J n_{k}$ is small, we can compute the QR decomposition of $R_{k L}$ and $R_{k R}$

$$
R_{k L}=Q_{k_{1}} R_{k 1}, \quad R_{k R}=Q_{k_{2}} R_{k 2}
$$

where $Q_{k_{1}} \in \mathbb{R}^{n_{x} \times\left(J n_{k}+n_{k}+1\right)}$ and $Q_{k_{2}} \in \mathbb{R}^{n_{y} \times\left(J n_{k}+n_{k}+1\right)}$ are orthogonal and $R_{k 1} \in$ $\mathbb{R}^{\left(J n_{k}+n_{k}+1\right) \times\left(J n_{k}+n_{k}+1\right)}$ and $R_{k 2} \in \mathbb{R}^{\left(J n_{k}+n_{k}+1\right) \times\left(J n_{k}+n_{k}+1\right)}$ are upper triangular matrices.

The Frobenius norm of the residual can then be written as

$$
\begin{aligned}
\left\|R_{k}\right\|_{F} & =\sqrt{\operatorname{trace}\left(R_{k}^{\top} R_{k}\right)} \\
& =\sqrt{\operatorname{trace}\left(Q_{k 2} R_{k 2} R_{k 1}^{\top} Q_{k 1}^{\top} Q_{k_{1}} R_{k 1} R_{k 2}^{\top} Q_{k 2}^{\top}\right)} \\
& =\sqrt{\operatorname{trace}\left(R_{k 2} R_{k 1}^{\top} R_{k 1} R_{k 2}^{\top}\right)} .
\end{aligned}
$$

Since $J n_{k}$ is small, we then only need to compute a small matrix of size $\left(J n_{k}+n_{k}+\right.$ 1) $\times\left(J n_{k}+n_{k}+1\right)$ to obtain the Frobenius norm of $R_{k}$. However, this method will be expensive and memory-consuming if $J n_{k}$ is not small, which is usually the case in our problems.

A cheaper but less stringent stopping criterion that can be used when $J n_{k}$ is large is the successive relative difference

$$
\begin{equation*}
\Delta_{k}:=\frac{\left\|U_{k}-U_{k-1}\right\|}{\left\|U_{k}\right\|_{F}}=\frac{\left\|Y_{k}-\left[Y_{k-1} ; 0\right]\right\|_{F}}{\left\|Y_{k}\right\|_{F}} . \tag{3.34}
\end{equation*}
$$

For a chosen tolerance $\epsilon$, the iteration is terminated when $\Delta_{k}<\epsilon$. Since we never want to compute and store any vectors or matrices of size $n_{x} \times n_{y}$, this stopping criterion avoids the need to compute and store residuals of size $n_{x} \times n_{y}$. This stopping criterion is also used in [64].

### 3.2.4 Multi-RB algorithm

We now present the Multi-term reduced basis method for the matrix equation (3.21) obtained after the pre-processing phase, in Algorithm 1.

In Algorithm 1, we need to store the matrices $V_{k}^{\top} P_{r} V_{k}$ for $r=1, \ldots, J$ if using the Galerkin condition or $V_{k}^{\top} P_{i}^{\top} P_{r} V_{k}$ for $i=1, \ldots, J$ and $r=1, \ldots, J$ if using the PetrovGalerkin condition. We use standard Krylov subspace methods to solve the reduced problems. The initial guess chosen for the reduced problem at the $k$ th iteration is $\left[Y_{k-1}, 0\right]$.

```
Algorithm 1 Multi-term reduced basis method (Multi-RB)
    Input: the matrices \(P_{r}\) for \(r=1,2, \ldots, J, Q_{r}\) for \(r=0,1, \ldots, J\), vectors \(\tilde{\mathbf{f}}\) and \(\mathbf{g}_{0}\),
    parameters \(s_{i, r}\) with \(i=1, \ldots, n\) and \(r=1,2, \ldots, J\), a tolerance \(\epsilon\) and a percentage
    \(\vartheta \%\) for truncating the basis vectors.
    Set \(\mathbf{v}_{0}=\tilde{\mathbf{f}} /\|\tilde{\mathbf{f}}\|, V_{1}=\mathbf{v}_{0}, \Delta_{K}=1, i=1\) and \(k=1\)
    while \(\Delta_{k} \geqslant \epsilon\) do
    \(W_{k}:=\left[\left(P_{1}+s_{i, 1} I\right)^{-1} \mathbf{v}_{k-1}, \ldots,\left(P_{J}+s_{i, J} I\right)^{-1} \mathbf{v}_{k-1}\right]\)
    \(\widehat{W}_{k}=W_{k}-V_{k-1} V_{k-1}^{\top} W_{k}\)
        cant directions to expand \(V_{k}=\left[V_{k-1}, \mathbf{u}_{1}, \mathbf{u}_{2}, \ldots, \mathbf{u}_{m_{k}}\right]\)
        approximately for \(Y_{k}\)
        Compute \(\Delta_{k}\) in (3.34)
        \(\mathbf{v}_{k}=V_{k} \mathbf{e}_{k+1}\) which is the \((k+1)\) st column of \(V_{k}\)
        \(k=k+1\)
        \(i=i+1\)
        if \(i>n\) then
            \(i=1\)
        end if
    end while
    Output: solution factors \(V_{k}\) and \(Y_{k}\).
```

        Decompose \(\widehat{W}_{k}\) using SVD \(\widehat{W}_{k}=\sum_{i=1}^{J} \sigma_{i} \mathbf{u}_{i} \mathbf{v}_{i}^{\top}\) and keep \(\vartheta \%\) of the most signifi-
        Solve (3.30) using Galerkin condition or (3.32) using Petrov-Galerkin condition
    In line 4, we need the matrices $P_{r}$ to build the basis vectors and set up the reduced problems. Note that we do not need to explicitly form these matrices. For example, we pass $\widehat{P}_{r}:=K_{r}+\alpha_{r} K_{S}$ and $L_{K_{S}}$ to the solver when using the symmetric preconditioning strategy and then $P_{r}=L_{K_{S}}^{-1} \widehat{P}_{r} L_{K_{S}}^{-\top}$. Let $\hat{U}_{k}:=\left[\mathbf{u}_{1}, \mathbf{u}_{2}, \ldots, \mathbf{u}_{m_{k}}\right]$ contain the new basis vectors at the $k$-th iteration, where $\mathbf{u}_{1}, \mathbf{u}_{2}, \ldots, \mathbf{u}_{m_{k}}$ are constructed on line 6 . The matrix $V_{k}^{\top} P_{r} V_{k}$ can then be built by the following procedure

$$
\begin{aligned}
V_{k}^{\top} P_{r} V_{k} & =\left[\begin{array}{c}
V_{k-1}^{\top} \\
\hat{U}_{k}^{\top}
\end{array}\right] L_{K_{S}}^{-1} \widehat{P}_{r} L_{K_{S}}^{-\top}\left[\begin{array}{ll}
V_{k-1} & \hat{U}_{k}
\end{array}\right] \\
& =\left[\begin{array}{cc}
V_{k-1}^{\top} L_{K_{S}}^{-1} \widehat{P}_{r} L_{K_{S}}^{-\top} V_{k-1} & V_{k-1}^{\top} L_{K_{S}}^{-1} \widehat{P}_{r} L_{K_{S}}^{-\top} \widehat{U}_{k} \\
\hat{U}_{k}^{\top} L_{K_{S}}^{-1} \widehat{P}_{r} L_{K_{S}}^{-\top} V_{k-1} & \hat{U}_{k}^{\top} L_{K_{S}}^{-1} \widehat{P}_{r} L_{K_{S}}^{-\top} \widehat{U}_{k}
\end{array}\right],
\end{aligned}
$$

where the $(1,1)$ block is from the $(k-1)^{\text {st }}$ iteration, and we only need to assemble the $(1,2)$ block and the $(2,2)$ block and the $(2,1)$ block is the transpose of $(1,2)$ block. To build $W_{k}$, we need to solve the systems

$$
\left(P_{r}+s_{i, r} I\right) \mathbf{z}=\mathbf{v}_{k-1} .
$$

We do this by solving the following systems:

$$
\left(\widehat{P}_{r}+s_{i, r} K_{S}\right) \hat{\mathbf{z}}=L_{K_{S}} \mathbf{v}_{k-1}
$$

and then obtain $\mathbf{z}$ by

$$
\mathbf{z}=L_{K_{S}}^{\top} \hat{\mathbf{z}} .
$$

In the next two chapters, we will apply Multi-RB to the parameter-dependent linear elasticity problem using the two suggested preconditioning strategies.

## Chapter 4

## Multi-RB with Symmetric

## Preconditioning

In this chapter, we apply Multi-RB to the multi-term matrix equations (3.21) associated with the symmetric preconditioning strategy discussed in Section 3.2.1 with the preconditioner $K_{S}$ defined in (2.54) for the two test problems defined in Section 2.3.2. Recall that the shifted symmetrically preconditioned matrix equation is given by (3.13). The left matrices $P_{r}$ in (3.21) with the symmetric preconditioning strategy are defined as

$$
P_{r}=\widehat{K}_{r-1}+\alpha_{r-1} I, \quad r=1,2, \ldots, M+1,
$$

where the shifts $\alpha_{r}$ for $r=0,1, \ldots, M$ need to be chosen based on eigenvalue bounds of

$$
\widehat{K}_{r}:=L_{K_{S}}^{-1} K_{r} L_{K_{S}}^{-\top}, \quad r=0,1, \ldots, M,
$$

to ensure that the matrices $P_{r}$ for $r=1,2, \ldots, M+1$ are non-singular. Theoretical eigenvalue bounds for $\widehat{K}_{r}$ were given in Lemma 3.1.

### 4.1 Eigenvalues and choice of shifts

In order to select shifts for $\widehat{K}_{r}$ for $r=0,1, \ldots, M$, we now investigate the eigenvalues of all these matrices for Test problems 1-2. As discussed in Section 3.2.1, the first preconditioned left matrix $\hat{K}_{0}$ has three distinct eigenvalues in the set $\mathcal{S}_{0}$ defined in
(3.7). The eigenvalues of $\widehat{K}_{r}$, for $r=1,2, \ldots, M$, lie in the interval $\mathcal{S}_{r}$ defined in (3.8) when $e_{r}(\boldsymbol{x})$ has both positive and negative values or in (3.11) when $e_{r}(\boldsymbol{x})$ is strictly positive. Since $e_{0}(\boldsymbol{x})=1$ for the two chosen test problems, we have

$$
\mathcal{S}_{r}=\left[-\left\|e_{r}(\boldsymbol{x})\right\|_{\infty},\left\|e_{r}(\boldsymbol{x})\right\|_{\infty}\right],
$$

when $e_{r}(\boldsymbol{x})$ has both positive and negative values. However, the function $e_{1}(\boldsymbol{x})$ in Test problem 1 is strictly positive on the spatial domain $D$ and so in that case the eigenvalues of $\hat{K}_{1}$ lie in

$$
\mathcal{S}_{1}=\{0\} \cup\left[e_{1}^{\min },\left\|e_{1}(\boldsymbol{x})\right\|_{\infty}\right] .
$$

The bounds for the eigenvalues of the left matrices are independent of the finite element mesh size $h$, the Poisson ratio $\nu$ and the polynomial degree $d$.

Next, we compute the theoretical eigenvalue bounds for $\widehat{K}_{r}$ for $r=1, \ldots, M$ using numerically computed values of $\left\|e_{r}(\boldsymbol{x})\right\|_{\infty}$ on a spatial mesh on $D$ with grid level 6 . For Test problem 1 with $\sigma=0.085$ and $\sigma=0.17$, the computed theoretical eigenvalue bounds for $\widehat{K}_{r}$ for $r=1,2, \ldots, 7$ are recorded in Table 4.1. We see that the interval $\mathcal{S}_{r}$ contracts to zero as $r \rightarrow \infty$. This is because the eigenvalues $\lambda_{r}$ associated with the separable exponential covariance function (2.15) decay.

| $r$ | $\sigma=0.085$ | $\sigma=0.17$ |
| :---: | :---: | :---: |
| 1 | $\{0\} \cup[0.0789,0.1251]$ | $\{0\} \cup[0.1578,0.2502]$ |
| 2 | $[-0.0668,0.0668]$ | $[-0.1337,0.1337]$ |
| 3 | $[-0.0668,0.0668]$ | $[-0.1337,0.1337]$ |
| 4 | $[-0.0399,0.0399]$ | $[-0.0797,0.0797]$ |
| 5 | $[-0.0399,0.0399]$ | $[-0.0797,0.0797]$ |
| 6 | $[-0.0357,0.0357]$ | $[-0.0714,0.0714]$ |
| 7 | $[-0.0277,0.0277]$ | $[-0.0555,0.0555]$ |

Table 4.1: Computed theoretical eigenvalue bounds for $\hat{K}_{r}$, for $r=1,2, \ldots, 7$, in Test problem 1.

For Test problem 2, the computed theoretical eigenvalue bounds for $\widehat{K}_{r}$ for $r=$ $1,2, \ldots, 5$ are recorded in Table 4.2. Again, the intervals $\mathcal{S}_{r}$ contract to zero as $r \rightarrow \infty$ and in the fast decay case much more rapidly than in the slow decay case.

| $r$ | slow decay | fast decay |
| :---: | :---: | :---: |
| 1 | $[-0.5470,0.5470]$ | $[-0.8320,0.8320]$ |
| 2 | $[-0.1368,0.1368]$ | $[-0.0520,0.0520]$ |
| 3 | $[-0.0608,0.0608]$ | $[-0.0103,0.0103]$ |
| 4 | $[-0.0342,0.0342]$ | $[-0.0032,0.0032]$ |
| 5 | $[-0.0219,0.0219]$ | $[-0.0013,0.0013]$ |

Table 4.2: Computed theoretical eigenvalue bounds for $\hat{K}_{r}$, for $r=1,2, \ldots, 5$, in Test problem 2.


Figure 4.1: Numerical eigenvalues of $\hat{K}_{r}$ for $r=1,2, \ldots, 7$ in Test problem 1, computed on a spatial mesh with grid level 5 .


Figure 4.2: Numerical eigenvalues of $\widehat{K}_{r}$ for $r=1,2, \ldots, 5$ in Test problem 2, computed on a spatial mesh with grid level 5 .

We now compute the eigenvalues of $\widehat{K}_{r}$ numerically for these two test problems on a coarse spatial mesh with grid level 5 using eigs in MATLAB. For Test problem 1 with $\sigma=0.085,0.17$, the numerical eigenvalues of $\widehat{K}_{r}$, for $r=1,2, \ldots, 7$, are shown in Figure 4.1. For Test problem 2, the numerical eigenvalues of $\widehat{K}_{r}$, for $r=1,2, \ldots, 5$, are shown in Figure 4.2. For both test problems, we compared the numerically computed external eigenvalues with the computed theoretical eigenvalue bounds in Tables 4.1-4.2 and found good agreement. Next, we choose the shifts $\alpha_{r}$ for $\hat{K}_{r}$, for $r=0,1, \ldots, M$, for the two test problems based on the theoretical eigenvalue bounds.

For $\widehat{K}_{0}$, we choose $\alpha_{0}=\sqrt{5} / 2$ so that the smallest eigenvalue of $P_{1}=\widehat{K}_{0}+\alpha_{0} I$ is $1 / 2$. Then $P_{1}$ has three distinct real eigenvalues which lie in the following set

$$
\left\{\frac{1}{2}, \frac{2+\sqrt{5}}{2}, \frac{1+2 \sqrt{5}}{2}\right\} .
$$

For $r=1,2, \ldots, M$, we choose

$$
\alpha_{r}=1-\frac{\lambda_{r}^{\min }+\lambda_{r}^{\max }}{2}
$$

where $\lambda_{r}^{\min }$ and $\lambda_{r}^{\max }$ represent the minimum and maximum of the eigenvalue bounds of $\widehat{K}_{r}$. In this case, the eigenvalues of $P_{r}=\widehat{K}_{r-1}+\alpha_{r-1} I$ for $r=2, \ldots, M+1$ are centred around 1. For Test problem 1, we choose $\alpha_{1}=\left\|e_{1}(\boldsymbol{x})\right\|_{\infty} / 2$ and $\alpha_{r}=1$ for $r=2,3, \ldots, M$. The computed theoretical eigenvalue bounds for $P_{r}$, for $r=$ $2, \ldots, M+1$, are recorded in Table 4.3 using numerically computed values of $\left\|e_{r}(\boldsymbol{x})\right\|_{\infty}$, for $r=1,2, \ldots, M$, on a spatial mesh on $D$ with grid level 6 .

| $r$ | $\sigma=0.085$ | $\sigma=0.17$ |
| :---: | :---: | :---: |
| 2 | $\{0.9374\} \cup[1.0163,1.0626]$ | $\{0.8749\} \cup[1.0327,1.1251]$ |
| 3 | $[0.9332,1.0668]$ | $[0.8663,1.1337]$ |
| 4 | $[0.9332,1.0668]$ | $[0.8663,1.1337]$ |
| 5 | $[0.9601,1.0399]$ | $[0.9203,1.0797]$ |
| 6 | $[0.9601,1.0399]$ | $[0.9203,1.0797]$ |
| 7 | $[0.9643,1.0357]$ | $[0.9286,1.0714]$ |
| 8 | $[0.9723,1.0277]$ | $[0.9445,1.0555]$ |

Table 4.3: Computed theoretical eigenvalue bounds for $P_{r}=\widehat{K}_{r-1}+\alpha_{r-1} I$, for $r=$ $2, \ldots, 7,8$, in Test problem 1.

For Test problem 2, we choose $\alpha_{r}=1$ for $r=1,2, \ldots, M$. The computed theoretical eigenvalue bounds for $P_{r}$, for $r=2, \ldots, M+1$, are recorded in Table 4.4.

| $r$ | slow decay | fast decay |
| :---: | :---: | :---: |
| 2 | $[0.4530,1.5470]$ | $[0.1680,1.8320]$ |
| 3 | $[0.8632,1.1368]$ | $[0.9480,1.0520]$ |
| 4 | $[0.9392,1.0608]$ | $[0.9897,1.0103]$ |
| 5 | $[0.9658,1.0342]$ | $[0.9968,1.0032]$ |
| 6 | $[0.9781,1.0219]$ | $[0.9987,1.0013]$ |

Table 4.4: Computed theoretical eigenvalue bounds for $P_{r}=\hat{K}_{r-1}+\alpha_{r-1} I$, for $r=$ $2, \ldots, 5,6$, in Test problem 2.

### 4.2 Numerical experiments

In this section, we apply Multi-RB to the symmetrically preconditioned matrix equation (3.21) for Test problems 1-2.

We choose $\epsilon_{\text {outer }}=10^{-6}$ to be the stopping tolerance for the outer iteration. That is, Multi-RB will be terminated when $\Delta_{k}<\epsilon_{\text {outer }}$ where $\Delta_{k}$ is the successive relative difference defined in (3.34). Unless otherwise stated, we use the parameter-free strategy and use the backslash operator in MATLAB to construct the basis vectors to form the columns of $W_{k}$ in Algorithm 1 and keep $99 \%$ of the basis vectors at each iteration. Recall that a reduced problem also needs to be solved at each iteration. If the Galerkin condition is imposed, the reduced problem is (3.30) and the Kronecker form is defined in (3.31). If the Petrov-Galerkin condition is applied, the reduced problem is (3.32) and the Kronecker form is (3.33). The solutions to these reduced problems are approximated using MINRES and we call this the inner iteration. We set $Y_{0}=0$ and choose $Y_{k}=\left[Y_{k-1} ; 0\right]$ as the initial guess when we apply MINRES to the reduced problem at the $k$-th outer iteration. Let $\delta_{k}^{(s)}$ be 2-norm of the relative residual at the $s$-th iteration of the reduced problem. The inner iteration is terminated when the following stopping condition is satisfied

$$
\begin{equation*}
\delta_{k}^{(s)}<\epsilon_{\text {inner }} \Delta_{k-1}, \tag{4.1}
\end{equation*}
$$

where $\epsilon_{\text {inner }}$ is a chosen tolerance. Here, we choose $\epsilon_{\text {inner }}=10^{-3}$.
For all numerical experiments, we record the number of outer iterations $k$, the dimension $n_{k}$ of the approximation space generated, the estimated rank of the final approximation $X_{k}$, the average inner iteration count $i$ and the timing in seconds. We also record the final unpreconditioned relative residual $\operatorname{Res}_{k}$

$$
\operatorname{Res}_{k}:=\frac{\left\|L_{K_{S}} R_{k}\right\|_{F}}{\left\|L_{K_{S}} R_{0}\right\|_{F}}
$$

where $R_{k}$ is the $k$-th residual of the preconditioned matrix equation (3.13) with the preconditioner $K_{S}$ and $L_{K_{S}} R_{k}$ is the $k$-th residual of the unpreconditioned matrix equation (3.1), as well as the final preconditioned relative residual $\operatorname{Pres}_{1, k}$

$$
\operatorname{Pres}_{1, k}:=\frac{\left\|R_{k}\right\|_{F}}{\left\|R_{0}\right\|_{F}} .
$$

The preconditioned relative residual Pres $_{1, k}$ for the symmetrically preconditioned matrix equation (3.13) is comparable to the preconditioned relative residual (2.52) of the symmetrically preconditioned linear system (2.51) with the preconditioner $\mathcal{P}=$ $\mathcal{P}_{1}$ when we apply preconditioned MINRES. Unless otherwise stated, we apply the Galerkin condition to the residual $R_{k}$ to obtain the reduced problem.

Experiment 1 (varying $M$ and $d$ ) First, we fix the grid level to be 6 (giving $\left.n_{x}=10242\right)$ and let $\nu=0.4999$ and apply Multi-RB to Test problems $1-2$ with varying $M$ and $d$. For these problems, we record the iteration counts and the timings (in brackets) in the last column when we apply preconditioned MINRES to the Kronecker form (2.44) with the preconditioner $\mathcal{P}_{1}=I_{n_{y}} \otimes K_{S}$ and we terminate MINRES iterations when the preconditioned relative residual defined in (2.52) is less than Pres $_{1, k}$. In Tables 4.5-4.6, we can see that the iteration counts $k$ and the dimension $n_{k}$ are independent of the polynomial degree $d$ but depend on the number of parameters $M$ and standard deviation $\sigma$ when we apply Multi-RB to the symmetrically preconditioned matrix equation for Test problem 1. The dimension $n_{k}$ is smaller than $n_{y}$ when $M$ and $d$ are not too small, but much larger than the estimated rank of the final approximation $X_{k}$. The final unpreconditioned relative residual $\operatorname{Res}_{k}$ is roughly one order of magnitude higher than the successive relative difference $\Delta_{k}$. The final preconditioned relative residual Pres $_{1, k}$ has a good agreement with the successive relative difference $\Delta_{k}$. Recall the stopping condition is based on $\Delta_{k}$. Compared with MINRES, we can
see that Multi-RB was quicker for problems with $d=5$ to achieve the same accuracy in terms of the preconditioned relative residual.

| $M$ | $d$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{1, k}$ | MINRES |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 3 | 56 | 63 | 290 | 56 | 9 | 1.10 e 2 | $3.92 \mathrm{e}-5$ | $6.73 \mathrm{e}-6$ | $15(3.52 \mathrm{e} 1)$ |
|  | 4 | 126 | 63 | 290 | 93 | 9 | 1.10 e 2 | $3.92 \mathrm{e}-5$ | $6.73 \mathrm{e}-6$ | $16(8.13 \mathrm{e} 1)$ |
|  | 5 | 252 | 63 | 290 | 93 | 9 | 8.21 e 1 | $3.92 \mathrm{e}-5$ | $6.73 \mathrm{e}-6$ | $17(1.81 \mathrm{e} 2)$ |
| 7 | 3 | 120 | 79 | 408 | 119 | 9 | 1.33 e 2 | $2.74 \mathrm{e}-5$ | $5.95 \mathrm{e}-6$ | $16(7.79 \mathrm{e} 1)$ |
|  | 4 | 330 | 79 | 408 | 166 | 9 | 1.36 e 2 | $2.74 \mathrm{e}-5$ | $5.95 \mathrm{e}-6$ | $17(2.26 \mathrm{e} 2)$ |
|  | 5 | 792 | 79 | 408 | 168 | 9 | 1.40 e 2 | $2.74 \mathrm{e}-5$ | $5.95 \mathrm{e}-6$ | $18(5.70 \mathrm{e} 2)$ |
| 13 | 3 | 560 | 154 | 1078 | 402 | 9 | 5.77 e 2 | $4.95 \mathrm{e}-5$ | $3.36 \mathrm{e}-6$ | $17(3.77 \mathrm{e} 2)$ |
|  | 4 | 2380 | 154 | 1078 | 476 | 9 | 8.04 e 2 | $4.95 \mathrm{e}-5$ | $3.36 \mathrm{e}-6$ | $18(1.68 \mathrm{e} 3)$ |
|  | 5 | 8568 | 154 | 1078 | 482 | 9 | 1.52 e 3 | $4.95 \mathrm{e}-5$ | $3.36 \mathrm{e}-6$ | $18(6.08 \mathrm{e} 3)$ |

Table 4.5: Numerical results for Test problem 1 with fixed grid level $6, \nu=0.4999$ and $\sigma=0.085$ using the symmetric preconditioning strategy; Multi-RB solver vs preconditioned MINRES.

| $M$ | $d$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{1, k}$ | MINRES |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 3 | 56 | 92 | 441 | 56 | 12 | 1.26 e 2 | $2.19 \mathrm{e}-5$ | $4.24 \mathrm{e}-6$ | $21(4.72 \mathrm{e} 1)$ |
|  | 4 | 126 | 92 | 441 | 124 | 12 | 1.28 e 2 | $2.29 \mathrm{e}-5$ | $4.38 \mathrm{e}-6$ | $24(1.20 \mathrm{e} 2)$ |
|  | 5 | 252 | 92 | 441 | 168 | 12 | 1.31 e 2 | $2.31 \mathrm{e}-5$ | $4.41 \mathrm{e}-6$ | $24(2.40 \mathrm{e} 2)$ |
| 7 | 3 | 120 | 100 | 571 | 120 | 11 | 1.82 e 2 | $3.27 \mathrm{e}-5$ | $1.10 \mathrm{e}-5$ | $21(9.74 \mathrm{e} 1)$ |
|  | 4 | 330 | 100 | 571 | 268 | 12 | 1.89 e 2 | $3.36 \mathrm{e}-5$ | $1.12 \mathrm{e}-5$ | $23(2.90 \mathrm{e} 2)$ |
|  | 5 | 792 | 100 | 571 | 310 | 12 | 2.45 e 2 | $3.37 \mathrm{e}-5$ | $1.12 \mathrm{e}-5$ | $24(7.52 \mathrm{e} 2)$ |
| 13 | 3 | 560 | 185 | 1437 | 510 | 11 | 1.14 e 3 | $1.90 \mathrm{e}-5$ | $4.93 \mathrm{e}-6$ | $23(5.05 \mathrm{e} 2)$ |
|  | 4 | 2380 | 185 | 1437 | 802 | 12 | 1.58 e 3 | $1.94 \mathrm{e}-5$ | $5.07 \mathrm{e}-6$ | $25(2.33 \mathrm{e} 3)$ |
|  | 5 | 8568 | 185 | 1437 | 888 | 12 | 3.03 e 3 | $1.95 \mathrm{e}-5$ | $5.10 \mathrm{e}-6$ | $27(8.92 \mathrm{e} 3)$ |

Table 4.6: Numerical results for Test problem 1 with fixed grid level $6, \nu=0.4999$ and $\sigma=0.17$ using the symmetric preconditioning strategy; Multi-RB solver vs standard MINRES.

| $M$ | $d$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{1, k}$ | MINRES |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 3 | 56 | 248 | 1234 | 56 | 18 | 4.96 e 2 | $1.24 \mathrm{e}-4$ | $7.13 \mathrm{e}-6$ | $34(7.35 \mathrm{e} 1)$ |
|  | 4 | 126 | 248 | 1234 | 121 | 21 | 5.32 e 2 | $1.48 \mathrm{e}-4$ | $8.55 \mathrm{e}-6$ | $36(1.72 \mathrm{e} 2)$ |
|  | 5 | 252 | 262 | 1314 | 186 | 25 | 6.52 e 2 | $1.14 \mathrm{e}-4$ | $6.46 \mathrm{e}-6$ | $39(3.74 \mathrm{e} 2)$ |
| 7 | 3 | 120 | 247 | 1444 | 120 | 19 | 7.44 e 2 | $9.88 \mathrm{e}-5$ | $1.07 \mathrm{e}-5$ | $33(1.52 \mathrm{e} 2)$ |
|  | 4 | 330 | 247 | 1444 | 245 | 23 | 9.20 e 2 | $9.88 \mathrm{e}-5$ | $1.07 \mathrm{e}-5$ | $36(4.52 \mathrm{e} 2)$ |
|  | 5 | 792 | 247 | 1444 | 318 | 27 | 1.24 e 3 | $1.04 \mathrm{e}-4$ | $1.09 \mathrm{e}-5$ | $37(1.14 \mathrm{e} 3)$ |
| 13 | 3 | 560 | 255 | 1852 | 383 | 30 | 3.54 e 3 | $8.59 \mathrm{e}-5$ | $7.33 \mathrm{e}-6$ | $34(7.34 \mathrm{e} 2)$ |
|  | 4 | 2380 | 258 | 1874 | 534 | 39 | 1.03 e 4 | $9.07 \mathrm{e}-5$ | $7.51 \mathrm{e}-6$ | $37(3.38 \mathrm{e} 3)$ |
|  | 5 | 8568 | 258 | 1874 | 602 | 48 | 3.60 e 4 | $9.35 \mathrm{e}-5$ | $7.64 \mathrm{e}-6$ | $39(1.30 \mathrm{e} 4)$ |

Table 4.7: Numerical results for the slow decay case in Test problem 2 with fixed grid level 6 and $\nu=0.4999$ using the symmetric preconditioning strategy; Multi-RB solver vs standard MINRES.

| $M$ | $d$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{1, k}$ | MINRES |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 3 | 56 | 357 | 1542 | 50 | 66 | 2.09 e 3 | $2.15 \mathrm{e}-4$ | $1.11 \mathrm{e}-5$ | $49(1.05 \mathrm{e} 2)$ |
|  | 4 | 126 | 337 | 1635 | 80 | 86 | 2.75 e 3 | $2.77 \mathrm{e}-4$ | $1.48 \mathrm{e}-5$ | $52(2.50 \mathrm{e} 2)$ |
|  | 5 | 252 | 600 | $\mathbf{2 7 3 5}$ | 114 | $\mathbf{1 4 3}$ | 1.95 e 4 | $1.25 \mathrm{e}-4$ | $6.00 \mathrm{e}-6$ | $61(9.41 \mathrm{e} 2)$ |
| 7 | 3 | 120 | 300 | 1366 | 76 | 49 | 1.57 e 3 | $1.77 \mathrm{e}-4$ | $9.47 \mathrm{e}-6$ | $49(2.88 \mathrm{e} 2)$ |
|  | 4 | 330 | 333 | 1552 | 115 | 59 | 2.56 e 3 | $1.80 \mathrm{e}-4$ | $8.99 \mathrm{e}-6$ | $55(8.88 \mathrm{e} 2)$ |
|  | 5 | 792 | 333 | 1552 | 141 | 76 | 3.45 e 3 | $2.45 \mathrm{e}-4$ | $1.24 \mathrm{e}-5$ | $58(1.73 \mathrm{e} 3)$ |
| 13 | 3 | 560 | 242 | 1138 | 118 | 52 | 1.80 e 3 | $3.27 \mathrm{e}-4$ | $2.12 \mathrm{e}-5$ | $46(9.78 \mathrm{e} 2)$ |
|  | 4 | 2380 | 280 | 1351 | 159 | 64 | 4.96 e 3 | $2.80 \mathrm{e}-4$ | $1.57 \mathrm{e}-5$ | $52(4.80 \mathrm{e} 3)$ |
|  | 5 | 8568 | 320 | 1617 | 191 | 74 | 2.14 e 4 | $2.27 \mathrm{e}-4$ | $1.24 \mathrm{e}-5$ | $58(2.19 \mathrm{e} 4)$ |

Table 4.8: Numerical results for the fast decay case in Test problem 2 with fixed grid level 6 and $\nu=0.4999$ using the symmetric preconditioning strategy; Multi-RB solver vs standard MINRES. Bold numbers indicate unexpectedly large iteration counts.

In Tables 4.7-4.8, we see that the number of outer iterations $k$ is nearly independent of $M$ for the slow decay case of Test problem 2 and it decreases as $M$ grows for the fast decay case of Test problem 2. Note that the relative difference $\Delta_{k}$ is not guaranteed to
decrease at each iteration. The final unpreconditioned relative residual $R e s_{k}$ is roughly two orders of magnitude high than the final successive relative difference $\Delta_{k}$. The final preconditioned relative residual Pres $_{1, k}$ has a good agreement with the successive relative difference $\Delta_{k}$ in most cases. The number of outer iterations becomes large for the fast decay case of Test problem 2 with $M=5$ and $d=5$ and so does the inner iteration count.

Experiment 2 (varying grid level $l$ ) We now apply Multi-RB to Test problem 1 with $\sigma=0.17$ and the slow decay case of Test problem 2 on finite element meshes with varying grid level $l=4,5,6,7$. We fix $\nu=0.4999, M=7$ and $d=4$ (giving $\left.n_{y}=330\right)$ and record the numerical results in Tables 4.9-4.10. We observe that the number of outer iterations $k$ and the dimension $n_{k}$ are nearly independent of grid level $l$ or equivalently, mesh size $h$.

| $l$ | $n_{x}$ | $k$ | $n_{k}$ | rank | $i$ | time $(\mathrm{s})$ | Res $_{k}$ | Pres $_{1, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 4 | 642 | 90 | 580 | 266 | 13 | 2.43 e 1 | $3.66 \mathrm{e}-6$ | $8.80 \mathrm{e}-7$ |
| 5 | 2562 | 100 | 612 | 272 | 14 | 4.96 e 1 | $2.33 \mathrm{e}-5$ | $1.06 \mathrm{e}-5$ |
| 6 | 10242 | 100 | 571 | 268 | 12 | 1.89 e 2 | $3.36 \mathrm{e}-5$ | $1.12 \mathrm{e}-5$ |
| 7 | 40962 | 108 | 587 | 258 | 13 | 3.59 e 3 | $1.60 \mathrm{e}-5$ | $3.60 \mathrm{e}-6$ |

Table 4.9: Numerical results for Test problem 1 with varying grid level $l$ when $\nu=0.4999, \sigma=0.17, M=7$ and $d=4$ using the symmetric preconditioning strategy.

| $l$ | $n_{x}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{1, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 4 | 642 | 99 | 642 | 214 | 19 | 3.59 e 1 | $3.53 \mathrm{e}-8$ | $5.18 \mathrm{e}-9$ |
| 5 | 2562 | 240 | 1513 | 250 | 32 | 9.50 e 2 | $6.69 \mathrm{e}-5$ | $6.94 \mathrm{e}-6$ |
| 6 | 10242 | 247 | 1444 | 245 | 23 | 9.20 e 2 | $9.88 \mathrm{e}-5$ | $1.07 \mathrm{e}-5$ |
| 7 | 40962 | 218 | 1172 | 242 | 17 | 7.52 e 3 | $1.39 \mathrm{e}-4$ | $1.64 \mathrm{e}-5$ |

Table 4.10: Numerical results for the slow decay case of Test problem 2 with varying grid level $l$ when $\nu=0.4999, M=7$ and $d=4$ using the symmetric preconditioning strategy.

Experiment 3 (varying Poisson ratio) We record the numerical results obtained from applying Multi-RB to Test problem 1 with $\sigma=0.17$ and the slow decay
case of Test problem 2 with varying Poisson ratio $\nu$ in Tables 4.11-4.12. We fix the grid level $l=6, M=7$ and total polynomial degree $d=4$. For both problems, the number of outer iterations $k$ and the dimension $n_{k}$ decrease as the Poisson ratio $\nu$ approaches 0.5 . The solver is robust in the incompressible limit. The final unpreconditioned relative residual $R e s_{k}$ is of one order of magnitude higher than the relative difference for Test problem 1 when $\sigma=0.17$ and two orders of magnitude higher than the relative difference for the slow decay case in Test problem 2.

| $\nu$ | $k$ | $n_{k}$ | rank | $i$ | time $(\mathrm{s})$ | Res $_{k}$ | Pres $_{1, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 0.4 | 115 | 674 | 304 | 11 | 2.41 e 2 | $3.18 \mathrm{e}-5$ | $9.02 \mathrm{e}-6$ |
| 0.49 | 112 | 658 | 286 | 12 | 2.24 e 2 | $1.97 \mathrm{e}-5$ | $5.09 \mathrm{e}-6$ |
| 0.499 | 100 | 583 | 276 | 12 | 1.98 e 2 | $3.26 \mathrm{e}-5$ | $9.75 \mathrm{e}-6$ |
| 0.4999 | 100 | 571 | 268 | 12 | 1.89 e 2 | $3.36 \mathrm{e}-5$ | $1.12 \mathrm{e}-5$ |

Table 4.11: Numerical results for Test problem 1 with varying Poisson ratio $\nu$ when $l=6, \sigma=0.17, M=7$ and $d=4$ using the symmetric preconditioning strategy.

| $\nu$ | $k$ | $n_{k}$ | rank | $i$ | time $(\mathrm{s})$ | Res $_{k}$ | Pres $_{1, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 0.4 | 305 | 1881 | 304 | 31 | 1.93 e 3 | $1.37 \mathrm{e}-4$ | $9.53 \mathrm{e}-6$ |
| 0.49 | 313 | 1897 | 294 | 54 | 4.24 e 3 | $1.12 \mathrm{e}-4$ | $8.18 \mathrm{e}-6$ |
| 0.499 | 262 | 1546 | 266 | 20 | 1.30 e 3 | $1.03 \mathrm{e}-4$ | $1.06 \mathrm{e}-5$ |
| 0.4999 | 247 | 1444 | 245 | 23 | 9.20 e 2 | $9.88 \mathrm{e}-5$ | $1.07 \mathrm{e}-5$ |

Table 4.12: Numerical results for the slow decay case of Test problem 2 with varying Poisson ratio $\nu$ when $l=6, M=7$ and $d=4$ using the symmetric preconditioning strategy.

Experiment 4 (keeping fewer basis vectors) In some of the above experiments, we observed that the dimension $n_{k}$ is much greater than the estimated rank of the final approximation. This suggests that the selected basis vectors are not optimal. Next, we try to remedy this by keeping fewer basis vectors at each iteration. In Tables 4.13-4.15, we record the numerical results for Test problem 1 when $\sigma=0.17$ and two cases of Test problem 2 with fixed $l=6, \nu=0.4999, M=7$ and $d=4$. We can see that the dimension $n_{k}$ is somewhat reduced for Test problem 1 when $\sigma=0.17$ and for
the slow decay case of Test problem 2 when we reduce $\vartheta$ from 99 to 96 but increases for the fast decay case of Test problem 2. However, the average inner iteration count becomes larger and it becomes less accurate in terms of $\operatorname{Res}_{k}$ when we reduce $\vartheta$ from 99 to 96 for the fast decay case of Test problem 2.

| $\vartheta$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{1, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 96 | 90 | 419 | 254 | 12 | 1.75 e 2 | $4.41 \mathrm{e}-5$ | $1.05 \mathrm{e}-5$ |
| 97 | 89 | 431 | 255 | 12 | 1.71 e 2 | $7.02 \mathrm{e}-5$ | $1.25 \mathrm{e}-5$ |
| 98 | 97 | 506 | 257 | 12 | 1.95 e 2 | $2.79 \mathrm{e}-5$ | $5.92 \mathrm{e}-6$ |
| 99 | 100 | 571 | 268 | 12 | 1.89 e 2 | $3.36 \mathrm{e}-5$ | $1.12 \mathrm{e}-5$ |

Table 4.13: Numerical results for Test problem 1 with varying $\vartheta$ when $l=6, \sigma=$ $0.17, \nu=0.4999 M=7$ and $d=4$ using the symmetric preconditioning strategy.

| $\vartheta$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{1, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 96 | 210 | 1027 | 236 | 44 | 7.62 e 2 | $9.20 \mathrm{e}-5$ | $5.78 \mathrm{e}-6$ |
| 97 | 220 | 1118 | 239 | 31 | 7.00 e 2 | $9.42 \mathrm{e}-5$ | $6.09 \mathrm{e}-6$ |
| 98 | 237 | 1271 | 243 | 20 | 6.71 e 2 | $1.44 \mathrm{e}-4$ | $1.65 \mathrm{e}-5$ |
| 99 | 247 | 1444 | 245 | 23 | 9.20 e 2 | $9.88 \mathrm{e}-5$ | $1.07 \mathrm{e}-5$ |

Table 4.14: Numerical results for the slow decay case of Test problem 2 with varying $\vartheta$ when $l=6, \nu=0.4999, M=7$ and $d=4$ using the symmetric preconditioning strategy.

| $\vartheta$ | $k$ | $n_{k}$ | rank | $i$ | time $(\mathrm{s})$ | Res $_{k}$ | Pres $_{1, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 96 | 502 | 1947 | 121 | 146 | 1.20 e 4 | $1.17 \mathrm{e}-3$ | $5.55 \mathrm{e}-5$ |
| 97 | 494 | 2019 | 120 | 194 | 1.57 e 4 | $1.80 \mathrm{e}-4$ | $8.23 \mathrm{e}-6$ |
| 98 | 374 | 1594 | 115 | 138 | 6.18 e 3 | $2.08 \mathrm{e}-4$ | $1.01 \mathrm{e}-5$ |
| 99 | 333 | 1552 | 115 | 59 | 2.56 e 3 | $1.80 \mathrm{e}-4$ | $8.99 \mathrm{e}-6$ |

Table 4.15: Numerical results for the fast decay case of Test problem 2 with varying $\vartheta$ when $l=6, \nu=0.4999, M=7$ and $d=4$ using the symmetric preconditioning strategy.

Experiment 5 (the multi-parameter strategy) The choice of basis vectors is also affected by the choice of parameters $s_{i, r}$ in (3.26) with $i=1, \ldots, n$ and $r=$
$1,2, \ldots, M+1$. In Tables 4.16-4.18, we fix the Poisson ratio $\nu=0.4999$, polynomial degree $d=4$ and use the multi-parameter strategy described in Section 3.2.2 to choose 5 distinct parameters $s_{1, r}, s_{2, r}, \ldots, s_{5, r}$ for every matrix $P_{r}$.

| $M$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{1, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 126 | 76 | 358 | 124 | 12 | 1.02 e 2 | $4.55 \mathrm{e}-5$ | $1.25 \mathrm{e}-5$ |
| 7 | 330 | 96 | 534 | 267 | 14 | 1.78 e 2 | $3.38 \mathrm{e}-5$ | $1.40 \mathrm{e}-5$ |
| 13 | 2380 | 183 | 1377 | 801 | 37 | 2.31 e 3 | $2.32 \mathrm{e}-5$ | $5.16 \mathrm{e}-6$ |

Table 4.16: Numerical results for Test problem 1 with varying $M$ when $l=6$, $\sigma=0.17, \nu=0.4999$ and $d=4$ using the multi-parameter strategy and the symmetric preconditioning strategy.

| $M$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{1, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 126 | 243 | 1161 | 121 | 18 | 4.74 e 2 | $1.11 \mathrm{e}-4$ | $6.72 \mathrm{e}-6$ |
| 7 | 330 | 269 | 1488 | 246 | 14 | 8.60 e 2 | $1.06 \mathrm{e}-4$ | $1.15 \mathrm{e}-5$ |
| 13 | 2380 | 231 | 1555 | 525 | 17 | 2.49 e 3 | $1.21 \mathrm{e}-4$ | $1.73 \mathrm{e}-5$ |

Table 4.17: Numerical results for the slow decay case of Test problem 2 with varying $M$ when $l=6, \nu=0.4999$ and $d=4$ using the multi-parameter strategy and the symmetric preconditioning strategy.

| $M$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time $(\mathrm{s})$ | Res $_{k}$ | Pres $_{1, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 126 | 871 | 3537 | 85 | $\mathbf{1 9 9}$ | 7.03 e 4 | $1.05 \mathrm{e}-4$ | $4.46 \mathrm{e}-6$ |
| 7 | 330 | 541 | 2225 | 120 | $\mathbf{1 2 5}$ | 1.63 e 4 | $1.86 \mathrm{e}-4$ | $8.78 \mathrm{e}-6$ |
| 13 | 2380 | 406 | 1694 | 165 | 44 | 1.03 e 4 | $3.51 \mathrm{e}-4$ | $1.78 \mathrm{e}-5$ |

Table 4.18: Numerical results for the fast decay case of Test problem 2 with varying $M$ when $l=6, \nu=0.4999$ and $d=4$ using the multi-parameter strategy and the symmetric preconditioning strategy. Bold numbers indicate unexpectedly large iteration counts.

For Test problem 1 with $\sigma=0.17$ and the slow decay case of Test problem 2, the iteration counts $k$ and dimension $n_{k}$ are reduced compared with the numerical results
in Tables 4.6-4.7 when we use the parameter-free strategy. However, the iteration counts $k$ and dimension $n_{k}$ are increased compared with the numerical results in Table 4.8 for the fast decay case of Test problem 2.

Experiment 6 (inner MINRES iterations) In some of the above experiments, it was observed that at some outer iterations, the inner MINRES solver required a large number of inner iterations to meet the selected stopping condition. This resulted in the overall timings being extremely slow e.g., when $M=5$ and $d=5$ in Table 4.8 as well as when $M=5$ and $M=7$ in Table 4.18. In Figure 4.3, we plot the number of inner iterations required at each iteration for the latter experiment. Recall that when we apply a Galerkin condition to $R_{k}$ with the symmetric preconditioning strategy, the well-posedness of the reduced problem (3.30) is not guaranteed. We checked the flag of MINRES in MATLAB at each iteration and found no issues. However, by examining the eigenvalues of the coefficient matrix associated with the reduced problem we did find problems with ill-conditioning at some iterations. For example, the smallest absolute value of the eigenvalues is $2.6 \times 10^{-3}$ and the largest is $8.4 \times 10^{8}$ at the 25 -th iteration when we solve Test problem 1 with $\sigma=0.17, \nu=0.4999$, $M=6$ and $d=4$ using the multi-parameter strategy.


Figure 4.3: Inner iteration counts for the fast decay case of Test problem 2 when $l=6, \nu=0.4999$ and $d=4$ using multi-parameter strategy.

Experiment 7 (choice of projection strategy) We show the convergence history of the relative difference $\Delta_{k}$, the preconditioned relative residual Pres $_{1, k}$ and unpreconditioned relative residual $\operatorname{Res}_{k}$ for some small problems with spatial grid level

5 (giving $\left.n_{x}=2562\right), M=7$ and total polynomial degree $d=4$ (giving $\left.n_{y}=330\right)$ using two types of projection strategy in Figures 4.4-4.7. Here, we set $\epsilon_{\text {outer }}=10^{-8}$.

(a) $\sigma=0.085$

(b) $\sigma=0.17$

Figure 4.4: Convergence history for Test problem 1 when we apply the Galerkin condition with $\nu=0.4999, l=5, M=7$ and $d=4$.


Figure 4.5: Convergence history for Test problem 2 when we apply the Galerkin condition with $\nu=0.4999, l=5, M=7$ and $d=4$.

First, we apply the Galerkin condition to $R_{k}$ and show the convergence history in Figures 4.4-4.5. Since the coefficient matrix of the associated preconditioned linear system is indefinite, applying the Galerkin condition to $R_{k}$ does not minimise the residual. From the convergence history, we can see that the preconditioned relative residual and unpreconditioned relative residual are not monotonically decreasing. For both test problems, we can see that the successive relative difference $\Delta_{k}$ is the smallest
among these three distinct errors. For Test problem 1, the unpreconditioned relative residual $\operatorname{Res}_{k}$ is nearly two orders of magnitude higher than the successive relative difference $\Delta_{k}$. For Test problem 2, $\operatorname{Res}_{k}$ is roughly two orders of magnitude higher than $\Delta_{k}$. For both test problems, the preconditioned relative residual Pres $_{1, k}$ has a good agreement with the successive relative difference $\Delta_{k}$. The relative residuals $R e s_{k}$ and Pres $_{1, k}$ for Test problem 2 are more erratic than Test problem 1.


Figure 4.6: Convergence history for Test problem 1 when we apply the PetrovGalerkin condition with $\nu=0.4999, l=5, M=7$ and $d=4$.

(a) slow decay

(b) fast decay

Figure 4.7: Convergence history for Test problem 2 when we apply the PetrovGalerkin condition with $\nu=0.4999, l=5, M=7$ and $d=4$.

Next, we apply the Petrov-Galerkin condition to $R_{k}$ at every outer iteration. In Figures 4.6-4.7, we show the convergence history of $\operatorname{Pres}_{1, k}, \operatorname{Res}_{k}$ and $\Delta_{k}$ at every
iteration for the same problem showed in Figures 4.4-4.5. Again, we use the parameterfree strategy and use the backslash operator in MATLAB to construct the basis vectors and keep the $99 \%$ most significant directions. That is the singular vectors associated with the $99 \%$ largest singular values. We can see that the preconditioned relative residual Pres $_{1, k}$ is now monotonically decreasing and so is the unpreconditioned relative residual $\operatorname{Res}_{k}$. The relative difference $\Delta_{k}$ also becomes smoother than when applying the Galerkin condition. For both test problems, the unpreconditioned relative residual $R e s_{k}$ is about two orders of magnitude higher than the successive relative difference $\Delta_{k}$.

| $M$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{1, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 126 | 77 | 370 | 123 | 10 | 1.32 e 2 | $8.57 \mathrm{e}-5$ | $6.97 \mathrm{e}-6$ |
| 7 | 330 | 93 | 528 | 264 | 10 | 2.64 e 2 | $1.15 \mathrm{e}-4$ | $9.25 \mathrm{e}-6$ |
| 13 | 2380 | 185 | 1437 | 839 | 10 | 1.02 e 4 | $4.84 \mathrm{e}-5$ | $3.38 \mathrm{e}-6$ |

Table 4.19: Numerical results for Test problem 1 with varying $M$ when $l=6$, $\nu=0.4999, \sigma=0.17$ and $d=4$ when we apply the Petrov-Galerkin condition.

| $M$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{1, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 126 | 248 | 1234 | 126 | 12 | 7.39 e 2 | $1.57 \mathrm{e}-4$ | $7.15 \mathrm{e}-6$ |
| 7 | 330 | 239 | 1394 | 245 | 12 | 1.47 e 3 | $1.54 \mathrm{e}-4$ | $7.97 \mathrm{e}-6$ |
| 13 | 2380 | 212 | 1524 | 519 | 12 | 1.57 e 4 | $1.94 \mathrm{e}-4$ | $1.19 \mathrm{e}-5$ |

Table 4.20: Numerical results for the slow decay case of Test problem 2 with varying $M$ when $l=6, \nu=0.4999$ and $d=4$ when we apply the Petrov-Galerkin condition.

| $M$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{1, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 126 | 357 | 1542 | 79 | 21 | 1.42 e 3 | $2.78 \mathrm{e}-4$ | $1.36 \mathrm{e}-5$ |
| 7 | 330 | 341 | 1589 | 115 | 21 | 3.13 e 3 | $1.47 \mathrm{e}-4$ | $6.92 \mathrm{e}-6$ |
| 13 | 2380 | 272 | 1304 | 157 | 20 | 2.68 e 4 | $3.22 \mathrm{e}-4$ | $1.50 \mathrm{e}-5$ |

Table 4.21: Numerical results for the fast decay case of Test problem 2 with varying $M$ when $l=6, \nu=0.4999$ and $d=4$ when we apply the Petrov-Galerkin condition.

In Tables 4.19-4.21, we record the numerical results obtained when we apply the Petrov-Galerkin condition to test problems with grid level $l=6, \nu=0.4999$ and $d=4$. Compared with the results in Tables 4.6-4.8, the number of outer iterations $k$ and the dimension $n_{k}$ are barely changed but the timings for problems with large $M$ have increased significantly. We can see the average inner MINRES iteration counts $i$ are stable now when we vary $M$ and lower than the results obtained when we apply the Galerkin condition. Since the reduced problems are well-conditioned when we apply the Petrov-Galerkin condition, a large number of iterations is not needed for the inner MINRES solver.

| Problem | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{1, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| $P 1$ | 252 | 369 | 1603 | 107 | 22 | 2.06 e 3 | $3.11 \mathrm{e}-4$ | $1.60 \mathrm{e}-5$ |
| $P 2$ | 126 | 361 | 1337 | 78 | 20 | 1.27 e 3 | $7.08 \mathrm{e}-4$ | $3.76 \mathrm{e}-5$ |
| $P 3$ | 330 | 382 | 1492 | 114 | 20 | 3.63 e 3 | $5.12 \mathrm{e}-4$ | $2.52 \mathrm{e}-5$ |

Table 4.22: Numerical results for problems $P 1, P 2$ and $P 3$ when we apply the PetrovGalerkin condition.


Figure 4.8: MINRES inner iteration counts for $P 2$ and $P 3$ when we apply the PetrovGalerkin condition.

Recall that the inner iteration counts are problematic for problems when $M=5$ and $d=5$ in Table 4.8 as well as $M=5$ and $M=7$ in Table 4.18. Let us denote these problems by $P 1, P 2$ and $P 3$, respectively. We now apply the Petrov-Galerkin
condition to these problems and record the numerical results in Table 4.22. Compared with the results obtained when we applied the Galerkin condition, we can see that the number of outer iterations $k$, the dimension $n_{k}$, the average number of inner iteration $i$ and the timings are significantly reduced. Compared with the former numerical results, the accuracy in terms of the unpreconditioned relative residual $R e s_{k}$ stays within the same order of magnitude.

For $P 2$ and $P 3$, we plot the number of inner iterations required at every outer iteration when we apply the Petrov-Galerkin condition in Figure 4.8. The maximum inner iteration count required is less than 35 , which is much less than when we applied the Galerkin condition.

Experiment 8 (large problems) In all the above experiments, $n_{x} \cdot n_{y}$ was relatively small in order to allow a comparison with MINRES on the original problem, and to permit the computation of residuals. However, recall that our main motivation for developing the reduced basis method is to solve large scale problems. We were unable to solve some large problems using preconditioned MINRES due to memory limitations (see Table 2.18 in Chapter 2). We applied Multi-RB with our symmetric preconditioning strategy to these problems but the machine we use still runs out of memory. This is because the chosen block-diagonal preconditioner $K_{S}$ has a dense block of size $n_{p} \times n_{p}$, which exhausts available memory when fine finite element meshes are used.


Figure 4.9: The relative difference $\Delta_{k}$ against the dimension $n_{k}$ for Test problem 1 when $\sigma=0.085, \nu=0.4999, l=6, M=5$ and $d=3$.

We also experimented with a different symmetric preconditioner, namely the sparser
preconditioner $K_{S, \text { approx }}$ in (2.58). However, in experiments using the Galerkin condition, we observed that the relative difference $\Delta_{k}$ decreased more slowly and almost stagnated. For example, in Figure 4.9, we plot the relative difference $\Delta_{k}$ obtained for Test problem 1 when $\sigma=0.085, \nu=0.4999, l=6, M=5$ and $d=3$.

### 4.3 Conclusions

When we apply Multi-RB with our chosen symmetric preconditioning strategy to large problems that cannot be solved using preconditioned MINRES, we still encounter problems with memory. This is because the dense block of the chosen preconditioner $K_{S}$ requires a high storage for large problems on a fine mesh.

For small problems, we found that the number of iterations $k$ and the dimension $n_{k}$ required to meet the chosen stopping tolerance are independent of the total polynomial degree $d$, finite element grid level $l$ and the Poisson ratio $\nu$, but depend on the number $M$ of parameters and standard deviation. The unpreconditioned relative residual $\operatorname{Res}_{k}$ is one to two orders of magnitude higher than the preconditioned relative residual Pres $_{1, k}$ and the preconditioned relative residual Pres $_{1, k}$ has a good agreement with the successive relative difference $\Delta_{k}$ in most cases.

In all numerical experiments under the default settings described at the beginning of Section 4.2, the dimension $n_{k}$ is much greater than the estimated rank of $X_{k}$ for both test problems, which suggests that selected basis vectors are not optimal. However, keeping fewer basis vectors at each iteration or using the multi-parameter strategy to choose $s_{i, r}$ did not remedy this.

When we use the Galerkin condition to build the reduced problem, the convergence of the successive relative difference $\Delta_{k}$, the preconditioned relative residual $\operatorname{Pres}_{1, k}$ and the unpreconditioned relative residual $\operatorname{Res}_{k}$ are not stable. That is, these errors increase drastically at some iterations. In addition, the reduced problems are not wellconditioned at some iterations and so a large number of inner iterations are required. To fix this, we applied the Petrov-Galerkin condition to build the reduced problem. The convergence history becomes smoother than imposing the Galerkin condition, but the computational cost is expensive because the number of terms in the reduced problem (3.32) is much larger, and the convergence rate is not improved obviously
for most cases. However, for some problematic cases when we apply the Galerkin condition, applying the Petrov-Galerkin condition can make the inner iteration counts stable and so the convergence can be obviously improved.

## Chapter 5

## Multi-RB with Left

## Preconditioning

In this chapter, we apply Multi-RB to the matrix equation (3.21) associated with the left preconditioning strategy discussed in Section 3.2.1 with the saddle point preconditioner $K_{0}$. The shifted left preconditioned matrix equation is defined in (3.20) and in that case the left matrices $P_{r}$ in (3.21) are defined as

$$
P_{r}=\widetilde{K}_{r}+\alpha_{r} I, \quad r=1,2, \ldots, M
$$

where the shifts $\alpha_{r}$ for $r=1, \ldots, M$ need to be chosen based on the eigenvalues of

$$
\widetilde{K}_{r}:=K_{0}^{-1} K_{r}, \quad r=1, \ldots, M,
$$

to ensure that $P_{r}$ for $r=1, \ldots, M$ are non-singular. The theoretical eigenvalue bounds for $\widetilde{K}_{r}$ are given in Lemma 3.3 and they are the same as the theoretical eigenvalue bounds for $\widehat{K}_{r}$ given in Lemma 3.1 or Lemma 3.2 when $e_{r}(\boldsymbol{x})$ is strictly positive. However, the first symmetrically preconditioned left matrix $\hat{K}_{0}$ in the matrix equation (3.6) has three distinct nonzero eigenvalues whereas the first left preconditioned left matrix in the matrix equation (3.15) is the $n_{x} \times n_{x}$ identity matrix with eigenvalues all equal to one.

### 5.1 Eigenvalues and choice of shifts

The shifts $\alpha_{r}$ are chosen based on the eigenvalues of $\widetilde{K}_{r}$, for $r=1, \ldots, M$. We know that the theoretical bounds for the eigenvalues for $\widetilde{K}_{r}$ are the same as those
for $\hat{K}_{r}$ for $r=1, \ldots, M$. Therefore, if we use the same numerically computed values of $\left\|e_{r}(\boldsymbol{x})\right\|_{\infty}$ as we used to compute the theoretical eigenvalue bounds for $\widehat{K}_{r}$, the theoretical eigenvalue bounds are the same as in Tables 4.1-4.2.

| $r$ | $\sigma=0.085$ | $\sigma=0.17$ |
| :---: | :---: | :---: |
| 1 | $\{0\} \cup[0.0812,0.1251]$ | $\{0\} \cup[0.1624,0.2501]$ |
| 2 | $[-0.0667,0.0667]$ | $[-0.1334,0.1334]$ |
| 3 | $[-0.0667,0.0667]$ | $[-0.1334,0.1334]$ |
| 4 | $[-0.0397,0.0397]$ | $[-0.0794,0.0795]$ |
| 5 | $[-0.0397,0.0397]$ | $[-0.0794,0.0795]$ |
| 6 | $[-0.0355,0.0355]$ | $[-0.0711,0.0711]$ |
| 7 | $[-0.0276,0.0275]$ | $[-0.0553,0.0550]$ |

Table 5.1: Numerical eigenvalue bounds for $\widetilde{K}_{r}$, for $r=1, \ldots, 7$, in Test problem 1.

| $r$ | slow decay | fast decay |
| :---: | :---: | :---: |
| 1 | $[-0.5456,0.5430]$ | $[-0.8299,0.8260]$ |
| 2 | $[-0.1364,0.1364]$ | $[-0.0519,0.0519]$ |
| 3 | $[-0.0602,0.0602]$ | $[-0.0102,0.0102]$ |
| 4 | $[-0.0339,0.0339]$ | $[-0.0032,0.0032]$ |
| 5 | $[-0.0217,0.0217]$ | $[-0.0013,0.0013]$ |

Table 5.2: Numerical eigenvalue bounds for $\widetilde{K}_{r}$, for $r=1, \ldots, 5$, in Test problem 2.

Next, we compute the eigenvalues of $\widetilde{K}_{r}$ numerically for the two chosen test problems on a coarse spatial mesh with grid level 5 using eigs in MATLAB and record the eigenvalue bounds in Tables 5.1-5.2. Compared with the computed theoretical eigenvalue bounds in Tables 4.1-4.2, we can see that the theoretical eigenvalue bounds are very sharp. The numerical eigenvalues of $\widetilde{K}_{r}$ for these two test problems are also plotted in Figures 5.1-5.2.

Since the theoretical eigenvalue bounds for $\widetilde{K}_{r}$ are sharp, we can choose the shifts $\alpha_{r}$ based on the computed theoretical eigenvalue bounds. Similar to the choice of $\alpha_{r}$ for $\widehat{K}_{r}$ in Section 4.1, we choose $\alpha_{1}=\left\|e_{1}(\boldsymbol{x})\right\|_{\infty} / 2$ and $\alpha_{r}=1$, for $r=2, \ldots, M$, in

Test problem 1 and choose $\alpha_{r}=1$, for $r=1, \ldots, M$, in Test problem 2. In this case, the eigenvalues of $P_{r}=\widetilde{K}_{r}+\alpha_{r} I$ for each $r=1,2, \ldots, M$ are centred around 1 .

(a) $\sigma=0.085$

(b) $\sigma=0.17$

Figure 5.1: Numerical eigenvalues of $\widetilde{K}_{r}$ for $r=1,2, \ldots, 7$ in Test problem 1 computed on a spatial mesh with $l=5$.


Figure 5.2: Numerical eigenvalues of $\widetilde{K}_{r}$ for $r=1,2, \ldots, 5$ in Test problem 2 computed on a spatial mesh with grid $l=5$.

### 5.2 Numerical experiments

In this section, we apply Multi-RB to the left preconditioned matrix equation (3.21) for Test problems $1-2$. We choose $\epsilon_{\text {outer }}=10^{-6}$ and the stopping condition for MultiRB is chosen to be $\Delta_{k}<\epsilon_{\text {outer }}$. Unless otherwise stated, we use the parameter-free
strategy and use the backslash solver in MATLAB to solve the systems required to construct the basis vectors to form the columns of $W_{k}$ in Algorithm 1 and keep the $99 \%$ most significant directions. That is, we keep the singular vectors associated with the $99 \%$ largest singular values. We impose the Galerkin condition at each iteration and solve the reduced problem (3.30). We use QMR (see [29, 71]) to solve the Kronecker form (3.31) of the reduced problem iteratively and we call this the inner iteration. We set $Y_{0}=0$ and choose $Y_{k}=\left[Y_{k-1} ; 0\right]$ as the initial guess when we apply QMR to the reduced problem at the $k$ th iteration. Let $\delta_{k}^{(s)}$ be the 2-norm of relative residual of 2 -norm at the $s$-th iteration of the reduced problem. The stopping condition we use for the inner iteration is (4.1) and we choose $\epsilon_{\text {inner }}=10^{-3}$.

For all numerical experiments in this chapter, we record the number of outer iteration $k$, the dimension $n_{k}$ of the approximation space generated, the estimated rank of the final approximation $X_{k}$, the average inner iteration count $i$ and the timing in seconds. We also record the final unpreconditioned relative residual $R e s_{k}$

$$
\operatorname{Res}_{k}:=\frac{\left\|K_{0} R_{k}\right\|_{F}}{\left\|K_{0} R_{k}\right\|_{F}}
$$

where $R_{k}$ is the $k$-th residual of the preconditioned matrix equation (3.20) with the preconditioner $K_{0}$ and $K_{0} R_{k}$ is the $k$-th residual of the unpreconditioned matrix equation (3.1), as well as the final preconditioned relative residual Pres $_{2, k}$

$$
\operatorname{Pres}_{2, k}:=\frac{\left\|R_{k}\right\|_{F}}{\left\|R_{0}\right\|_{F}} .
$$

Note that computing unpreconditioned residuals allows a direct comparison with the results obtained in Chapter 4. Next, we carry out some numerical experiments to examine the performance of Multi-RB with the left preconditioning strategy.

Experiment 1 (varying $M$ and $d$ ) First, we fix the grid level to be 6 ( $n_{x}=$ 10242) and let $\nu=0.4999$ and apply Multi-RB to Test problems $1-2$ with varying $M$ and $d$.

In Tables 5.3-5.4, we record the numerical results for Test problem 1. The number of iterations $k$ and the dimension $n_{k}$ are nearly independent of the number $M$ of parameters and the total polynomial degree $d$ but depend on the standard deviation $\sigma$. The estimated rank of $X_{k}$ is closer to $n_{k}$ when $M$ and $d$ grow. The unpreconditioned relative residual $\operatorname{Res}_{k}$ is one to two orders of magnitude higher than the successive
relative difference $\Delta_{k}$. The preconditioned relative residual Pres $_{2, k}$ and the successive relative difference $\Delta_{k}$ have a good agreement.

| $M$ | $d$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{2, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 3 | 56 | 19 | 96 | 56 | 7 | 1.77 e 1 | $1.57 \mathrm{e}-5$ | $1.13 \mathrm{e}-6$ |
|  | 4 | 126 | 19 | 96 | 79 | 7 | 1.76 e 1 | $1.79 \mathrm{e}-5$ | $1.33 \mathrm{e}-6$ |
|  | 5 | 252 | 19 | 96 | 80 | 7 | 1.78 e 1 | $1.80 \mathrm{e}-5$ | $1.34 \mathrm{e}-6$ |
| 7 | 3 | 120 | 20 | 141 | 110 | 7 | 2.83 e 2 | $2.64 \mathrm{e}-5$ | $1.52 \mathrm{e}-6$ |
|  | 4 | 330 | 20 | 141 | 129 | 7 | 2.83 e 1 | $2.64 \mathrm{e}-5$ | $1.52 \mathrm{e}-6$ |
|  | 5 | 792 | 20 | 141 | 130 | 8 | 2.98 e 1 | $2.77 \mathrm{e}-5$ | $1.60 \mathrm{e}-6$ |
| 13 | 3 | 560 | 25 | 320 | 292 | 7 | 8.67 e 1 | $7.97 \mathrm{e}-5$ | $1.58 \mathrm{e}-6$ |
|  | 4 | 2380 | 26 | 332 | 325 | 8 | 1.15 e 2 | $7.93 \mathrm{e}-5$ | $1.46 \mathrm{e}-6$ |
|  | 5 | 8568 | 26 | 332 | 326 | 8 | 1.75 e 2 | $7.96 \mathrm{e}-5$ | $1.47 \mathrm{e}-6$ |

Table 5.3: Numerical results for Test problem 1 with fixed grid level $6, \nu=0.4999$ and $\sigma=0.085$ using Multi-RB with the left preconditioning strategy.

| $M$ | $d$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{2, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 3 | 56 | 38 | 191 | 56 | 10 | 3.57 e 1 | $5.62 \mathrm{e}-5$ | $2.05 \mathrm{e}-6$ |
|  | 4 | 126 | 40 | 201 | 120 | 10 | 3.86 e 1 | $6.48 \mathrm{e}-5$ | $2.08 \mathrm{e}-6$ |
|  | 5 | 252 | 41 | 206 | 155 | 11 | 4.00 e 1 | $6.97 \mathrm{e}-5$ | $2.18 \mathrm{e}-6$ |
| 7 | 3 | 120 | 33 | 232 | 120 | 10 | 4.74 e 1 | $7.11 \mathrm{e}-5$ | $2.01 \mathrm{e}-6$ |
|  | 4 | 330 | 35 | 246 | 213 | 11 | 5.28 e 1 | $8.72 \mathrm{e}-5$ | $2.10 \mathrm{e}-6$ |
|  | 5 | 792 | 38 | 267 | 243 | 11 | 6.28 e 1 | $8.77 \mathrm{e}-5$ | $1.75 \mathrm{e}-6$ |
| 13 | 3 | 560 | 37 | 475 | 413 | 10 | 1.51 e 2 | $4.90 \mathrm{e}-4$ | $1.85 \mathrm{e}-6$ |
|  | 4 | 2380 | 42 | 539 | 325 | 11 | 2.42 e 2 | $6.50 \mathrm{e}-4$ | $1.56 \mathrm{e}-6$ |
|  | 5 | 8568 | 43 | 552 | 552 | 12 | 5.35 e 2 | $7.49 \mathrm{e}-4$ | $1.72 \mathrm{e}-6$ |

Table 5.4: Numerical results for Test problem 1 with fixed grid level $6, \nu=0.4999$ and $\sigma=0.17$ using Multi-RB with the left preconditioning strategy.

In Tables 5.5-5.6, we record the numerical results for Test problem 2. This time, the number of iterations $k$ decreases as we increase the number $M$ of parameters and increases as we increase the total polynomial degree $d$. The dimension $n_{k}$ grows as
$d$ grows in both cases and it is much larger than the estimated rank of $X_{k}$ especially in the fast decay case. The unpreconditioned relative residual $R e s_{k}$ is observed to be two to four orders of magnitude higher than the successive relative difference $\Delta_{k}$. The preconditioned relative residual Pres $_{2, k}$ is at most one order of magnitude higher than the successive relative difference $\Delta_{k}$.

| $M$ | $d$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time $(\mathrm{s})$ | Res $_{k}$ | Pres $_{2, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 3 | 56 | 92 | 460 | 56 | 15 | 8.75 e 1 | $1.27 \mathrm{e}-3$ | $5.77 \mathrm{e}-6$ |
|  | 4 | 126 | 108 | 540 | 120 | 17 | 1.11 e 2 | $3.08 \mathrm{e}-3$ | $7.95 \mathrm{e}-6$ |
|  | 5 | 252 | 114 | 570 | 186 | 17 | 1.34 e 2 | $4.91 \mathrm{e}-3$ | $9.91 \mathrm{e}-6$ |
| 7 | 3 | 120 | 54 | 376 | 120 | 15 | 1.15 e 2 | $2.76 \mathrm{e}-4$ | $2.45 \mathrm{e}-6$ |
|  | 4 | 330 | 60 | 418 | 223 | 16 | 9.84 e 1 | $7.10 \mathrm{e}-4$ | $4.13 \mathrm{e}-6$ |
|  | 5 | 792 | 75 | 523 | 286 | 18 | 1.55 e 2 | $1.27 \mathrm{e}-3$ | $3.48 \mathrm{e}-6$ |
| 13 | 3 | 560 | 39 | 465 | 314 | 15 | 1.44 e 2 | $4.46 \mathrm{e}-4$ | $2.14 \mathrm{e}-6$ |
|  | 4 | 2380 | 47 | 565 | 408 | 17 | 3.21 e 2 | $1.02 \mathrm{e}-3$ | $2.22 \mathrm{e}-6$ |
|  | 5 | 8568 | 49 | 590 | 454 | 17 | 8.52 e 2 | $1.88 \mathrm{e}-3$ | $2.93 \mathrm{e}-6$ |

Table 5.5: Numerical results for the slow decay case in Test problem 2 with fixed grid level 6 and $\nu=0.4999$ using Multi-RB with the left preconditioning strategy.

| $M$ | $d$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{2, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 3 | 56 | 209 | 834 | 56 | 24 | 2.12 e 2 | $1.15 \mathrm{e}-2$ | $1.11 \mathrm{e}-5$ |
|  | 4 | 126 | 255 | 1025 | 79 | 27 | 3.39 e 2 | $1.86 \mathrm{e}-2$ | $1.83 \mathrm{e}-5$ |
|  | 5 | 252 | 361 | 1470 | 111 | 30 | 8.60 e 2 | $1.48 \mathrm{e}-2$ | $1.03 \mathrm{e}-5$ |
| 7 | 3 | 120 | 135 | 652 | 75 | 23 | 1.93 e 2 | $8.71 \mathrm{e}-3$ | $9.67 \mathrm{e}-6$ |
|  | 4 | 330 | 212 | 1059 | 112 | 27 | 5.27 e 2 | $1.11 \mathrm{e}-2$ | $7.45 \mathrm{e}-6$ |
|  | 5 | 792 | 265 | 1339 | 146 | 30 | 1.54 e 3 | $1.09 \mathrm{e}-2$ | $7.05 \mathrm{e}-6$ |
| 13 | 3 | 560 | 104 | 661 | 314 | 23 | 3.70 e 2 | $4.61 \mathrm{e}-3$ | $5.38 \mathrm{e}-6$ |
|  | 4 | 2380 | 136 | 903 | 155 | 26 | 1.62 e 3 | $9.22 \mathrm{e}-3$ | $7.31 \mathrm{e}-6$ |
|  | 5 | 8568 | 207 | 1458 | 197 | 30 | 1.79 e 4 | $9.19 \mathrm{e}-3$ | $5.37 \mathrm{e}-6$ |

Table 5.6: Numerical results for the fast decay case in Test problem 2 with fixed grid level 6 and $\nu=0.4999$ using Multi-RB with the left preconditioning strategy.

For both test problems, the average inner iteration count $i$ is nearly independent of $M$ and $d$ and we did not observe any large counts as observed in Figure 4.3 when we applied Multi-RB with the symmetric preconditioning strategy. However, the accuracy in terms of the unpreconditioned relative residual is worse than when we used the symmetric preconditioning strategy. We compared plots of expectations and variances of the final approximation obtained using Multi-RB with the plots obtained using preconditioned MINRES for Test problem 2 and observed that they are consistent.

Experiment 2 (varying grid level $l$ ) Next, we apply Multi-RB to Test problem 1 with $\sigma=0.17$ and the slow decay case of Test problem 2 with fixed $\nu=0.4999$, $M=7$ and $d=4$ on finite element meshes with varying grid level $l=4,5,6,7$. The numerical results are recorded in Tables 5.7-5.8. We can see that the number of outer iterations $k$ and the dimension $n_{k}$ are nearly independent of the grid level $l$.

| $l$ | $n_{x}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{2, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 4 | 642 | 31 | 218 | 191 | 10 | 3.26 e 0 | $1.21 \mathrm{e}-4$ | $1.51 \mathrm{e}-6$ |
| 5 | 2562 | 38 | 267 | 219 | 11 | 4.96 e 1 | $1.67 \mathrm{e}-4$ | $1.74 \mathrm{e}-6$ |
| 6 | 10242 | 35 | 246 | 213 | 11 | 5.28 e 1 | $8.72 \mathrm{e}-5$ | $2.10 \mathrm{e}-6$ |
| 7 | 40962 | 35 | 246 | 211 | 11 | 3.00 e 2 | $8.12 \mathrm{e}-5$ | $1.97 \mathrm{e}-6$ |

Table 5.7: Numerical results for Test problem 1 with varying grid level $l$ when $\nu=0.4999, \sigma=0.17, M=7$ and $d=4$ using the left preconditioning strategy.

| $l$ | $n_{x}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{2, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 4 | 642 | 42 | 289 | 160 | 15 | 6.98 e 0 | $1.84 \mathrm{e}-4$ | $9.69 \mathrm{e}-7$ |
| 5 | 2562 | 64 | 447 | 224 | 16 | 2.79 e 1 | $7.98 \mathrm{e}-4$ | $2.53 \mathrm{e}-6$ |
| 6 | 10242 | 60 | 418 | 223 | 16 | 9.84 e 1 | $7.10 \mathrm{e}-4$ | $4.13 \mathrm{e}-6$ |
| 7 | 40962 | 60 | 418 | 220 | 16 | 4.89 e 2 | $3.25 \mathrm{e}-4$ | $3.42 \mathrm{e}-6$ |

Table 5.8: Numerical results for the slow decay case in Test problem 2 with varying grid level $l$ when $\nu=0.4999, M=7$ and $d=4$ using the left preconditioning strategy.

Experiment 3 (varying Poisson ratio $\nu$ ) We record the numerical results obtained from applying Multi-RB with the left preconditioning strategy to Test problem 1 with $\sigma=0.17$ and the slow decay case of Test problem 2 with fixed $l=6, M=7$,
$d=4$ and varying Poisson ratio $\nu$ in Tables 5.9-5.10. We can see that the number of iterations $k$ and the dimension $n_{k}$ decrease as $\nu$ approaches 0.5 , which means the solver is robust in the incompressible limit.

| $\nu$ | $k$ | $n_{k}$ | rank | $i$ | time $(\mathrm{s})$ | Res $_{k}$ | Pres $_{2, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 0.4 | 39 | 274 | 241 | 10 | 5.53 e 1 | $7.69 \mathrm{e}-5$ | $1.71 \mathrm{e}-6$ |
| 0.49 | 36 | 253 | 225 | 10 | 5.74 e 1 | $8.84 \mathrm{e}-5$ | $2.40 \mathrm{e}-6$ |
| 0.499 | 35 | 246 | 215 | 10 | 5.20 e 1 | $8.79 \mathrm{e}-5$ | $2.17 \mathrm{e}-6$ |
| 0.4999 | 35 | 246 | 213 | 11 | 5.28 e 1 | $8.72 \mathrm{e}-5$ | $2.10 \mathrm{e}-6$ |

Table 5.9: Numerical results for Test problem 1 with varying Poisson ratio $\nu$ when $l=6, \sigma=0.17, M=7$ and $d=4$ using the left preconditioning strategy.

| $\nu$ | $k$ | $n_{k}$ | rank | $i$ | time $(\mathrm{s})$ | Res $_{k}$ | Pres $_{2, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 0.4 | 85 | 592 | 273 | 15 | 1.51 e 2 | $5.51 \mathrm{e}-4$ | $3.01 \mathrm{e}-6$ |
| 0.49 | 77 | 538 | 255 | 16 | 1.30 e 2 | $4.66 \mathrm{e}-4$ | $3.32 \mathrm{e}-6$ |
| 0.499 | 74 | 516 | 236 | 17 | 1.29 e 2 | $4.16 \mathrm{e}-4$ | $2.72 \mathrm{e}-6$ |
| 0.4999 | 60 | 418 | 223 | 16 | 9.84 e 1 | $7.10 \mathrm{e}-4$ | $4.13 \mathrm{e}-6$ |

Table 5.10: Numerical results for the slow decay case in Test problem 2 with varying Poisson ratio $\nu$ when $l=6, M=7$ and $d=4$ using the left preconditioning strategy.

| $\vartheta$ | $k$ | $n_{k}$ | rank | $i$ | time $(\mathrm{s})$ | Res $_{k}$ | Pres $_{2, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 95 | 71 | 422 | 224 | 16 | 1.17 e 2 | $5.91 \mathrm{e}-4$ | $5.10 \mathrm{e}-6$ |
| 96 | 66 | 404 | 221 | 16 | 1.03 e 2 | $4.60 \mathrm{e}-4$ | $3.59 \mathrm{e}-6$ |
| 97 | 67 | 432 | 222 | 16 | 1.03 e 2 | $4.65 \mathrm{e}-4$ | $2.95 \mathrm{e}-6$ |
| 98 | 69 | 464 | 225 | 17 | 1.12 e 2 | $7.62 \mathrm{e}-4$ | $2.75 \mathrm{e}-6$ |
| 99 | 60 | 418 | 223 | 16 | 9.86 e 1 | $7.10 \mathrm{e}-4$ | $4.13 \mathrm{e}-6$ |

Table 5.11: Numerical results for the slow decay case of Test problem 2 with varying $\vartheta$ when $l=6, \nu=0.4999, M=7$ and $d=4$ using the left preconditioning strategy.

Experiment 4 (keeping fewer basis vectors) We observed that the dimension $n_{k}$ is much larger than the estimated rank of the final approximation for Test problem

2 , which suggests the selected basis vectors are not optimal. We now try to remedy this by keeping fewer basis vectors at each iteration. In Tables 5.11-5.12, we record the numerical results for Test problem 2 with fixed $l=6, \nu=0.4999, M=7$ and $d=4$. We can see that the dimension $n_{k}$ barely changes for the slow decay case but clearly decreases for the fast decay case if we keep fewer basis vectors at each iteration. The accuracy in terms of $\operatorname{Res}_{k}$ and $\operatorname{Pres}_{2, k}$ also barely changes.

| $\vartheta$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{2, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 95 | 265 | 760 | 107 | 25 | 4.79 e 2 | $1.22 \mathrm{e}-2$ | $1.28 \mathrm{e}-5$ |
| 96 | 170 | 510 | 103 | 23 | 2.42 e 2 | $2.13 \mathrm{e}-2$ | $6.03 \mathrm{e}-5$ |
| 97 | 256 | 892 | 109 | 25 | 5.43 e 2 | $2.00 \mathrm{e}-2$ | $1.79 \mathrm{e}-5$ |
| 98 | 238 | 941 | 110 | 25 | 5.61 e 2 | $1.83 \mathrm{e}-2$ | $1.70 \mathrm{e}-5$ |
| 99 | 212 | 1059 | 112 | 27 | 5.27 e 2 | $1.11 \mathrm{e}-2$ | $7.45 \mathrm{e}-6$ |

Table 5.12: Numerical results for the fast decay case of Test problem 2 with varying $\vartheta$ when $l=6, \nu=0.4999, M=7$ and $d=4$ using the left preconditioning strategy.

Experiment 5 (the multi-parameter strategy) The choice of basis vectors is also affected by the choice of parameters $s_{i, r}$ with $i=1, \ldots, n$ and $r=1,2, \ldots, M$. We record the numerical results obtained using Multi-RB with the multi-parameter strategy when $n=5$ for Test problem $1-2$ with varying $M$ and fixed $l=6, d=4$ and

| $M$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{2, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 126 | 19 | 96 | 78 | 7 | 1.86 e 1 | $1.81 \mathrm{e}-5$ | $1.36 \mathrm{e}-6$ |
| 7 | 330 | 20 | 141 | 129 | 7 | 3.04 e 1 | $2.77 \mathrm{e}-5$ | $1.62 \mathrm{e}-6$ |
| 13 | 2380 | 26 | 332 | 325 | 8 | 1.18 e 2 | $8.03 \mathrm{e}-5$ | $1.46 \mathrm{e}-6$ |

Table 5.13: Numerical results for Test problem 1 with varying $M$ when $l=6$, $\sigma=0.085, \nu=0.4999$ and $d=4$ using the multi-parameter strategy and the left preconditioning strategy.
$\nu=0.4999$. For Test problem 1, the number of iterations $k$ and the dimension $n_{k}$ barely change compared with the results in Tables 5.3-5.4 when using the parameterfree strategy. For some cases in the slow decay case of Test problem 2, the number of iterations $k$ and the dimension $n_{k}$ slightly decreased compared with the results in Table
5.5. For the fast decay case of Test problem 2, the number of iterations $k$ increases but the dimension $n_{k}$ decreases when $M=7,13$ compared with the results in Table 5.6 and it is still much larger than the estimated rank of $X_{k}$.

| $M$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{2, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 126 | 40 | 201 | 120 | 10 | 3.93 e 1 | $6.64 \mathrm{e}-5$ | $2.03 \mathrm{e}-6$ |
| 7 | 330 | 35 | 246 | 213 | 10 | 5.60 e 1 | $8.37 \mathrm{e}-5$ | $2.10 \mathrm{e}-6$ |
| 13 | 2380 | 42 | 538 | 538 | 11 | 2.39 e 2 | $6.98 \mathrm{e}-4$ | $1.60 \mathrm{e}-6$ |

Table 5.14: Numerical results for Test problem 1 with varying $M$ when $l=6$, $\sigma=0.17, \nu=0.4999$ and $d=4$ using the multi-parameter strategy and the left preconditioning strategy.

| $M$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time $(\mathrm{s})$ | Res $_{k}$ | Pres $_{2, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 126 | 108 | 540 | 120 | 17 | 1.16 e 2 | $2.57 \mathrm{e}-3$ | $6.29 \mathrm{e}-6$ |
| 7 | 330 | 60 | 416 | 223 | 16 | 1.03 e 2 | $6.73 \mathrm{e}-4$ | $3.98 \mathrm{e}-6$ |
| 13 | 2380 | 45 | 540 | 403 | 16 | 3.05 e 2 | $7.92 \mathrm{e}-4$ | $2.49 \mathrm{e}-6$ |

Table 5.15: Numerical results for the slow decay case of Test problem 2 with varying $M$ when $l=6, \nu=0.4999$ and $d=4$ using the multi-parameter strategy and the left preconditioning strategy.

| $M$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{2, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 126 | 321 | 1098 | 79 | 27 | 4.52 e 2 | $1.86 \mathrm{e}-2$ | $1.18 \mathrm{e}-5$ |
| 7 | 330 | 232 | 900 | 109 | 26 | 5.35 e 2 | $1.32 \mathrm{e}-2$ | $1.06 \mathrm{e}-5$ |
| 13 | 2380 | 182 | 869 | 154 | 25 | 2.12 e 3 | $1.14 \mathrm{e}-2$ | $8.24 \mathrm{e}-6$ |

Table 5.16: Numerical results for the fast decay case of Test problem 2 with varying $M$ when $l=6, \nu=0.4999$ and $d=4$ using the multi-parameter strategy and the left preconditioning strategy.

Experiment 6 (convergence history) We show the convergence history of the preconditioned relative residual $\operatorname{Pres}_{2, k}$, the unpreconditioned relative residual $\operatorname{Res}_{k}$ and the successive relative difference $\Delta_{k}$ against the dimension $n_{k}$ in Figures 5.3-5.4.

Here, the stopping tolerance is chosen to be $\epsilon_{\text {outer }}=10^{-8}$. For Test problem 1, the unpreconditioned relative residual $\operatorname{Res}_{k}$ is roughly three orders of magnitude higher than the successive relative difference $\Delta_{k}$ and the preconditioned relative residual $\operatorname{Pres}_{2, k}$ has a good agreement with the relative difference $\Delta_{k}$. However, we see the unpreconditioned relative residual is effectively stagnating in Figure 5.3a. For Test problem 2, the unpreconditioned relative residual $R e s_{k}$ is three to four orders of magnitude higher than the successive relative difference $\Delta_{k}$ and $\operatorname{Pres}_{2, k}$ also has a good agreement with


Figure 5.3: Convergence history for Test problem 1 with $\nu=0.4999, l=5, M=7$ and $d=4$ using the left preconditioning strategy.

(a) slow decay

(b) fast decay

Figure 5.4: Convergence history for Test problem 2 with $\nu=0.4999, l=5, M=7$ and $d=4$ using the left preconditioning strategy.
$\Delta_{k}$. Compared with the convergence history for the same problems in Figures $4.4-$ 4.5 using the symmetric preconditioning strategy and the Galerkin condition, the convergence history is much more stable. However, the difference between Res ${ }_{k}$ and $\Delta_{k}$ becomes much larger. We observed that solution plots for these problems are consistent with plots obtained using preconditioned MINRES. As we mentioned in the beginning of Section 3.2.3, the iterative error is usually controlled in a norm that is compatible with the norm of the discretisation error. Since the preconditioned relative residual has a good agreement with the relative difference, our stopping condition will be a good choice if one can prove the norm associated with the preconditioner $K_{0}$ is compatible with the norm of the discretisation error.

Experiment 7 (large problems) Recall that we were unable to solve large problems using preconditioned MINRES or Multi-RB with the symmetric preconditioning strategy due to memory constraints. We applied Multi-RB with the left preconditioning strategy to problems with $l=9$ (giving $n_{x}=655362$ ), $M=15$ and $d=4$ (giving $n_{y}=3876$ ). We set $\epsilon_{\text {outer }}=10^{-8}$ and $\nu=0.4999$. The total problem size is $2,540,183,112 \times 2,540,183,112$. The numerical results are recorded in Table 5.17. We do not run out of memory when using Multi-RB with the left preconditioning strategy. However, we cannot compute relative residuals and analyse the accuracy.

| Test problem | case | $k$ | $n_{k}$ | rank | $i$ | time (s) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 1 | $\sigma=0.085$ | 65 | 926 | 876 | 9 | 4.34 e 4 |
|  | $\sigma=0.17$ | 97 | 1392 | 1364 | 13 | 6.75 e 4 |
| 2 | slow decay | 96 | 1339 | 1027 | 19 | 6.90 e 4 |
|  | fast decay | 320 | 2879 | 368 | 31 | 2.44 e 5 |

Table 5.17: Numerical results obtained using Multi-RB with the left preconditioning strategy for test problems with $l=9, M=15, d=4$ and $\nu=0.4999$.

### 5.3 Conclusions

Using Multi-RB with the left preconditioning strategy we are able to solve larger problems that cannot be solved using either preconditioned MINRES or Multi-RB with the chosen symmetric preconditioning strategy.

When we apply Multi-RB with the left preconditioning strategy to Test problem 1 , the number of outer iterations $k$ and the dimension $n_{k}$ are independent of the finite element grid level $l$, the total polynomial degree $d$ and the Poisson ratio $\nu$ but depend on the number $M$ of parameters and the standard deviation $\sigma$ for Test problem 1. The estimated rank of the final approximation $X_{k}$ is also close to the dimension $n_{k}$ as $M$ and $d$ increase. When we apply Multi-RB with the left preconditioning strategy to Test problem 2, the number of iterations $k$ and the dimension $n_{k}$ are still independent of $l$ and $\nu$. The number of iterations $k$ decreases as we increase $M$ and increases as we increase $d$. The dimension $n_{k}$ of the constructed approximation space is larger than the estimated rank of $X_{k}$. However, keeping fewer basis vectors at each iteration or using the multi-parameter strategy did not remedy this.

Based on the convergence history of smaller problems, we observe that the unpreconditioned relative residual $R e s_{k}$ is approximately three to four orders of magnitude higher than the successive relative difference and the preconditioned relative residual $\operatorname{Pres}_{2, k}$ has a good agreement with the successive relative difference $\Delta_{k}$. In all experiments, we imposed the Galerkin condition on the residual to obtain the reduced problems. When we compare the convergence history in Figures 5.3-5.4 with that in Figures $4.4-4.5$ when we apply Multi-RB using the symmetric preconditioning strategy and imposing the Galerkin condition, we can see the convergence history is much smoother but the difference between $\operatorname{Res}_{k}$ and $\Delta_{k}$ becomes larger. Unlike the symmetric preconditioning strategy, we did not observe any large inner iteration counts when solving reduced problems.

## Chapter 6

## Parameter-dependent groundwater flow problem

In this chapter, we apply the reduced basis method introduced in Chapter 3 to the multi-term matrix equation associated with another parameter-dependent PDE problem. Our aim is to investigate if Multi-RB can be generalised to solve other indefinite problems with the same structure as the parameter-dependent linear elasticity problem. In Section 1.2, we briefly introduced the parameter-dependent groundwater flow problem with uncertain permeability coefficient

$$
\begin{align*}
A^{-1}(\boldsymbol{x}, \boldsymbol{y}) \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{y})+\nabla p(\boldsymbol{x}, \boldsymbol{y}) & =0, & & \boldsymbol{x} \in D, \boldsymbol{y} \in \Gamma,  \tag{6.1}\\
\nabla \cdot \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{y}) & =f(\boldsymbol{x}), & & \boldsymbol{x} \in D, \boldsymbol{y} \in \Gamma, \tag{6.2}
\end{align*}
$$

with the mixed boundary conditions

$$
\begin{array}{rlrl}
p(\boldsymbol{x}, \boldsymbol{y}) & =g(\boldsymbol{x}), & \boldsymbol{x} \in \partial D_{D}, \boldsymbol{y} \in \Gamma, \\
\boldsymbol{u}(\boldsymbol{x}, \boldsymbol{y}) \cdot \boldsymbol{n} & =0, & & \boldsymbol{x} \in \partial D_{N}, \boldsymbol{y} \in \Gamma . \tag{6.4}
\end{array}
$$

In this problem, the permeability coefficient $A^{-1}(\boldsymbol{x}, \boldsymbol{y})$ is modelled as a parameterdependent function of the form

$$
\begin{equation*}
A^{-1}(\boldsymbol{x}, \boldsymbol{y})=a_{0}(\boldsymbol{x})+\sum_{r=1}^{M} a_{r}(\boldsymbol{x}) y_{r}, \quad \boldsymbol{x} \in D, \quad \boldsymbol{y} \in \Gamma . \tag{6.5}
\end{equation*}
$$

We want to solve for the fluid velocity $\boldsymbol{u}: D \times \Gamma \rightarrow \mathbb{R}^{2}$ and the pressure $p: D \times \Gamma \rightarrow \mathbb{R}$.
We now define appropriate function spaces for the velocity $\boldsymbol{u}$ and the pressure $p$.

For the velocity $\boldsymbol{u}$, we choose the vector-valued solution space

$$
\mathcal{V}:=L_{\rho}^{2}\left(\Gamma, \mathcal{H}_{0, N}(D)\right),
$$

where

$$
\mathcal{H}_{0, N}:=\left\{\boldsymbol{v} \in\left(L^{2}(D)\right)^{2} ; \nabla \cdot \boldsymbol{v} \in L^{2}(D), \boldsymbol{v} \cdot \boldsymbol{n}=0 \text { on } \partial D_{N}\right\} .
$$

For the pressure $p$, we choose the scalar-valued solution space

$$
\mathcal{Q}:=L_{\rho}^{2}\left(\Gamma, L^{2}(D)\right)
$$

The weak formulation can be written as: find $\boldsymbol{u} \in \mathcal{V}$ and $p \in \mathcal{Q}$ such that

$$
\begin{array}{rlrl}
a(\boldsymbol{u}, \boldsymbol{v})+b(\boldsymbol{v}, p) & =\ell(\boldsymbol{v}), & \forall \boldsymbol{v} & \in \mathcal{V}, \\
b(\boldsymbol{u}, q) & =m(q), & q \in \mathcal{Q}, \tag{6.7}
\end{array}
$$

where $a(\cdot, \cdot): \mathcal{V} \times \mathcal{V} \rightarrow \mathbb{R}$ and $b(\cdot, \cdot): \mathcal{V} \times \mathcal{Q} \rightarrow \mathbb{R}$ are defined as

$$
\begin{aligned}
a(\boldsymbol{u}, \boldsymbol{v}) & :=\int_{\Gamma} \rho(\boldsymbol{y}) \int_{D} A^{-1}(\boldsymbol{x}, \boldsymbol{y}) \boldsymbol{u}(\boldsymbol{x}, \boldsymbol{y}) \cdot \boldsymbol{v}(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d} \boldsymbol{x} \mathrm{~d} \boldsymbol{y} \\
b(\boldsymbol{v}, q) & :=-\int_{\Gamma} \rho(\boldsymbol{y}) \int_{D} q(\boldsymbol{x}, \boldsymbol{y})(\nabla \cdot \boldsymbol{v}(\boldsymbol{x}, \boldsymbol{y})) \mathrm{d} \boldsymbol{x} \mathrm{~d} \boldsymbol{y}
\end{aligned}
$$

and the functions $\ell: V \rightarrow \mathbb{R}$ and $m: \mathcal{Q} \rightarrow \mathbb{R}$ are defined as

$$
\begin{aligned}
\ell(\boldsymbol{v}) & :=-\int_{\Gamma} \rho(\boldsymbol{y}) \int_{\partial D_{D}} g(\boldsymbol{x}) \boldsymbol{v}(\boldsymbol{x}, \boldsymbol{y}) \cdot \boldsymbol{n} \mathrm{d} s \mathrm{~d} \boldsymbol{y} \\
m(q) & :=-\int_{\Gamma} \rho(\boldsymbol{y}) \int_{D} f(\boldsymbol{x}) q(\boldsymbol{x}, \boldsymbol{y}) \mathrm{d} \boldsymbol{x} \mathrm{~d} \boldsymbol{y}
\end{aligned}
$$

To ensure the well-posedness of the weak formulation, analogous to Assumption 2.18 for $E(\boldsymbol{x}, \boldsymbol{y})$, we also make the following assumption for $A^{-1}(\boldsymbol{x}, \boldsymbol{y})$ :

Assumption 6.1. We assume that there exist two constants $A_{\min }, A_{\max } \in \mathbb{R}^{+}$such that

$$
0<A_{\min } \leqslant A(\boldsymbol{x}, \boldsymbol{y}) \leqslant A_{\max }<\infty, \quad \text { a.e. in } D \times \Gamma .
$$

We choose each parameter $y_{r}$ to be the image of a uniform random variable $\xi_{r} \sim$ $U(-1,1)$ with the constant density function $\rho\left(y_{r}\right)=1 / 2$ and assume that the random variables $\xi_{r}$, for $r=1,2 \ldots, M$ are independent. We then have $y_{r} \in \Gamma_{r}=[-1,1]$ and $\boldsymbol{y} \in \Gamma=[-1,1]^{M}$.

Next, we apply a stochastic Galerkin mixed finite element scheme to the weak formulation (6.6)-(6.7). We choose $D$ to be a square domain and partition it into a uniform mesh of square elements of edge length $h$. Let $l$ be the grid level, then the number of elements is $2^{l} \times 2^{l}$. For the parametric approximation, we choose $S_{d}$ to be the set of multivariate polynomials on $\Gamma$ of total degree less than or equal to $d$ as we did for the parameter-dependent linear elasticity problem. For the spatial approximation, we choose the approximation space $\boldsymbol{V}_{h}$ for the velocity $\boldsymbol{u}$ to be the set of lowest order Raviart-Thomas finite element functions (see [65])

$$
\boldsymbol{V}_{h}:=\operatorname{span}\left\{\boldsymbol{\phi}_{1}(\boldsymbol{x}), \boldsymbol{\phi}_{2}(\boldsymbol{x}), \ldots, \boldsymbol{\phi}_{n_{u}}(\boldsymbol{x})\right\}
$$

and the approximation space for the pressure $p$ to be the set of piecewise constant functions on some spatial mesh on $D$

$$
Q_{h}:=\operatorname{span}\left\{\varphi_{1}(\boldsymbol{x}), \varphi_{2}(\boldsymbol{x}), \ldots, \varphi_{n_{p}}(\boldsymbol{x})\right\} .
$$

See [27] for more details.
After applying the stochastic Galerkin mixed finite element method, we obtain a linear system with the Kronecker product structure (1.38). The coefficient matrix of the linear system is symmetric and indefinite. The matrices $G_{r} \in \mathbb{R}^{n_{y} \times n_{y}}, r=$ $0,1, \ldots, M$, are sparse and symmetric and are defined the same as in (2.39)-(2.40). The finite element matrices $K_{r} \in \mathbb{R}^{n_{x} \times n_{x}}$, where $n_{x}=n_{u}+n_{p}$, have a simpler $2 \times 2$ block structure and are defined as

$$
K_{0}:=\left[\begin{array}{cc}
A_{0} & B^{\top} \\
B & 0
\end{array}\right], \quad K_{r}:=\left[\begin{array}{cc}
A_{r} & 0 \\
0 & 0
\end{array}\right]
$$

where $A_{r} \in \mathbb{R}^{n_{u} \times n_{u}}, r=0,1, \ldots, M$ are weighted mass matrices associated with the coefficients $a_{r}(\boldsymbol{x})$ in (6.5) and $B \in \mathbb{R}^{n_{p} \times n_{u}}$ is of full row rank. The entries of $A_{r}$, for $r=0,1, \ldots, M$, are defined as

$$
A_{r}(i, j):=\int_{D} a_{r}(\boldsymbol{x}) \boldsymbol{\phi}_{i}(\boldsymbol{x}) \cdot \boldsymbol{\phi}_{j}(\boldsymbol{x}) \mathrm{d} \boldsymbol{x}, \quad i, j=1, \ldots, n_{u} .
$$

The matrix $A_{0}$ is defined in terms of $a_{0}(\boldsymbol{x})$ which needs to be a positive function and so $A_{0}$ needs to be symmetric and positive definite. The entries of $B \in \mathbb{R}^{n_{p} \times n_{u}}$ are defined as

$$
B(i, j):=-\int_{D} \varphi_{i}(\boldsymbol{x}) \nabla \cdot \boldsymbol{\phi}_{j}(\boldsymbol{x}) \mathrm{d} \boldsymbol{x}, \quad i=1, \ldots, n_{p}, j=1, \ldots, n_{u} .
$$

The right hand side $\mathbf{h} \in \mathbb{R}^{n_{y} n_{x}}$ has the structure

$$
\mathbf{h}:=\mathbf{g}_{0} \otimes\left[\begin{array}{l}
\mathbf{b} \\
\mathbf{f}
\end{array}\right]
$$

where $\mathbf{g}_{0}$ is the first column of $G_{0}$ and the entries of $\mathbf{g}$ and $\mathbf{f}$ are defined as

$$
\begin{aligned}
\mathbf{b}(i):=-\int_{\partial D_{D}} g(\boldsymbol{x}) \boldsymbol{\phi}(\boldsymbol{x}) \cdot \boldsymbol{n} \mathrm{d} s, & i=1, \ldots, n_{u} \\
\mathbf{f}(j):=-\int_{D} f(\boldsymbol{x}) \varphi_{j}(\boldsymbol{x}) \mathrm{d} \boldsymbol{x}, & j=1, \ldots, n_{p}
\end{aligned}
$$

We can also reformulate the Kronecker system as a multi-term matrix equation of the form in (1.40). Since $G_{0}$ is the $n_{y} \times n_{y}$ identity matrix and $G_{r}$ for $r=1, \ldots, M$ are symmetric, we can write the matrix equation as

$$
\begin{equation*}
K_{0} X+\sum_{r=1}^{M} K_{r} X G_{r}=H \tag{6.8}
\end{equation*}
$$

where $X=\operatorname{array}\{\mathbf{x}\}$ and $H=\operatorname{array}\{\mathbf{h}\}$.
We now introduce two test problems for which are taken from [58].
Test problem 3. We choose the spatial domain $D:=[-1,1] \times[-1,1]$ and define the coefficient function $A^{-1}(\boldsymbol{x}, \boldsymbol{y})$ is defined as in (6.5) with

$$
a_{0}(\boldsymbol{x})=1, \quad a_{r}(\boldsymbol{x})=\sqrt{3} \sqrt{\lambda_{r}} \eta_{r}(\boldsymbol{x}), \quad r=1,2, \ldots, M,
$$

where $\left(\lambda_{r}, \eta_{r}(\boldsymbol{x})\right)$ are the eigenpairs associated with the separable exponential covariance function (2.15). We fix the standard deviation $\sigma=0.1$ and consider two cases of the boundary conditions and source function $f(\boldsymbol{x})$. In case 1 , we choose $f(\boldsymbol{x})=0$ with $p=1$ on $\{-1\} \times[-1,1], p=0$ on $\{1\} \times[-1,1]$ and homogeneous Neumann conditions for the velocity on $(-1,1) \times\{-1,1\}$. In case 2, we choose $f(\boldsymbol{x})=1$ and $p=0$ on the entire $\partial D$.

Test problem 4. We choose the spatial domain $D=[0,1] \times[0,1]$ and define $A^{-1}(\boldsymbol{x}, \boldsymbol{y})$ as in (6.5) with

$$
a_{0}(\boldsymbol{x})=1, \quad a_{r}(\boldsymbol{x})=\gamma_{r} \cos \left(2 \pi \varrho_{1}(r) x_{1}\right) \cos \left(2 \pi \varrho_{2}(r) x_{2}\right),
$$

where $\gamma=0.832 r^{-4}$ and $\varrho_{1}(r), \varrho_{2}(r)$ are defined as in Example 2.17, which is the fast decay case in Test problem 2. Again, we consider two cases of boundary conditions
and source functions $f(\boldsymbol{x})$. In case 1 , we choose $f(\boldsymbol{x})=0$ with $p=1$ on $\{0\} \times[0,1], p=$ 0 on $\{1\} \times[0,1]$ and homogeneous Neumann conditions for the velocity on $(0,1) \times\{0,1\}$. In case 2, we choose $f(\boldsymbol{x})=1$ and $p=0$ on the entire $\partial D$.

(a) Expectations

(b) Variances

Figure 6.1: Expectations and variances of components of the pressure $p$ and the velocity $\boldsymbol{u}$ when $l=6, M=7$ and $d=4$ for case 1 in Test problem 3.

In Figure 6.1, we plot approximations of expectations and variances of the pressure $p$ and the velocity $\boldsymbol{u}$ using preconditioned MINRES with stopping tolerance $10^{-8}$. The preconditioner we use is a block-diagonal matrix of the form

$$
\begin{equation*}
\mathcal{P}=I_{n_{y}} \otimes K_{S, \text { approx }}, \tag{6.9}
\end{equation*}
$$

with

$$
K_{S, \text { approx }}:=\left[\begin{array}{cc}
A_{0, \text { diag }} & 0 \\
0 & S_{0, \text { approx }}
\end{array}\right],
$$

where $A_{0, \text { diag }}$ is a diagonal of the matrix $A_{0}$ and $S_{0, \text { approx }}:=B A_{0, \text { diag }}^{-1} B^{\top}$. The theoretical eigenvalue bounds of the preconditioned linear system are analysed in [27].

For the two cases in Test problems 3-4, the right hand side $\mathbf{h}$ of the linear system or $H$ of the matrix equation are different. In [81], different pole choices are discussed when Lyapunov equations are solved using the rational Krylov projection method and the pole choices are strongly influenced by the right hand side of the matrix equation. Here, we want to investigate if the performance of Multi-RB are influenced by the right hand side of our multi-term matrix equation.

For the spatial degrees of freedom, the number $n_{u}$ associated with the velocity $\boldsymbol{u}$ is the number of finite element edges whose normal components need to be determined and the number $n_{p}$ associated with the pressure $p$ is the number of finite elements. The number $n_{u}$ depends on the boundary conditions. For example, the normal components of the finite element edges on the Neumann boundary in case 1 are zero due to the homogeneous Neumann boundary condition and so the number $n_{u}$ is less than in case 2.

### 6.1 Symmetric preconditioning

In this section, we apply symmetric preconditioning with a positive definite preconditioner to the matrix equation (6.8) for the parameter-dependent groundwater flow problem. The approach is similar to the one outlined in Section 3.2.1 for the parameterdependent linear elasticity problem but with a different preconditioner. After preconditioning, we choose shifts and then we apply Multi-RB to the shifted matrix equation.

### 6.1.1 Pre-processing phase

We consider a Schur complement based block-diagonal preconditioner, which is defined as

$$
K_{S}:=\left[\begin{array}{cc}
A_{0} & 0  \tag{6.10}\\
0 & S_{0}
\end{array}\right]
$$

where $S_{0}$ is the exact Schur complement of $K_{0}$

$$
\begin{equation*}
S_{0}:=B A_{0}^{-1} B^{\top} . \tag{6.11}
\end{equation*}
$$

Note that $K_{S, \text { approx }}$ in (6.9) is a sparse approximation of $K_{S}$. We now use the above preconditioner $K_{S}$ and apply the preconditioning strategy stated in Section 3.2.1 to the matrix equation (6.8). First, we factorise $K_{S}=L_{K_{S}} L_{K_{S}}^{\top}$ exploiting Cholesky factorisation. We then apply $L_{K_{S}}^{-1}$ on the left of both sides of the matrix equation (6.8) and replace $X$ with $\hat{X}=L_{K_{S}}^{\top} X$ to obtain

$$
\begin{equation*}
\widehat{K}_{0} \hat{X}+\sum_{r=1}^{M} \widehat{K}_{r} \widehat{X} G_{r}=\hat{H} \tag{6.12}
\end{equation*}
$$

where the matrices $\widehat{K}_{r}, \hat{H}$ are defined by

$$
\widehat{K}_{r}:=L_{K_{S}}^{-1} K_{r} L_{K_{S}}^{-\top}, \quad \widehat{H}:=L_{K_{S}}^{-1} H
$$

Similar to (3.13), we choose $M+1$ shifts $\alpha_{r}$ for $r=0,1, \ldots, M$ and rewrite (6.12) as

$$
\begin{equation*}
\widehat{X}\left(-\alpha_{0} I-\sum_{r=1}^{M} \alpha_{r} G_{r}\right)+\sum_{r=0}^{M}\left(\widehat{K}_{r}+\alpha_{r} I\right) \hat{X} G_{r}=\hat{H} \tag{6.13}
\end{equation*}
$$

The shifts are chosen so that each $\widehat{K}_{r}+\alpha_{r} I$ has positive eigenvalues.

| $r$ | Test problem 3 | Test problem 4 |
| :---: | :---: | :---: |
| 0 | $\{-0.6180,1,1.6180\}$ | $\{-0.6180,1,1.6180\}$ |
| 1 | $\{0\} \cup[0.0957,0.1471]$ | $[-0.8280,0.8280]$ |
| 2 | $[-0.0786,0.0786]$ | $[-0.0517,0.0517]$ |
| 3 | $[-0.0786,0.0786]$ | $[-0.0101,0.0101]$ |
| 4 | $[-0.0468,0.0466]$ | $[-0.0032,0.0032]$ |
| 5 | $[-0.0468,0.0466]$ | $[-0.0013,0.0013]$ |
| 6 | $[-0.0419,0.0419]$ | $[-0.0006,0.0006]$ |
| 7 | $[-0.0326,0.0326]$ | $[-0.0003,0.0003]$ |

Table 6.1: Numerical eigenvalue bounds for $\hat{K}_{r}$ in Test problems 3-4.

We now investigate the eigenvalues of $\widehat{K}_{r}$ in (6.12) and choose shifts $\alpha_{r}$ based on them. We record the extremal eigenvalues of $\hat{K}_{r}$, for $\hat{K}_{r}$, for $r=0,1, \ldots, 7$, for Test
problem 3-4 in Table 6.1. These values are computed for problems on a mesh with grid level 5 using the eigs function in MATLAB. For both test problems, the first preconditioned matrix $\widehat{K}_{0}$ has three distinct eigenvalues. For Test problem 3, the eigenvalues of $\widehat{K}_{1}$ are non-negative and the minimum and maximum eigenvalues of $\widehat{K}_{r}$, for $r=2, \ldots, 7$, are symmetric about zero. For Test problem 4, the minimum and maximum eigenvalues of $\widehat{K}_{r}$, for $r=1,2, \ldots, 7$ are also symmetric about zero. We choose $\alpha_{0}=0.5-\lambda_{0}^{\min }$ for $\hat{K}_{0}$ and $\alpha_{r}=1-\left(\lambda_{r}^{\min }+\lambda_{r}^{\max }\right) / 2$ for $\hat{K}_{r}$, for $r=1,2, \ldots, M$, where $\lambda_{r}^{\min }$ and $\lambda_{r}^{\max }$ represent the minimum and maximum eigenvalues of $\hat{K}_{r}$. For both problems, we choose $\alpha_{0}=1.6180$. For Test problem 3, we choose $\alpha_{1}=0.0735$ and $\alpha_{r}=1$ for $r=2, \ldots, M$. For Test problem 4, we choose $\alpha_{r}=1$ for $r=1,2, \ldots, M$.

Next, we apply Multi-RB to the shifted matrix equation (6.13) for Test problems 3-4.

### 6.1.2 Iterative solution phase

Recall that the details of the iterative solution phase were presented in Section 3.2.2 and the Multi-RB method was outlined in Algorithm 1. Unless otherwise stated, we make the following choice for all numerical experiments. The stopping condition for Multi-RB (the outer iteration) is $\Delta_{k}<\epsilon_{\text {outer }}$ and we choose $\epsilon_{\text {outer }}=10^{-6}$. We use the parameter-free strategy and use the backslash operator in MATLAB to construct the basis vectors and keep the $99 \%$ most significant directions. That is the singular vectors associated with the $99 \%$ largest singular values. We can impose either the Galerkin condition and solve the reduced problem (3.30) with the Kronecker form in (3.31), or impose the Petrov-Galerkin condition and solve the reduced problem (1.49) with the Kronecker form in (3.33). We use MINRES to solve these reduced problems and this is called the inner iteration. We set $Y_{0}=0$ and choose $Y_{k}=\left[Y_{k-1} ; 0\right]$ as the initial guess when we apply MINRES to the reduced problem at the $k$-th outer iteration. Let $\delta_{k}^{(s)}$ be the 2-norm of the relative residual at the $s$-th iteration of the reduced problem. The inner iteration is terminated when

$$
\begin{equation*}
\delta_{k}^{(s)}<\epsilon_{\text {inner }} \Delta_{k-1}, \tag{6.14}
\end{equation*}
$$

and we choose $\epsilon_{\text {inner }}=10^{-3}$.
Let $R_{k}$ denote the $k$-th residual of the preconditioned matrix equation (6.12) so
that $L_{K_{S}} R_{k}$ is the $k$-th residual of the unpreconditioned matrix equation (6.8). We define $R e s_{k}$ as the Frobenius norm of $k$-th unpreconditioned relative residual

$$
\operatorname{Res}_{k}:=\frac{\left\|L_{K_{S}} R_{k}\right\|_{F}}{\left\|L_{K_{S}} R_{0}\right\|_{F}}
$$

and $\operatorname{Pres}_{3, k}$ as the Frobenius norm of $k$-th preconditioned relative residual

$$
\operatorname{Pres}_{3, k}:=\frac{\left\|R_{k}\right\|_{F}}{\left\|R_{0}\right\|_{F}}
$$

For all numerical experiments, we record the final iteration count $k$, the dimension $n_{k}$ of the approximation space generated, the estimated rank of the final approximation $X_{k}$, the average inner iteration count $i$, the timing in seconds, as well as $\operatorname{Pres}_{3, k}$ and $R e s_{k}$.

Experiment 1 (varying $M$ and $d$ ) First, we fix the grid level to be 6 and apply Multi-RB with the symmetric preconditioning strategy to Test problems 3-4 with varying $M$ and $d$.

| $M$ | $d$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{3, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 3 | 56 | 32 | 120 | 46 | 8 | 6.66 e 2 | $4.69 \mathrm{e}-6$ | $4.81 \mathrm{e}-6$ |
|  | 4 | 126 | 32 | 120 | 54 | 8 | 6.48 e 2 | $4.70 \mathrm{e}-6$ | $4.82 \mathrm{e}-6$ |
|  | 5 | 252 | 32 | 120 | 55 | 9 | 6.58 e 2 | $4.70 \mathrm{e}-6$ | $4.82 \mathrm{e}-6$ |
| 7 | 3 | 120 | 44 | 191 | 87 | 8 | 1.16 e 3 | $2.50 \mathrm{e}-6$ | $1.86 \mathrm{e}-6$ |
|  | 4 | 330 | 44 | 191 | 98 | 9 | 1.18 e 3 | $2.51 \mathrm{e}-6$ | $1.87 \mathrm{e}-6$ |
|  | 5 | 792 | 44 | 191 | 99 | 9 | 1.16 e 3 | $2.51 \mathrm{e}-6$ | $1.87 \mathrm{e}-6$ |
| 13 | 3 | 560 | 21 | 104 | 98 | 8 | 9.19 e 2 | $1.74 \mathrm{e}-4$ | $1.08 \mathrm{e}-4$ |
|  | 4 | 2380 | 26 | 130 | 112 | 9 | 1.14 e 3 | $8.34 \mathrm{e}-5$ | $6.44 \mathrm{e}-5$ |
|  | 5 | 8568 | 26 | 130 | 112 | 9 | 1.16 e 3 | $8.34 \mathrm{e}-5$ | $6.44 \mathrm{e}-5$ |

Table 6.2: Numerical results for case 1 in Test problem 3 with fixed grid level 6 (giving $n_{x}=12288$ ) using Multi-RB with the symmetric preconditioning strategy.

In Tables 6.2-6.3, we record the numerical results for the two cases in Test problem 3. For both cases, we can see the number of outer iterations $k$ and the dimension $n_{k}$ are independent of the polynomial degree $d$ but depend on the number $M$ of parameters. The dimension $n_{k}$ is smaller than $n_{y}$ when $M$ and $d$ are not too small.

As $M$ and $d$ increase, the dimension $n_{k}$ approaches the estimated rank of the final approximation $X_{k}$. It requires more iterations $k$ and dimensions $n_{k}$ in case 2 than in case 1 to achieve the stopping condition.

| $M$ | $d$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time $(\mathrm{s})$ | Res $_{k}$ | Pres $_{3, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 3 | 56 | 52 | 209 | 55 | 9 | 1.09 e 3 | $5.07 \mathrm{e}-5$ | $5.66 \mathrm{e}-6$ |
|  | 4 | 126 | 52 | 209 | 91 | 9 | 1.13 e 3 | $5.09 \mathrm{e}-5$ | $5.68 \mathrm{e}-6$ |
|  | 5 | 252 | 52 | 209 | 97 | 9 | 1.08 e 3 | $5.09 \mathrm{e}-5$ | $5.68 \mathrm{e}-6$ |
| 7 | 3 | 120 | 53 | 235 | 117 | 9 | 1.40 e 3 | $1.62 \mathrm{e}-4$ | $1.25 \mathrm{e}-5$ |
|  | 4 | 330 | 53 | 235 | 161 | 9 | 1.40 e 3 | $1.62 \mathrm{e}-4$ | $1.25 \mathrm{e}-5$ |
|  | 5 | 792 | 53 | 235 | 161 | 9 | 1.40 e 3 | $1.62 \mathrm{e}-4$ | $1.25 \mathrm{e}-5$ |
| 13 | 3 | 560 | 85 | 494 | 332 | 9 | 3.72 e 3 | $2.25 \mathrm{e}-4$ | $1.64 \mathrm{e}-5$ |
|  | 4 | 2380 | 85 | 494 | 384 | 10 | 3.79 e 3 | $2.26 \mathrm{e}-4$ | $1.65 \mathrm{e}-5$ |
|  | 5 | 8568 | 85 | 494 | 389 | 10 | 3.97 e 3 | $2.26 \mathrm{e}-4$ | $1.65 \mathrm{e}-5$ |

Table 6.3: Numerical results for case 2 in Test problem 3 with fixed grid level 6 (giving $n_{x}=12416$ ) using Multi-RB with the symmetric preconditioning strategy.

| $M$ | $d$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{3, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 3 | 56 | 79 | 250 | 36 | 46 | 1.65 e 3 | $2.95 \mathrm{e}-5$ | $1.53 \mathrm{e}-5$ |
|  | 4 | 126 | 87 | 281 | 48 | 60 | 1.79 e 3 | $4.63 \mathrm{e}-5$ | $2.47 \mathrm{e}-5$ |
|  | 5 | 252 | 110 | 367 | 56 | 63 | 2.27 e 3 | $5.52 \mathrm{e}-5$ | $2.92 \mathrm{e}-5$ |
| 7 | 3 | 120 | 67 | 224 | 39 | 41 | 1.74 e 3 | $1.50 \mathrm{e}-5$ | $5.41 \mathrm{e}-6$ |
|  | 4 | 330 | 79 | 265 | 49 | 61 | 2.13 e 3 | $1.89 \mathrm{e}-5$ | $8.03 \mathrm{e}-6$ |
|  | 5 | 792 | 84 | 282 | 55 | 44 | 2.20 e 3 | $2.47 \mathrm{e}-5$ | $1.06 \mathrm{e}-5$ |
| 13 | 3 | 560 | 85 | 328 | 48 | 45 | 3.73 e 3 | $3.42 \mathrm{e}-6$ | $1.07 \mathrm{e}-6$ |
|  | 4 | 2380 | 86 | 329 | 59 | 57 | 3.99 e 3 | $6.79 \mathrm{e}-6$ | $2.25 \mathrm{e}-6$ |
|  | 5 | 8568 | 86 | 329 | 65 | 46 | 4.47 e 3 | $9.98 \mathrm{e}-6$ | $3.45 \mathrm{e}-6$ |

Table 6.4: Numerical results for case 1 in Test problem 4 with fixed grid level 6 (giving $n_{x}=12288$ ) using Multi-RB with the symmetric preconditioning strategy.

| $M$ | $d$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time $(\mathrm{s})$ | Res $_{k}$ | Pres $_{3, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 3 | 56 | 63 | 209 | 43 | 24 | 1.27 e 3 | $1.46 \mathrm{e}-4$ | $9.81 \mathrm{e}-6$ |
|  | 4 | 126 | 78 | 261 | 58 | 29 | 1.60 e 3 | $1.08 \mathrm{e}-4$ | $5.18 \mathrm{e}-6$ |
|  | 5 | 252 | 84 | 281 | 71 | 32 | 1.73 e 3 | $1.02 \mathrm{e}-4$ | $5.49 \mathrm{e}-6$ |
| 7 | 3 | 120 | 70 | 250 | 55 | 25 | 1.95 e 3 | $4.92 \mathrm{e}-5$ | $2.76 \mathrm{e}-6$ |
|  | 4 | 330 | 70 | 250 | 69 | 28 | 1.84 e 3 | $8.19 \mathrm{e}-5$ | $4.53 \mathrm{e}-6$ |
|  | 5 | 792 | 70 | 250 | 78 | 30 | 1.83 e 3 | $1.11 \mathrm{e}-4$ | $6.15 \mathrm{e}-6$ |
| 13 | 3 | 560 | 57 | 212 | 71 | 22 | 2.49 e 3 | $9.76 \mathrm{e}-5$ | $5.31 \mathrm{e}-6$ |
|  | 4 | 2380 | 57 | 212 | 83 | 24 | 2.51 e 3 | $1.67 \mathrm{e}-4$ | $8.96 \mathrm{e}-6$ |
|  | 5 | 8568 | 57 | 212 | 93 | 26 | 2.62 e 3 | $2.30 \mathrm{e}-4$ | $1.24 \mathrm{e}-5$ |

Table 6.5: Numerical results for case 2 in Test problem 4 with fixed grid level 6 (giving $n_{x}=12416$ ) using Multi-RB with the symmetric preconditioning strategy.

In Tables 6.4-6.5, we record the numerical results for the two cases in Test problem 4. For case 1, we choose $\epsilon_{\text {inner }}=10^{-4}$ because Multi-RB is stagnated at the 2nd outer iteration when we choose $\epsilon_{\text {inner }}=10^{-3}$. This happens because the inner stopping tolerance is too large for the reduced problem so that we finally obtain $Y_{2}=\left[Y_{1} ; 0\right]$ at the 2nd outer iteration. Again, for both cases, we can see the number of outer iterations $k$ and the dimension $n_{k}$ are independent of the polynomial degree $d$ but slightly depend on the number $M$ of parameters. The dimension $n_{k}$ is smaller than $n_{y}$ when $M$ and $d$ are not too small but unlike for Test problem 3, it is always much greater than the estimated rank of $X_{k}$. It requires more outer iterations $k$ and dimension $n_{k}$ in case 1 than in case 2.

For all cases, Res $_{k}$ is greater than $\operatorname{Pres}_{3, k}$ and Pres $_{3, k}$ has a good agreement with the successive relative difference $\Delta_{k}$. For both Test problems, we see that $\operatorname{Res}_{k}$ is roughly one order of magnitude higher than $\Delta_{k}$ in case 1 and $R e s_{k}$ is roughly two orders of magnitude higher than $\Delta_{k}$ in case 2.

Experiment 2 (varying grid level) We apply Multi-RB to the two cases in Test problems 3-4 on finite element meshes with varying grid level $l=5,6,7$. We fix $M=7$ and $d=4$ (giving $n_{y}=330$ ) and record the numerical results in Tables 6.6-6.9.

| $l$ | $n_{x}$ | $k$ | $n_{k}$ | rank | $i$ | time $(\mathrm{s})$ | Res $_{k}$ | Pres $_{3, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 3072 | 44 | 193 | 105 | 9 | 4.53 e 1 | $6.34 \mathrm{e}-6$ | $2.83 \mathrm{e}-6$ |
| 6 | 12288 | 44 | 191 | 98 | 9 | 1.18 e 3 | $2.51 \mathrm{e}-6$ | $1.87 \mathrm{e}-6$ |
| 7 | 49152 | 13 | 53 | 46 | 7 | 1.08 e 4 | $1.06 \mathrm{e}-4$ | $6.64 \mathrm{e}-5$ |

Table 6.6: Numerical results for case 1 in Test problem 3 with varying grid level $l$ when $M=7$ and $d=4$ using the symmetric preconditioning strategy.

| $l$ | $n_{x}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{3, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 3136 | 52 | 231 | 161 | 9 | 5.25 e 1 | $1.47 \mathrm{e}-4$ | $1.47 \mathrm{e}-5$ |
| 6 | 12416 | 53 | 235 | 161 | 9 | 1.18 e 3 | $2.51 \mathrm{e}-6$ | $1.87 \mathrm{e}-6$ |
| 7 | 49408 | 53 | 236 | 160 | 12 | 4.42 e 4 | $1.83 \mathrm{e}-4$ | $1.17 \mathrm{e}-5$ |

Table 6.7: Numerical results for case 2 in Test problem 3 with varying grid level $l$ when $M=7$ and $d=4$ using the symmetric preconditioning strategy.

| $l$ | $n_{x}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{3, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 3072 | 84 | 287 | 53 | 31 | 1.03 e 2 | $3.29 \mathrm{e}-5$ | $9.90 \mathrm{e}-6$ |
| 6 | 12288 | 79 | 265 | 49 | 61 | 2.13 e 3 | $1.89 \mathrm{e}-5$ | $8.03 \mathrm{e}-6$ |
| 7 | 49152 | 67 | 217 | 46 | 49 | 5.62 e 4 | $1.25 \mathrm{e}-5$ | $5.76 \mathrm{e}-6$ |

Table 6.8: Numerical results for case 1 in Test problem 4 with varying grid level $l$ when $M=7$ and $d=4$ using the symmetric preconditioning strategy.

| $l$ | $n_{x}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{3, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 3136 | 57 | 205 | 68 | 23 | 6.42 e 1 | $1.51 \mathrm{e}-4$ | $1.19 \mathrm{e}-5$ |
| 6 | 12416 | 70 | 250 | 69 | 28 | 1.84 e 3 | $8.19 \mathrm{e}-5$ | $4.53 \mathrm{e}-6$ |
| 7 | 49408 | 67 | 234 | 69 | 41 | 5.58 e 4 | $1.28 \mathrm{e}-4$ | $5.36 \mathrm{e}-6$ |

Table 6.9: Numerical results for case 2 in Test problem 4 with varying grid level $l$ when $M=7$ and $d=4$ using the symmetric preconditioning strategy.

In most cases, we observe that the number of outer iterations $k$ and the dimension $n_{k}$ are nearly independent of grid level $l$. However, the number of outer iterations $k$
and dimension $n_{k}$ for case 1 in Test problem 3 when $l=7$ is clearly less than those when $l=5,6$. We observed the plots of expectations and variances of approximations obtained for all problems in Table 6.6 using MATLAB. The plots for $l=5,6$ are consistent with Figure 6.1 but the plot for $l=7$ is not consistent. In Figure 6.2, we plot the convergence history of $\Delta_{k}$ for these three problems. We see that there is a big drop at the last iteration for $l=7$. For $l=5,6$, it happens at the same iteration but $\Delta_{k}$ does not reach $10^{-6}$ at that iteration. To fix this issue, we solve the problem with grid level $l=7$ with a smaller tolerance $\epsilon_{\text {outer }}=10^{-7}$ and checked plots for the final approximation again. The plots are consistent with Figure 6.1.


Figure 6.2: Convergence history of $\Delta_{k}$ for case 1 in Test problem 3 with $M=7$ and $d=4$.

Experiment 3 (choice of projection strategy) We plot the convergence history of the relative difference $\Delta_{k}$, the preconditioned relative residual Pres $_{3, k}$ and unpreconditioned relative residual $\operatorname{Res}_{k}$ for some small problems with spatial grid level $5, M=7$ and $d=4$ using two types of projection strategy in Figures 6.3-6.6. In this experiment, we set $\epsilon_{\text {outer }}=10^{-8}$.

First, we apply the Galerkin condition to $R_{k}$ and the convergence history is shown in Figures 6.3-6.4. We see that the preconditioned relative residual $\operatorname{Pres}_{3, k}$ and unpreconditioned relative residual $R e s_{k}$ are not monotonically decreasing. The relative residuals Res $_{k}$ and Pres $_{3, k}$ for Test problem 4 are more erratic than Test problem 3. The successive relative difference $\Delta_{k}$ is the smallest and its history is the most erratic. The unpreconditioned relative residual $R e s_{k}$ is roughly one order of magnitude higher than the successive relative difference $\Delta_{k}$ in case 1 and one to two orders of magnitude higher than $\Delta_{k}$ in case 2. The preconditioned relative residual $\operatorname{Pres}_{3, k}$ has
a good agreement with the successive relative difference $\Delta_{k}$. For case 1 in both test problems, the unpreconditioned relative residual $R e s_{k}$ is closer to the preconditioned relative residual Pres $_{3, k}$ than for case 2. Recall that the right hand sides of the matrix equations for these two cases are different. This indicates that the accuracy is influenced by the right hand side of the matrix equation.

(a) case 1

(b) case 2

Figure 6.3: Convergence history for Test problem 3 when we apply the Galerkin condition with $l=5, M=7$ and $d=4$.

(a) case 1

(b) case 2

Figure 6.4: Convergence history for Test problem 4 when we apply the Galerkin condition with $l=5, M=7$ and $d=4$.

Next, we apply the Petrov-Galerkin condition to $R_{k}$ and plot the convergence history in Figures 6.5-6.6.


Figure 6.5: Convergence history for Test problem 3 when we apply the PetrovGalerkin condition with $l=5, M=7$ and $d=4$.

(a) case 1

(b) case 2

Figure 6.6: Convergence history for Test problem 4 when we apply the PetrovGalerkin condition with $l=5, M=7$ and $d=4$.

The preconditioned relative residual Pres $_{3, k}$ is now monotonically decreasing. The relative difference $\Delta_{k}$ also becomes less erratic than when we apply the Galerkin condition. For all cases, Pres $_{3, k}$ has a good agreement with $\Delta_{k}$. The unpreconditioned relative residual $R e s_{k}$ is still roughly one order of magnitude higher than the successive relative difference $\Delta_{k}$ in case 1 and one to two orders of magnitude higher than $\Delta_{k}$ in case 2 for both test problems. In other words, Res $_{k}$ is closer to $\operatorname{Pres}_{3, k}$ and $\Delta_{k}$ in case 1 than in case 2 for both test problems, which also happens when we apply the Galerkin condition. This indicates that the accuracy is still influenced by the right
hand side vector.

| $M$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time $(\mathrm{s})$ | Res $_{k}$ | Pres $_{3, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 126 | 32 | 120 | 55 | 8 | 6.56 e 2 | $5.21 \mathrm{e}-6$ | $3.47 \mathrm{e}-6$ |
| 7 | 330 | 44 | 191 | 88 | 8 | 1.15 e 3 | $2.08 \mathrm{e}-6$ | $1.31 \mathrm{e}-6$ |
| 13 | 2380 | 21 | 104 | 102 | 7 | 9.47 e 2 | $1.44 \mathrm{e}-4$ | $9.10 \mathrm{e}-5$ |

Table 6.10: Numerical results for case 1 in Test problem 3 with varying $M$ when $l=6$ and $d=4$ when we apply the Petrov-Galerkin condition.

| $M$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{3, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 126 | 52 | 209 | 88 | 9 | 1.09 e 3 | $4.18 \mathrm{e}-5$ | $3.97 \mathrm{e}-6$ |
| 7 | 330 | 53 | 235 | 148 | 9 | 1.40 e 3 | $1.06 \mathrm{e}-4$ | $8.77 \mathrm{e}-6$ |
| 13 | 2380 | 85 | 494 | 373 | 9 | 4.52 e 3 | $1.41 \mathrm{e}-4$ | $1.14 \mathrm{e}-5$ |

Table 6.11: Numerical results for case 2 in Test problem 3 with varying $M$ when $l=6$ and $d=4$ when we apply the Petrov-Galerkin condition.

In Tables 6.10-6.13, we record the numerical results obtained when we apply the Petrov-Galerkin condition for Test problems 3-4 with fixed $d=4$. Compared with the numerical results in Tables 6.2-6.5 when we apply the Galerkin condition, we see the number of outer iterations $k$ and the dimension $n_{k}$ is barely changed for Test problem 3 but somewhat decrease for case 1 in Test problem 4. The timings and the accuracy in terms of $\operatorname{Res}_{k}$ and $\operatorname{Pres}_{3, k}$ are also barely changed.

| $M$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{3, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 126 | 74 | 237 | 46 | 26 | 1.58 e 3 | $4.46 \mathrm{e}-5$ | $2.35 \mathrm{e}-5$ |
| 7 | 330 | 48 | 153 | 51 | 25 | 1.25 e 3 | $9.51 \mathrm{e}-5$ | $4.78 \mathrm{e}-5$ |
| 13 | 2380 | 57 | 195 | 61 | 26 | 2.92 e 3 | $6.16 \mathrm{e}-5$ | $2.93 \mathrm{e}-5$ |

Table 6.12: Numerical results for case 1 in Test problem 4 with varying $M$ when $l=6$ and $d=4$ when we apply the Petrov-Galerkin condition.

| $M$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{3, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 126 | 63 | 209 | 57 | 21 | 1.33 e 3 | $1.67 \mathrm{e}-4$ | $1.20 \mathrm{e}-5$ |
| 7 | 330 | 67 | 239 | 69 | 21 | 1.76 e 3 | $6.03 \mathrm{e}-5$ | $3.99 \mathrm{e}-6$ |
| 13 | 2380 | 57 | 212 | 82 | 19 | 3.03 e 3 | $1.12 \mathrm{e}-4$ | $6.90 \mathrm{e}-6$ |

Table 6.13: Numerical results for case 2 in Test problem 4 with varying $M$ when $l=6$ and $d=4$ when we apply the Petrov-Galerkin condition.

Experiment 4 (large problems) We also applied Multi-RB to some larger problems with $l=9, M=15$ and $d=4$, which gives the total problem size about $3.05 e 9 \times 3.05 e 9$. As for the linear elasticity problem in Section 4.2 , the machine we use runs out of memory again. This is due to the storage cost for the dense block $S_{0}$ in $K_{S}$ when a fine finite element mesh is used.

### 6.2 Left preconditioning

In this section, we use $K_{0}$ as the preconditioner and apply the left preconditioning strategy to the matrix equation (6.8). In [58], Multi-RB with the same preconditioning strategy are also applied to the matrix equation associated with the problem (6.1)-(6.2). In this section, we use the same strategy and we will investigate the unpreconditioned relative residual and the convergence history of three different errors for some small problems. After preconditioning, we choose shifts and apply Multi-RB to the modified equation.

### 6.2.1 Pre-processing phase

We choose $K_{0}$ as the preconditioner and apply $K_{0}^{-1}$ on both sides of the matrix equation (6.8)

$$
\begin{equation*}
X+\sum_{r=1}^{M} \widetilde{K}_{r} X G_{r}=\widetilde{H} \tag{6.15}
\end{equation*}
$$

where $\widetilde{K}_{r}$ and $\widetilde{H}$ are defined by

$$
\widetilde{K}_{r}:=K_{0}^{-1} K_{r}, \quad \widetilde{H}:=K_{0}^{-1} H .
$$

Similar to (3.20), we choose $M$ shifts $\alpha_{r}$ to ensure that the eigenvalues of the matrices $\widetilde{K}_{r}+\alpha_{r} I$ are positive for each $r=1,2, \ldots, M$. This gives

$$
\begin{equation*}
X\left(I-\sum_{r=1}^{M} \alpha_{r} G_{r}\right)+\sum_{r=1}^{M}\left(\widetilde{K}_{r}+\alpha_{r} I\right) X G_{r}=\widetilde{H} . \tag{6.16}
\end{equation*}
$$

We now investigate eigenvalues of for $\widetilde{K}_{r}$ in (6.15) and choose shifts $\alpha_{r}$ based on them. In Table 6.14, we record the extremal eigenvalues for Test problems 3-4. These eigenvalues are computed for problems on a mesh with grid level 5 using the eigs function in MATLAB. Compared with the extremal eigenvalues of $\widehat{K}_{r}$ in Table 6.1, we see that they are the same for $r=1,2, \ldots, M$. However, there is an extra term $\hat{K}_{0}$ with three distinct eigenvalues in the symmetrically preconditioned matrix equation (6.12). Here, we choose $\alpha_{r}=1-\left(\lambda_{r}^{\min }+\lambda_{r}^{\max }\right) / 2$ for $\widetilde{K}_{r}$, for $r=1,2, \ldots, M$. For Test problem 3, we choose $\alpha_{1}=0.0735$ and $\alpha_{r}=1$ for $r=2, \ldots, M$. For Test problem 4, we choose $\alpha_{r}=1$ for $r=1,2, \ldots, M$.

| $r$ | Test problem 3 | Test problem 4 |
| :---: | :---: | :---: |
| 1 | $\{0\} \cup[0.0970,0.1471]$ | $[-0.8280,0.8280]$ |
| 2 | $[-0.0785,0.0785]$ | $[-0.0517,0.0517]$ |
| 3 | $[-0.0784,0.0784]$ | $[-0.0101,0.0101]$ |
| 4 | $[-0.0468,0.0466]$ | $[-0.0032,0.0032]$ |
| 5 | $[-0.0466,0.0466]$ | $[-0.0013,0.0013]$ |
| 6 | $[-0.0418,0.0418]$ | $[-0.0006,0.0006]$ |
| 7 | $[-0.0326,0.0326]$ | $[-0.0003,0.0003]$ |

Table 6.14: Numerical eigenvalue bounds for $\widetilde{K}_{r}$ in Test problems 3-4.

Next, we apply Multi-RB to the shifted matrix equation (6.16) for Test problems 3-4.

### 6.2.2 Iterative solution phase

We now apply Multi-RB to (6.16). For all numerical experiments we use the same settings as described in Section 5.2. We terminate the outer iteration when $\Delta_{k}<$ $\epsilon_{\text {outer }}$ where $\epsilon_{\text {outer }}=10^{-6}$. We use the parameter-free strategy and use the backslash operator in MATLAB to construct the basis vectors and keep the $99 \%$ most significant
directions. That is the singular vectors associated with the $99 \%$ largest singular values. We impose the Galerkin condition at each iteration and use QMR to solve the reduced problem (3.30). We set $Y_{0}=0$ and choose $Y_{k}=\left[Y_{k-1} ; 0\right]$ as the initial guess when we apply QMR to the reduced problem at the $k$-th outer iteration. Let $\delta_{k}^{(s)}$ be the 2-norm of relative residual at the $s$-th inner iteration of the reduced problem. The stopping condition we use for the inner iteration is (4.1) and we choose $\epsilon_{\text {inner }}=10^{-3}$.

We record the number of outer iterations $k$, the dimension $n_{k}$ of the approximation space generated, the estimated rank of the final approximation $X_{k}$, the average inner iteration count $i$ and the timing in seconds. We also record the final unpreconditioned relative residual $\operatorname{Res}_{k}$

$$
\operatorname{Res}_{k}:=\frac{\left\|K_{0} R_{k}\right\|_{F}}{\left\|K_{0} R_{k}\right\|_{F}}
$$

where $R_{k}$ is the $k$-th residual of the preconditioned matrix equation (6.15) with the preconditioner $K_{0}$ and $K_{0} R_{k}$ is the $k$-th residual of the unpreconditioned matrix equation (6.8), as well as the final preconditioned relative residual Pres $_{4, k}$

$$
\operatorname{Pres}_{4, k}:=\frac{\left\|R_{k}\right\|_{F}}{\left\|R_{0}\right\|_{F}}
$$

Experiment 1 (varying $M$ and $d$ ) First, we fix the grid level to be 6 and apply Multi-RB to the matrix equation (6.16) for Test problems 3-4.

In Tables 6.15-6.16, we record numerical results for Test problem 3. For both cases, the number of outer iterations $k$ and the dimension $n_{k}$ are independent of the polynomial degree $d$ but slightly depend on the number $M$ of parameters. The dimension $n_{k}$ is smaller than $n_{y}$ when $M$ and $d$ are not too small and is close to the estimated rank of the final approximation $X_{k}$. The values of $k$ and $n_{k}$ are very similar for the two cases.

In Tables 6.17-6.18, we record numerical results for Test problem 4. The number of outer iterations $k$ and the dimension $n_{k}$ are still independent of the polynomial degree $d$ but depend on the number $M$ of parameters. The dimension $n_{k}$ is smaller than $n_{y}$ when $M$ and $d$ are not too small but it is always greater than the estimated rank of $X_{k}$. The values of $k$ and $n_{k}$ are still similar for the two cases.

| $M$ | $d$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{4, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 3 | 56 | 7 | 35 | 34 | 5 | 3.92 e 0 | $8.34 \mathrm{e}-6$ | $2.01 \mathrm{e}-6$ |
|  | 4 | 126 | 9 | 45 | 40 | 6 | 5.08 e 0 | $5.51 \mathrm{e}-6$ | $1.43 \mathrm{e}-6$ |
|  | 5 | 252 | 9 | 45 | 40 | 6 | 5.14 e 0 | $5.51 \mathrm{e}-6$ | $1.43 \mathrm{e}-6$ |
| 7 | 3 | 120 | 9 | 63 | 59 | 6 | 7.99 e 0 | $2.35 \mathrm{e}-6$ | $5.99 \mathrm{e}-7$ |
|  | 4 | 330 | 9 | 63 | 59 | 6 | 8.19 e 0 | $2.39 \mathrm{e}-6$ | $6.14 \mathrm{e}-7$ |
|  | 5 | 792 | 9 | 63 | 59 | 6 | 8.77 e 0 | $2.40 \mathrm{e}-6$ | $6.15 \mathrm{e}-7$ |
| 13 | 3 | 560 | 14 | 163 | 145 | 7 | 3.36 e 1 | $1.06 \mathrm{e}-6$ | $2.82 \mathrm{e}-7$ |
|  | 4 | 2380 | 14 | 163 | 147 | 7 | 6.00 e 1 | $1.10 \mathrm{e}-6$ | $2.91 \mathrm{e}-7$ |
|  | 5 | 8568 | 14 | 163 | 147 | 7 | 1.44 e 2 | $1.10 \mathrm{e}-6$ | $2.92 \mathrm{e}-7$ |

Table 6.15: Numerical results for case 1 in Test problem 3 with fixed grid level 6 (giving $n_{x}=12288$ ) using Multi-RB with the left preconditioning strategy.

| $M$ | $d$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{4, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 3 | 56 | 9 | 46 | 40 | 5 | 5.05 e 0 | $1.80 \mathrm{e}-4$ | $1.33 \mathrm{e}-6$ |
|  | 4 | 126 | 9 | 46 | 44 | 6 | 5.10 e 0 | $1.84 \mathrm{e}-4$ | $1.35 \mathrm{e}-6$ |
|  | 5 | 252 | 9 | 46 | 44 | 6 | 5.30 e 0 | $1.84 \mathrm{e}-4$ | $1.35 \mathrm{e}-6$ |
| 7 | 3 | 120 | 9 | 64 | 64 | 6 | 7.81 e 0 | $2.36 \mathrm{e}-4$ | $1.60 \mathrm{e}-6$ |
|  | 4 | 330 | 9 | 64 | 64 | 6 | 7.95 e 0 | $2.41 \mathrm{e}-4$ | $1.63 \mathrm{e}-6$ |
|  | 5 | 792 | 9 | 64 | 64 | 6 | 8.93 e 0 | $2.41 \mathrm{e}-4$ | $1.63 \mathrm{e}-6$ |
| 13 | 3 | 560 | 14 | 172 | 172 | 6 | 3.52 e 1 | $7.23 \mathrm{e}-5$ | $4.87 \mathrm{e}-7$ |
|  | 4 | 2380 | 14 | 172 | 172 | 7 | 6.36 e 1 | $7.53 \mathrm{e}-5$ | $5.08 \mathrm{e}-7$ |
|  | 5 | 8568 | 14 | 172 | 172 | 7 | 1.64 e 2 | $7.55 \mathrm{e}-5$ | $5.09 \mathrm{e}-7$ |

Table 6.16: Numerical results for case 2 in Test problem 3 with fixed grid level 6 (giving $n_{x}=12416$ ) using Multi-RB with the left preconditioning strategy.

For both test problems, the preconditioned relative residual Pres $_{4, k}$ has a good agreement with $\Delta_{k}$. The unpreconditioned relative residual $R e s_{k}$ has a good agreement with the successive relative difference $\Delta_{k}$ in case 1 and the unpreconditioned relative residual $\operatorname{Res}_{k}$ is roughly two orders of magnitude higher than the successive relative difference $\Delta_{k}$ in case 2. The unpreconditioned relative residual Res ${ }_{k}$ in case 1 is closer
to $\Delta_{k}$ than in case 2, which also happens when we use the symmetric preconditioning strategy. However, the values of $k$ and $n_{k}$ are more similar for case 1 and case 2 than in the symmetric symmetric preconditioning case.

| $M$ | $d$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{4, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 3 | 56 | 19 | 66 | 30 | 15 | 8.82 e 0 | $4.76 \mathrm{e}-5$ | $1.04 \mathrm{e}-5$ |
|  | 4 | 126 | 38 | 157 | 41 | 20 | 2.47 e 1 | $9.21 \mathrm{e}-6$ | $2.13 \mathrm{e}-6$ |
|  | 5 | 252 | 43 | 181 | 47 | 21 | 3.73 e 1 | $7.05 \mathrm{e}-6$ | $1.60 \mathrm{e}-6$ |
| 7 | 3 | 120 | 15 | 67 | 33 | 15 | 9.66 e 0 | $1.61 \mathrm{e}-5$ | $3.54 \mathrm{e}-6$ |
|  | 4 | 330 | 22 | 104 | 42 | 18 | 1.90 e 1 | $3.99 \mathrm{e}-6$ | $8.86 \mathrm{e}-7$ |
|  | 5 | 792 | 22 | 104 | 47 | 18 | 2.50 e 1 | $6.18 \mathrm{e}-6$ | $1.38 \mathrm{e}-6$ |
| 13 | 3 | 560 | 20 | 138 | 43 | 16 | 3.82 e 1 | $1.10 \mathrm{e}-6$ | $2.44 \mathrm{e}-7$ |
|  | 4 | 2380 | 20 | 138 | 48 | 17 | 8.45 e 1 | $2.53 \mathrm{e}-6$ | $5.62 \mathrm{e}-7$ |
|  | 5 | 8568 | 20 | 138 | 51 | 18 | 2.67 e 2 | $3.99 \mathrm{e}-6$ | $8.90 \mathrm{e}-7$ |

Table 6.17: Numerical results for case 1 in Test problem 4 with fixed grid level 6 (giving $n_{x}=12288$ ) using Multi-RB with the left preconditioning strategy.

| $M$ | $d$ | $n_{y}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{4, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 3 | 56 | 30 | 122 | 34 | 16 | 1.58 e 1 | $1.89 \mathrm{e}-4$ | $1.18 \mathrm{e}-6$ |
|  | 4 | 126 | 32 | 128 | 45 | 18 | 1.91 e 1 | $2.23 \mathrm{e}-4$ | $1.40 \mathrm{e}-6$ |
|  | 5 | 252 | 32 | 128 | 52 | 19 | 2.27 e 1 | $2.70 \mathrm{e}-4$ | $1.69 \mathrm{e}-6$ |
| 7 | 3 | 120 | 21 | 93 | 41 | 15 | 1.38 e 1 | $2.40 \mathrm{e}-4$ | $1.56 \mathrm{e}-6$ |
|  | 4 | 330 | 21 | 93 | 50 | 16 | 1.61 e 1 | $3.80 \mathrm{e}-4$ | $2.47 \mathrm{e}-6$ |
|  | 5 | 792 | 30 | 146 | 56 | 19 | 4.71 e 1 | $1.18 \mathrm{e}-4$ | $7.17 \mathrm{e}-7$ |
| 13 | 3 | 560 | 21 | 117 | 52 | 15 | 3.40 e 1 | $1.27 \mathrm{e}-4$ | $8.04 \mathrm{e}-7$ |
|  | 4 | 2380 | 23 | 137 | 62 | 17 | 8.92 e 1 | $1.07 \mathrm{e}-4$ | $6.60 \mathrm{e}-7$ |
|  | 5 | 8568 | 23 | 137 | 66 | 18 | 2.91 e 2 | $1.56 \mathrm{e}-4$ | $9.66 \mathrm{e}-7$ |

Table 6.18: Numerical results for case 2 in Test problem 4 with fixed grid level 6 (giving $n_{x}=12416$ ) using Multi-RB with the left preconditioning strategy.

Experiment 2 (varying grid level) We now apply Multi-RB to Test problems $\mathbf{3} \mathbf{- 4}$ on finite element meshes with varying grid level $l=5,6,7$. The numerical results
for fixed $M=7$ and $d=4$ are recorded in Tables 6.19-6.22. We see that the number of outer iterations $k$ and the dimension $n_{k}$ are nearly independent of grid level $l$.

| $l$ | $n_{x}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{4, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 3072 | 9 | 63 | 60 | 6 | 2.00 e 0 | $2.98 \mathrm{e}-6$ | $9.81 \mathrm{e}-7$ |
| 6 | 12288 | 9 | 63 | 59 | 6 | 8.19 e 0 | $2.39 \mathrm{e}-6$ | $6.14 \mathrm{e}-7$ |
| 7 | 49152 | 9 | 63 | 56 | 6 | 3.78 e 1 | $1.83 \mathrm{e}-6$ | $3.84 \mathrm{e}-7$ |

Table 6.19: Numerical results for case 1 in Test problem 3 with varying grid level $l$ when $M=7$ and $d=4$ using the left preconditioning strategy.

| $l$ | $n_{x}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{4, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 3136 | 13 | 92 | 89 | 7 | 3.30 e 0 | $4.20 \mathrm{e}-5$ | $1.12 \mathrm{e}-6$ |
| 6 | 12416 | 9 | 64 | 64 | 6 | 7.95 e 0 | $2.41 \mathrm{e}-4$ | $1.63 \mathrm{e}-6$ |
| 7 | 49408 | 9 | 64 | 64 | 6 | 3.84 e 1 | $5.48 \mathrm{e}-4$ | $9.59 \mathrm{e}-7$ |

Table 6.20: Numerical results for case 2 in Test problem 3 with varying grid level $l$ when $M=7$ and $d=4$ using the left preconditioning strategy.

| $l$ | $n_{x}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{4, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 3072 | 19 | 85 | 44 | 16 | 4.82 e 0 | $6.67 \mathrm{e}-6$ | $2.05 \mathrm{e}-6$ |
| 6 | 12288 | 22 | 104 | 42 | 18 | 1.90 e 1 | $3.99 \mathrm{e}-6$ | $8.86 \mathrm{e}-7$ |
| 7 | 49152 | 19 | 83 | 37 | 17 | 6.57 e 1 | $7.65 \mathrm{e}-6$ | $1.21 \mathrm{e}-6$ |

Table 6.21: Numerical results for case 1 in Test problem 4 with varying grid level $l$ when $M=7$ and $d=4$ using the left preconditioning strategy.

| $l$ | $n_{x}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) | Res $_{k}$ | Pres $_{4, k}$ |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 5 | 3072 | 29 | 134 | 52 | 17 | 1.12 e 1 | $7.32 \mathrm{e}-5$ | $1.56 \mathrm{e}-6$ |
| 6 | 12288 | 21 | 93 | 50 | 16 | 1.61 e 1 | $3.80 \mathrm{e}-4$ | $2.47 \mathrm{e}-6$ |
| 7 | 49152 | 22 | 103 | 48 | 17 | 7.83 e 1 | $6.46 \mathrm{e}-4$ | $1.10 \mathrm{e}-6$ |

Table 6.22: Numerical results for case 2 in Test problem 4 with varying grid level $l$ when $M=7$ and $d=4$ using the left preconditioning strategy.

Experiment 3 (convergence history) We plot the convergence history of the preconditioned relative residual Pres $_{4, k}$, the unpreconditioned relative residual Res $_{k}$ and the successive relative difference $\Delta_{k}$ against the dimension $n_{k}$ in Figures 6.7-6.8.


Figure 6.7: Convergence history for Test problem 3 with $l=5, M=7$ and $d=4$ using the left preconditioning strategy.


Figure 6.8: Convergence history for Test problem 4 with $l=5, M=7$ and $d=4$ using the left preconditioning strategy.

The stopping tolerance in all cases is $\epsilon_{\text {outer }}=10^{-8}$. For case 1 in both test problems, the unpreconditioned relative residual $\operatorname{Res}_{k}$ is roughly one order of magnitude higher than the successive relative difference $\Delta_{k}$. For case 2 in both test problems, the unpreconditioned relative residual $\operatorname{Res}_{k}$ is roughly two orders of magnitude higher than the successive relative difference $\Delta_{k}$. As we observed for the symmetric preconditioning
strategy, Res $s_{k}$ is closer to $\Delta_{k}$ in case 1 than in case 2 for both test problems. The accuracy is influenced by the right hand side. Compared with the convergence history for the same problems in Figures 6.3-6.4 using the symmetric preconditioning strategy and the Galerkin condition, the convergence history is much more stable.

The convergence of the successive relative difference $\Delta_{k}$ is investigated with stopping condition $\epsilon_{\text {outer }}=10^{-6}$ for Test problems 3-4 with different values of $M$ in [58]. We also checked the convergence history of $\Delta_{k}$ for those problems and the plots are consistent with those in [58].

Experiment 4 (large problems) We now apply Multi-RB with the left preconditioning strategy to some large problems with $l=9, M=15$ and $d=4$. The total problem size is $3,048,210,432 \times 3,048,210,432$ in case 1 or $3,052,179,456 \times$ $3,052,179,456$ in case 2. The stopping tolerance is chosen to be $\epsilon_{\text {outer }}=10^{-8}$. We record the numerical results in Table 6.23. We do not run out of memory when using Multi-RB with the left preconditioning strategy.

| Test problem | case | $n_{x}$ | $k$ | $n_{k}$ | rank | $i$ | time (s) |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 3 | 1 | 786432 | 26 | 341 | 339 | 9 | 8.77 e 3 |
|  | 2 | 787456 | 27 | 374 | 374 | 8 | 9.29 e 3 |
| 4 | 1 | 786432 | 29 | 251 | 82 | 22 | 7.97 e 3 |
|  | 2 | 787456 | 27 | 229 | 104 | 20 | 7.32 e 3 |

Table 6.23: Numerical results obtained using Multi-RB with the left preconditioning strategy for Test problems $\mathbf{3} \mathbf{- 4}$ with $l=9, M=15$ and $d=4$.

### 6.3 Conclusions

In this chapter, we apply Multi-RB with two preconditioning strategies to the matrix equation associated with (6.1)-(6.4). From numerical results we obtained, there are some similarities and differences with observations for the parameter-dependent linear elasticity problem.

The similarities are:

1. For large problems that cause memory issues when we use MINRES, we are now able to solve them using Multi-RB with the left preconditioning strategy.

However, it was not feasible to use the suggested symmetric preconditioning strategy because the chosen preconditioner has a dense block whose storage costs become problematic when the finite element grid level is large.
2. For both preconditioning strategies, the number of outer iterations $k$ and the dimension $n_{k}$ required to meet the stopping condition are nearly independent of the finite element grid level $l$ and the polynomial degree $d$ but depend on the number $M$ of parameters. When $M$ and $d$ are not small, the dimension $n_{k}$ is less than $n_{y}$ and much less than $n_{x}$.
3. When we apply Multi-RB using the symmetric preconditioning strategy with the Galerkin condition, we observe that the unpreconditioned relative residual $R e s_{k}$, the preconditioned relative residual and the relative difference $\Delta_{k}$ are erratic. If we impose the Petrov-Galerkin condition, the preconditioned relative residual decreases monotonically and so does $R e s_{k}$. When we use the left preconditioning strategy with the Galerkin condition, the errors all decrease monotonically. In particular, the relative difference $\Delta_{k}$ (which controls the stopping condition) is more stable than when using symmetric preconditioning strategy.

The differences are:

1. When we use the left preconditioning strategy for the parameter-dependent linear elasticity problem, the difference between unpreconditioned relative residual $R e s_{k}$ and the relative difference $\Delta_{k}$ is much greater than when using the symmetric preconditioning strategy. However, for the parameter-dependent groundwater flow problem, the difference between $\operatorname{Res}_{k}$ and $\Delta_{k}$ for the two preconditioning strategies are nearly the same.
2. For the parameter-dependent linear elasticity problem, the dimension $n_{k}$ required to meet the stopping condition is much greater than the estimated rank of $X_{k}$ when using the symmetric preconditioning strategy. The dimension $n_{k}$ is close to the estimated rank of $X_{k}$ for Test problem 1 but larger than the estimated rank of $X_{k}$ for Test problem 2 when using the left preconditioning strategy. For both preconditioning strategies when solving the parameter-dependent groundwater flow problem, the dimension $n_{k}$ approaches the estimated rank of $X_{k}$ when $M$ and
$d$ are not small for Test problem 3 but is always larger than the estimated rank of $X_{k}$ for Test problem 4. For Test problems 3-4, the difference between $R e s_{k}$ and $\Delta_{k}$ in case 1 is smaller than that in case 2 . This suggests that the accuracy may be influenced by the right-hand side vector.

## Chapter 7

## Conclusions

In this thesis, we extended the Multi-RB algorithm that was proposed in [64] for linear multi-term matrix equations (LMTMEs) associated with symmetric and positive definite problems to LMTMEs associated with symmetric and indefinite problems. The matrix equations we investigated are associated with the saddle point structured linear systems that arise when we apply stochastic Galerkin mixed finite element methods (SGMFEMs) to PDEs with inputs that depend on uncertain parameters. These problems are challenging to solve due to their extremely large size. Indeed, we are unable to solve them on standard desktop computers using standard Krylov subspace methods due to memory constraints.

We applied SGMFEMs to two parameter-dependent PDE problems, namely a parameter-dependent linear elasticity problem and a parameter-dependent groundwater flow problem. Before applying the Multi-RB algorithm, we preconditioned the associated matrix equations using two preconditioning strategies, a symmetric preconditioning strategy and a left preconditioning strategy. After preconditioning, we introduced shifts to ensure that the eigenvalues of all the left matrices appearing in the modified matrix equation are positive. The approximation space we chose for MultiRB is based on the shifted left matrices and the construction is inspired by rational Krylov subspace methods for a two-term matrix equation. We were able to solve some extremely large LMTMEs using Multi-RB with our chosen left preconditioning strategy. However, we still encountered memory issues when using our chosen symmetric preconditioning strategy.

When applying symmetric preconditioning, we investigated two projection techniques based on Galerkin and Petrov-Galerkin constraints. When we imposed the Galerkin condition, the convergence history was observed to be very unstable and the inner iteration counts were very large at some iterations. However, imposing the Petrov-Galerkin condition resolved these issues because the resulting reduced problems are well-conditioned and the Frobenius norm of the preconditioned relative residual at every iteration is minimised. When we used the left preconditioning strategy, we imposed the Galerkin condition and the convergence history of the errors was observed to be much smoother than when using the symmetric preconditioning strategy. We concluded that Multi-RB with left preconditioning seems more successful than with the symmetric preconditioning strategy.

To conclude, we highlight some directions for future work. First, we would like to explore different strategies for choosing the parameters $s_{i, r}$ when constructing basis vectors for Multi-RB in Line 4 of Algorithm 1. A better choice might improve the convergence of the method. Second, based on observations for our groundwater flow test problems, it may be worthwhile investigating how the right-hand side of the matrix equation influences the accuracy and performance of Multi-RB. This could first be investigated numerically by considering more test problems with different source functions and boundary conditions. A sensible strategy would be to start with a simpler parameter-dependent PDE with only one solution variable such as the scalar elliptic PDE (2.53) before considering problems with saddle point structure such as the parameter-dependent linear elasticity problem and the parameter-dependent groundwater flow problem. Finally, it would be interesting to determine whether it is possible to apply the Multi-RB method to LMTMEs that arise when we apply SGMFEMs to more complicated PDEs/systems of PDEs arising in engineering applications, such as fluid flow problems, nonlinear elasticity problems, and problems with multiple sources of uncertainty.

## Bibliography

[1] Athanasios C. Antoulas. Approximation of Large-Scale Dynamical Systems. Advances in Design and Control. Society for Industrial and Applied Mathematics SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104, Philadelphia, PA, 2005.
[2] Ivo Babuška. The finite element method with Lagrangian multipliers. Numerische Mathematik, 20(3):179-192, 1973.
[3] Ivo Babuška and Manil Suri. On locking and robustness in the finite element method. SIAM Journal on Numerical Analysis, 29(5):1261-1293, 1992.
[4] Ivo Babuška, Raúl 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.
[5] Jonas Ballani and Lars Grasedyck. A projection method to solve linear systems in tensor format. Numerical Linear Algebra with Applications, 20(1):27-43, 2013.
[6] S Barnett and C Storey. Some applications of the Lyapunov matrix equation. IMA Journal of Applied Mathematics, 4(1):33-42, 1968.
[7] Richard H. Bartels and George W Stewart. Solution of the matrix equation $A X+X B=C[F 4]$. Communications of the ACM, 15(9):820-826, 1972.
[8] Peter Benner, Ren-Cang Li, and Ninoslav Truhar. On the ADI method for Sylvester equations. Journal of Computational and Applied Mathematics, 233(4):1035-1045, 2009.
[9] Peter Benner, Akwum Onwunta, and Martin Stoll. Low-rank solution of unsteady diffusion equations with stochastic coefficients. SIAM/ASA Journal on Uncertainty Quantification, 3(1):622-649, 2015.
[10] Daniele Boffi, Franco Brezzi, Michel Fortin, et al. Mixed Finite Element Methods and Applications, volume 44. Springer, 2013.
[11] Dietrich Braess. Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics. Cambridge University Press, 2007.
[12] Tobias Breiten, Valeria Simoncini, and Martin Stoll. Low-rank solvers for fractional differential equations. Electronic Transactions on Numerical Analysis, 45:107-132, 2016.
[13] Franco Brezzi. On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers. Publications Mathématiques et Informatique de Rennes, 8(S4):1-26, 1974.
[14] Franco Brezzi and Michel Fortin. Mixed and Hybrid Finite Element Methods, volume 15. Springer Science \& Business Media, 2012.
[15] Alexandra Bünger, Valeria Simoncini, and Martin Stoll. A low-rank matrix equation method for solving PDE-constrained optimization problems. SIAM Journal on Scientific Computing, 43(5):S637-S654, 2021.
[16] Alain Damlamian. Some remarks on Korn inequalities. Chinese Annals of Mathematics. Serie B, 39(2):335-344, 2018.
[17] Vladimir Druskin, Leonid Knizhnerman, and Valeria Simoncini. Analysis of the rational Krylov subspace and ADI methods for solving the Lyapunov equation. SIAM Journal on Numerical Analysis, 49(5):1875-1898, 2011.
[18] Vladimir Druskin, Chad Lieberman, and Mikhail Zaslavsky. On adaptive choice of shifts in rational Krylov subspace reduction of evolutionary problems. SIAM Journal on Scientific Computing, 32(5):2485-2496, 2010.
[19] Vladimir Druskin and Valeria Simoncini. Adaptive rational Krylov subspaces for large-scale dynamical systems. Systems $\mathcal{E}$ Control Letters, 60(8):546-560, 2011.
[20] Vladimir Druskin and Valeria Simoncini. Adaptive rational Krylov subspaces for large-scale dynamical systems. Systems and Control Letters, 60(8):546-560, 2011.
[21] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. Psychometrika, 1(3):211-218, 1936.
[22] Martin Eigel, Claude Jeffrey Gittelson, Christoph Schwab, and Elmar Zander. Adaptive stochastic Galerkin FEM. Computer Methods in Applied Mechanics and Engineering, 270:247-269, 2014.
[23] Nancy S Ellner and Eugene L Wachspress. New ADI model problem applications. In Proceedings of 1986 ACM Fall Joint Computer Conference, pages 528-534, 1986.
[24] Nancy S Ellner and Eugene L Wachspress. Alternating direction implicit iteration for systems with complex spectra. SIAM Journal on Numerical Analysis, 28(3):859-870, 1991.
[25] Howard C Elman, Darran G Furnival, and Catherine E Powell. H (div) preconditioning for a mixed finite element formulation of the diffusion problem with random data. Mathematics of Computation, pages 733-760, 2010.
[26] Howard C Elman, David J Silvester, and Andrew J Wathen. Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics. Oxford University Press, USA, 2014.
[27] Oliver G Ernst, Catherine Elizabeth 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.
[28] Oliver G Ernst and Elisabeth Ullmann. Stochastic Galerkin matrices. SIAM Journal on Matrix Analysis and Applications, 31(4):1848-1872, 2010.
[29] Roland W Freund and Noël M Nachtigal. QMR: a quasi-minimal residual method for non-hermitian linear systems. Numerische Mathematik, 60(1):315-339, 1991.
[30] 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.
[31] Vivette Girault and Pierre-Arnaud Raviart. Finite Element Methods for NavierStokes equations: Theory and Algorithms, volume 5. Springer Science \& Business Media, 2012.
[32] Gene Golub, Stephen Nash, and Charles Van Loan. A Hessenberg-Schur method for the problem $A X+X B=C$. IEEE Transactions on Automatic Control, 24(6):909-913, 1979.
[33] Gene H Golub and Charles F Van Loan. Matrix Computations. The Johns Hopkins University Press, 2013.
[34] Anne Greenbaum. Iterative Methods for Solving Linear Systems. SIAM, Philadelphia, PA, 1997.
[35] Leonard R Herrmann. Elasticity equations for incompressible and nearly incompressible materials by a variational theorem. AIAA Journal, 3(10):1896-1900, 1965.
[36] Magnus R Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving linear systems. Journal of Research of The National Bureau of Standards, 49(6):409, 1952.
[37] Roger A Horn and Charles R Johnson. Matrix Analysis. Cambridge University Press, 2012.
[38] Dan Y Hu and Lothar Reichel. Krylov-subspace methods for the Sylvester equation. Linear Algebra and its Applications, 172:283-313, 1992.
[39] Imad M Jaimoukha and Ebrahim M Kasenally. Krylov subspace methods for solving large Lyapunov equations. SIAM Journal on Numerical Analysis, 31(1):227251, 1994.
[40] Elias Jarlebring, Giampaolo Mele, Davide Palitta, and Emil Ringh. Krylov methods for low-rank commuting generalized Sylvester equations. Numerical Linear Algebra with Applications, 25(6):e2176, 2018.
[41] Khalide Jbilou. Low rank approximate solutions to large Sylvester matrix equations. Applied Mathematics and Computation, 177(1):365-376, 2006.
[42] Carsten Keller, Nicholas IM Gould, and Andrew J Wathen. Constraint preconditioning for indefinite linear systems. SIAM Journal on Matrix Analysis and Applications, 21(4):1300-1317, 2000.
[43] Arbaz Khan, Catherine E Powell, and David J Silvester. Robust a posteriori error estimators for mixed approximation of nearly incompressible elasticity. International Journal for Numerical Methods in Engineering, 119(1):18-37, 2019.
[44] 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):A402A421, 2019.
[45] Tamara G Kolda and Brett W Bader. Tensor decompositions and applications. SIAM Review, 51(3):455-500, 2009.
[46] Daniel Kressner and Petar Sirković. Truncated low-rank methods for solving general linear matrix equations. Numerical Linear Algebra with Applications, 22(3):564-583, 2015.
[47] Daniel Kressner and Christine Tobler. Low-rank tensor Krylov subspace methods for parametrized linear systems. SIAM Journal on Matrix Analysis and Applications, 32(4):1288-1316, 2011.
[48] Lev Davidovich Landau, Evgenij M Lifšic, Evegnii Mikhailovich Lifshitz, Arnold Markovich Kosevich, and Lev Petrovich Pitaevskii. Theory of Elasticity, volume 7. Butterworth-Heinemann, Oxford, 1986.
[49] Soren Larsen. Numerical analysis of elliptic partial differential equations with stochastic input data. PhD thesis, University of Maryland, College Park, 1986.
[50] V.I. Lebedev. On a Zolotarev problem in the method of alternating directions. USSR Computational Mathematics and Mathematical Physics, 17(2):58-76, 1977.
[51] Kookjin Lee and Howard C Elman. A preconditioned low-rank projection method with a rank-reduction scheme for stochastic partial differential equations. SIAM Journal on Scientific Computing, 39(5):S828-S850, 2017.
[52] Kookjin Lee, Howard C Elman, Catherine E Powell, and Dongeun Lee. Enhanced alternating energy minimization methods for stochastic Galerkin matrix equations. BIT Numerical Mathematics, pages 1-30, 2022.
[53] Kookjin Lee, Howard C Elman, and Bedrich Sousedik. A low-rank solver for the Navier-Stokes equations with uncertain viscosity. SIAM/ASA Journal on Uncertainty Quantification, 7(4):1275-1300, 2019.
[54] Yiding Lin and Valeria Simoncini. Minimal residual methods for large scale Lyapunov equations. Applied Numerical Mathematics, 72:52-71, 2013.
[55] Gabriel J Lord, Catherine E Powell, and Tony Shardlow. An Introduction to Computational Stochastic PDEs, volume 50 of Cambridge Texts in Applied Mathematics. Cambridge University Press, 2014.
[56] An Lu and Eugene L Wachspress. Solution of Lyapunov equations by alternating direction implicit iteration. Computers $\mathcal{E}$ Mathematics with Applications, 21(9):43-58, 1991.
[57] Malcolm F Murphy, Gene H Golub, and Andrew J Wathen. A note on preconditioning for indefinite linear systems. SIAM Journal on Scientific Computing, 21(6):1969-1972, 2000.
[58] Craig J Newsum. Efficient Reduced Basis Methods for Parameter-Dependent Saddle Point Problems. PhD thesis, The University of Manchester (United Kingdom), 2019.
[59] Christopher C Paige and Michael A Saunders. Solution of sparse indefinite systems of linear equations. SIAM Journal on Numerical Analysis, 12(4):617-629, 1975.
[60] Davide Palitta and Valeria Simoncini. Optimality properties of Galerkin and Petrov-Galerkin methods for linear matrix equations. Vietnam Journal of Mathematics, 48(4):791-807, 2020.
[61] Thilo Penzl. A cyclic low-rank Smith method for large sparse Lyapunov equations. SIAM Journal on Scientific Computing, 21(4):1401-1418, 1999.
[62] Catherine E Powell. Optimal preconditioning for mixed finite element formulation of second-order elliptic problems. PhD thesis, UMIST, 2003.
[63] 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.
[64] Catherine E Powell, David Silvester, and Valeria Simoncini. An efficient reduced basis solver for stochastic Galerkin matrix equations. SIAM Journal on Scientific Computing, 39(1):A141-A163, 2017.
[65] P. A. Raviart and J. M. Thomas. A mixed finite element method for 2-nd order elliptic problems. In Mathematical Aspects of Finite Element Methods, Lecture Notes in Mathematics, pages 292-315. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006.
[66] Axel Ruhe. Rational Krylov sequence methods for eigenvalue computation. Linear Algebra and its Applications, 58:391-405, 1984.
[67] Torgeir Rusten and Ragnar Winther. A preconditioned iterative method for saddlepoint problems. SIAM Journal on Matrix Analysis and Applications, 13(3):887-904, 1992.
[68] James Andrew John Rynn. On the Use of a Stochastic Galerkin Surrogate to Accelerate the Solution of PDE-Based Bayesian Inverse Problems. PhD thesis, The University of Manchester, 2020.
[69] Youcef Saad and X V Gv. Numerical solution of large Lyapunov equations. In in Signal Processing, Scattering and Operator Theory, and Numerical Methods, Proc. MTNS-89, pages 503-511. Birkhauser, 1990.
[70] 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.
[71] Yousef Saad. Iterative Methods for Sparse Linear Systems. SIAM, Philadelphia, PA, 2003.
[72] John Sabino. Solution of large-scale Lyapunov equations via the block modified smith methods. PhD thesis, Rice University, 2006.
[73] Jean Salençon. Handbook of Continuum Mechanics : General Concepts, Thermoelasticity. Springer, Berlin, 2001.
[74] Robert Scheichl. Iterative solution of saddle point problems using divergence-free finite elements with applications to groundwater flow. PhD thesis, University of Bath (United Kingdom), 2000.
[75] David Silvester, Alex Bespalov, and Catherine Powell. Stochastic incompressible flow and iterative solver software (S-IFISS) version 1.0, 2017. https://personalpages.manchester.ac.uk/staff/david.silvester/ ifiss/sifiss.html.
[76] David J Silvester and Valeria Simoncini. An optimal iterative solver for symmetric indefinite systems stemming from mixed approximation. ACM Transactions on Mathematical Software (TOMS), 37(4):1-22, 2011.
[77] Valeria Simoncini. A new iterative method for solving large-scale Lyapunov matrix equations. SIAM Journal on Scientific Computing, 29(3):1268-1288, 2007.
[78] Valeria Simoncini. Computational methods for linear matrix equations. SIAM Review, 58(3):377-441, 2016.
[79] RA Smith. Matrix equation $X A+B X=C$. SIAM Journal on Applied Mathematics, 16(1):198-201, 1968.
[80] Olga Taussky. Matrices $C$ with $C^{n} \rightarrow 0$. Journal of Algebra, 1(1):5-10, 1964.
[81] Ioana Teodora Vaduva. Novel iterative methods for solving matrix equations arising from partial differential equations. PhD thesis, University of Strathclyde, 2022.
[82] Charles F Van Loan. The ubiquitous Kronecker product. Journal of Computational and Applied Mathematics, 123(1-2):85-100, 2000.
[83] Eugene L Wachspress. Optimum alternating-direction-implicit iteration parameters for a model problem. Journal of the Society for Industrial and Applied Mathematics, 10(2):339-350, 1962.
[84] Eugene L Wachspress. Extended application of alternating direction implicit iteration model problem theory. Journal of the Society for Industrial and Applied Mathematics, 11(4):994-1016, 1963.
[85] Eugene L Wachspress. Iterative solution of the Lyapunov matrix equation. Applied Mathematics Letters, 1(1):87-90, 1988.
[86] Chunmei Wang, Junping Wang, Ruishu Wang, and Ran Zhang. A locking-free weak Galerkin finite element method for elasticity problems in the primal formulation. Journal of Computational and Applied Mathematics, 307:346-366, 2016.

