Optimal iterative solvers for linear systems with stochastic PDE origins: Balanced black-box stopping tests

Pranjal, Prasad

2017

MIMS EPrint: **2017.43**

Manchester Institute for Mathematical Sciences
School of Mathematics

The University of Manchester

Reports available from: http://eprints.maths.manchester.ac.uk/
And by contacting: The MIMS Secretary
School of Mathematics
The University of Manchester
Manchester, M13 9PL, UK

ISSN 1749-9097

Optimal iterative solvers for linear systems with stochastic PDE origins: Balanced black-box stopping tests

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

2017

Pranjal
School of Mathematics

Contents

List of Tables **6**

List of Figures **9**

List of Abbreviations **12**

List of Symbols **13**

Abstract **14**

Declaration **15**

Copyright Statement **16**

Dedication **17**

Acknowledgements **18**

A Researcher's Dilemma **19**

1 Introduction **20**

1.1 PDE models 20

1.2 Errors in numerical approximation 22

1.3 Finite element methods 23

1.4 Thesis objective 26

1.4.1 Problem statement 26

1.4.2 Solution methodology 26

1.4.3 Contribution to existing literature 27

1.5 Thesis organization 28

2 Balanced MINRES stopping for symmetric positive-definite systems 29

2.1 Parameter dependent PDEs . . . 31

2.2 Stochastic steady-state diffusion PDE . . . 33

2.3 An overview of MINRES . . . 35

2.3.1 MINRES strategy . . . 35

2.3.2 Convergence estimate for MINRES . . . 37

2.4 A fast iterative solver . . . 39

2.5 A balanced stopping test . . . 40

2.5.1 Error equation . . . 40

2.5.2 Tractable bounds on algebraic error . . . 41

2.5.3 Stopping criterion . . . 42

2.5.4 A posteriori error estimation . . . 42

2.5.5 Computational logistics . . . 45

2.6 Computational results . . . 45

2.6.1 Test problem 1 . . . 46

2.6.2 Test problem 2 . . . 50

2.7 Balanced stopping in CG . . . 52

2.8 Summary . . . 54

3 Balanced MINRES stopping for symmetric indefinite systems 55

3.1 Deterministic steady-state Stokes equations . . . 57

3.1.1 Weak formulation . . . 58

3.1.2 Mixed FEM formulation . . . 58

3.2 Block matrix form . . . 59

3.3 Block preconditioning . . . 60

3.4 A balanced stopping test . . . 61

3.4.1 Error equation . . . 62

3.4.2 Tractable bounds on algebraic error . . . 64

3.4.3 Stopping criterion . . . 65

3.4.4 A posteriori error estimation . . . 66

3.4.5 Computational logistics . . . 67

3.4.6 Cheap estimation of eigenvalues in stopping test . . . 67

3.4.7 Choice of stopping test 70
3.5 Computational results 72
3.5.1 Test Problem 1 72
3.5.2 Test Problem 2 77
3.6 Summary 79

4 Balanced iterative stopping for nonsymmetric systems I 81
4.1 Deterministic convection-diffusion equations 83
4.1.1 Weak formulation 84
4.1.2 Galerkin FEM formulation 85
4.1.3 Streamline diffusion FEM formulation 85
4.1.4 Matrix formulation 86
4.2 Fast Krylov solvers for nonsymmetric systems 88
4.2.1 An overview of GMRES 88
4.2.2 An overview of suboptimal Krylov solvers 91
4.3 A balanced stopping test 93
4.3.1 Error equation 93
4.3.2 Tractable bounds on algebraic error 95
4.3.3 Stopping criterion 96
4.3.4 A posteriori error estimation 97
4.3.5 Computational logistics 98
4.4 Computational results 98
4.5 Cheap eigenvalue estimation in stopping test 109
4.5.1 Solve the corresponding normal equations 109
4.5.2 Information from spectrum of F 110
4.5.3 Information from parameters of the problem 110
4.6 Summary 111

5 Balanced iterative stopping for nonsymmetric systems II 113
5.1 Deterministic Navier–Stokes equations 115
5.1.1 Weak formulation 115
5.1.2 Mixed FEM formulation 116
5.2 Nonlinear FEM iteration 116

5.2.1 Newton iteration . . . 117
5.2.2 Picard iteration . . . 117
5.2.3 Matrix formulation . . . 117
5.3 A balanced stopping test . . . 119
5.3.1 Error equation . . . 119
5.3.2 Tractable bounds on algebraic error . . . 120
5.3.3 Stopping criterion for linearized iteration . . . 121
5.3.4 Stopping criterion for nonlinear iteration . . . 121
5.3.5 A posteriori error estimation . . . 124
5.3.6 Computational logistics . . . 124
5.4 Computational results . . . 124
5.5 Summary . . . 128

6 Open questions **129**

Bibliography **131**

A Some definitions and theorems **141**
A.1 Linear algebra concepts . . . 141
A.2 Some special types of matrices . . . 143
A.3 Relevant theorems . . . 144

B Sample MATLAB runs of test problems in thesis **146**
B.1 Stochastic diffusion test problem 1 . . . 146
B.2 Stokes equations test problem 1 . . . 149
B.3 Convection-diffusion equations test problem . . . 153
B.4 Navier–Stokes equations test problem . . . 156

C CPUTIME comparisons of some test problems in chapter 2 **160**

D Eigenvalue behaviour of perturbed convection-diffusion operator **162**
D.1 Convection-diffusion eigenvalue problem . . . 162
D.2 Computational results . . . 164
D.3 Computational insights . . . 167

Word count 26885

List of Tables

2.1 Energy errors, a posteriori errors, and effectivity indices for diffusion test problem 1 with $m = 3$, $p = 2$, and $\sigma = 0.2$. 44

2.2 Energy errors, a posteriori errors, and effectivity indices for diffusion test problem 1 with $m = 3$, $h = 1/32$, and $\sigma = 0.2$. 44

2.3 Energy errors, a posteriori errors, and effectivity indices for diffusion test problem 1 with $m = 3$, $p = 2$, and $\sigma = 0.4$. 44

2.4 Energy errors, a posteriori errors, and effectivity indices for diffusion test problem 1 with $m = 3$, $h = 1/32$, and $\sigma = 0.4$. 45

2.5 Iteration counts and Rayleigh quotients estimates for diffusion test problem 1 with $\sigma = 0.3$, $m = 5$, and $p = 3$. 48

2.6 Iteration counts and Rayleigh quotients estimates for diffusion test problem 1 with $\sigma = 0.5$, $m = 5$, and $p = 3$. 48

2.7 Iteration counts and Rayleigh quotients estimates for diffusion test problem 1 with $\sigma = 0.5$, $m = 7$, and $p = 3$. 49

2.8 Iteration counts and Rayleigh quotients estimates for diffusion test problem 2 with slow decay, $m = 5$, and $p = 3$. 52

2.9 Iteration counts and Rayleigh quotients estimates for diffusion test problem 2 with fast decay, $m = 5$, and $p = 3$. 52

2.10 Comparison of iteration counts for preconditioned MINRES and CG for diffusion test problem 1 for $p = 3$. 53

2.11 Convergence of a posteriori approximation error at stopping point k^*_{CG} in preconditioned CG for diffusion test problem 1 for $p = 3$. 53

3.1 Actual approximation errors, a posteriori errors, and effectivity indices for $\boldsymbol{Q}_1$-$\boldsymbol{P}_0$ rectangular finite elements on uniform grids for Stokes test problem 1. 67

3.2 Comparison of literature and improved stopping tests for $\boldsymbol{Q}_2$-$\boldsymbol{P}_1$ finite elements on rectangular uniform grids for Stokes test problem 1. 69
3.3 MINRES iteration counts and errors along with extremal Ritz values and interior most harmonic Ritz values for block ideal preconditioning on uniform grids for Stokes test problem 1. 73
3.4 MINRES iteration counts and errors along with extremal Ritz values and interior most harmonic Ritz values for block AMG preconditioning on uniform grids for Stokes test problem 1. 73
3.5 MINRES iteration counts and errors along with extremal Ritz values and interior most harmonic Ritz values for block ideal preconditioning on $2^l \times (2^l \times 3)$ grids for Stokes test problem 2. 78
3.6 MINRES iteration counts and errors along with extremal Ritz values and interior most harmonic Ritz values for block AMG preconditioning on $2^l \times (2^l \times 3)$ grids for Stokes test problem 2. 78

4.1 Approximation errors, a posteriori errors, and effectivity indices for convection-diffusion test problem on uniform (left) and stretched (right) grids. 98
4.2 GMRES iteration counts & errors for DIAG (top) & ILU (bottom) preconditioning on uniform (left) & stretched (right) grids for discrete CD system. 102
4.3 GMRES iteration counts & errors for GMG (top) & AMG (bottom) preconditioning on uniform (left) & stretched (right) grids for discrete CD system. 103
4.4 BICGSTAB(2) iteration counts & errors for DIAG (top) & ILU (bottom) preconditioning on uniform (left) & stretched (right) grids for discrete CD system. 104
4.5 BICGSTAB(2) iteration counts & errors for GMG (top) & AMG (bottom) preconditioning on uniform (left) & stretched (right) grids for discrete CD system. 105
4.6 TFQMR iteration counts & errors for DIAG (top) & ILU (bottom) preconditioning on uniform (left) & stretched (right) grids for discrete CD system. 106

4.7 TFQMR iteration counts & errors for GMG (top) & AMG (bottom) preconditioning on uniform (left) & stretched (right) grids for discrete CD system. 107
4.8 Computed Θ from MATLAB `eigs` and its upper bound estimate for CD test problem on uniform grids for $\epsilon = 1/64$ (left) and $\epsilon = 1/200$ (right). 111

5.1 Navier–Stokes test problem solved using Newton iteration on a 16×64 grid with $\nu = 1/50$. 126
5.2 Navier–Stokes test problem solved using Newton iteration on a 32×96 grid with $\nu = 1/50$. 126
5.3 Navier–Stokes test problem solved using Newton iteration on a 64×192 grid with $\nu = 1/50$. 126
5.4 Navier–Stokes test problem solved using Newton iteration on a 16×64 grid with $\nu = 1/100$. 127
5.5 Navier–Stokes test problem solved using Newton iteration on a 32×96 grid with $\nu = 1/100$. 127
5.6 Navier–Stokes test problem solved using Newton iteration on a 64×192 grid with $\nu = 1/100$. 127

C.1 Iteration counts (cputimes in seconds) for diffusion test problem 1 with $\sigma = 0.3$, $m = 5$, and $p = 3$. 161
C.2 Iteration counts (cputimes in seconds) for diffusion test problem 1 with $\sigma = 0.5$, $m = 5$, and $p = 3$. 161
C.3 Iteration counts (cputimes in seconds) for diffusion test problem 2 with slow decay, $m = 5$, and $p = 3$. 161
C.4 Iteration counts (cputimes in seconds) for diffusion test problem 2 with fast decay, $m = 5$, and $p = 3$. 161

List of Figures

2.1 The SPD_MINRES algorithm expressed in pseudo-code. 43

2.2 Errors vs iteration number for preconditioned MINRES for diffusion test problem 1 with $h = 1/32$, $m = 5$, $p = 3$ | $\sigma = 0.3$ (left), $\sigma = 0.5$ (right). 47

2.3 Errors vs iteration number for preconditioned MINRES for diffusion test problem 1 with $m = 7$, $p = 3$, $\sigma = 0.5$ | $h = 1/16$ (left), $h = 1/32$ (right). 47

2.4 Computed Ritz values for diffusion test problem 1 with $m = 5$, $p = 3$ | $h = 1/16$ and $\sigma = 0.3$ (left), $h = 1/8$ and $\sigma = 0.5$ (right). 48

2.5 Errors vs iteration number for preconditioned MINRES for diffusion test problem 2 with $m = 5$, $p = 3$, $h = 1/32$ | slow decay (left), fast decay (right). 51

2.6 Computed Ritz values for diffusion test problem 2 with $m = 5$, $p = 3$, $h = 1/8$ | slow decay (left), fast decay (right). 51

3.1 The SADDLE_MINRES algorithm expressed in pseudo-code. 71

3.2 Errors vs iteration number for block ideal (left) and block AMG (right) preconditioned MINRES on a uniform grid $h = 1/128$ for Stokes test problem 1. 74

3.3 Computed Ritz values for block ideal (left) and block AMG (right) MINRES on a uniform grid $h = 1/128$ for Stokes test problem 1. 75

3.4 Computed harmonic Ritz values for block ideal (left) and block AMG (right) MINRES on a uniform grid $h = 1/128$ for Stokes test problem 1. 76

3.5 Computed discrete inf-sup constant for block ideal (left) and block AMG (right) preconditioned MINRES on a uniform grid $h = 1/128$ for Stokes test problem 1. 77

3.6 Errors vs iteration number for block ideal (left) and block AMG (right) preconditioned MINRES on a 128×384 grid for Stokes test problem 2. 79

4.1 FEM solution surface and contour plots from the MATLAB backslash solution on a uniform grid for $l = 7$. 99
4.2 Errors vs iteration number for GMRES with DIAG (top) and ILU (bottom) preconditioning on a uniform (left) and stretched (right) grid for $l = 7$. 102
4.3 Errors vs iteration number for GMRES with GMG (top) and AMG (bottom) preconditioning on a uniform (left) and stretched (right) grid for $l = 8$. 103
4.4 Errors vs iteration number for BICGSTAB(2) with DIAG (top) and ILU (bottom) preconditioning on a uniform (left) and stretched (right) grid for $l = 7$. 104
4.5 Errors vs iteration number for BICGSTAB(2) with GMG (top) and AMG (bottom) preconditioning on a uniform (left) and stretched (right) grid for $l = 8$. 105
4.6 Errors vs iteration number for TFQMR with DIAG (top) and ILU (bottom) preconditioning on a uniform (left) and stretched (right) grid for $l = 7$. 106
4.7 Errors vs iteration number for TFQMR with GMG (top) and AMG (bottom) preconditioning on a uniform (left) and stretched (right) grid for $l = 8$. 107
4.8 Computed and recurrence residuals 2-norm vs iteration number for GMG (left) and AMG (right) preconditioned BICGSTAB(2) on a uniform grid for $l = 8$. 108

5.1 The `NAVIER_NEWTON_GMRES` algorithm expressed in pseudo-code. 123
5.2 Errors vs iteration number for Navier–Stokes test problem on a 64×192 grid with $\nu = 1/100$ for Newton iteration (right) and linear (GMRES) iteration (left) at $l = 4$th Newton iteration. 128

D.1 Eigenvalue to shadow (left) and its perturbed values (right) for CD test problem on a 16×16 uniform grid. 164

D.2 Eigenvalue to shadow (left) and its perturbed values (right) for CD test problem on a 16×16 uniform grid. 165
D.3 Eigenvalue to shadow (left) and its perturbed values (right) for CD test problem on a 16×16 uniform grid. 165
D.4 Eigenvalue to shadow (left) and its perturbed values (right) for CD test problem on a 16×16 uniform grid. 166
D.5 Eigenvalue to shadow (left) and its perturbed values (right) for CD test problem on a 16×16 uniform grid. 166

List of Abbreviations

Term	Meaning
iff	If and Only If
PDE	Partial Differential Equation
FEM	Finite Element Method
UQ	Uncertainty Quantification
CG	Conjugate Gradient
MINRES	Minimal Residual
GMRES	Generalized Minimal Residual
BICGSTAB	Biconjugate Gradient Stabilized
TFQMR	Transpose Free Quasi-Minimal Residual
CD	Convection-Diffusion

List of Symbols

Symbol	Meaning
$\vert, :$	Such that
$:=$	Defined as
$\in$	Belongs to
$\forall$	For all
$\exists$	There exists
∞	Plus infinity
$\subset$	Subset
$\mathbb{R}$	Set of real numbers
$\mathbb{R}^d$	Real coordinate space of d dimensions
$\mathbb{R}^{m\times n}$	Set of real matrices of order $m \times n$
$\mathbb{C}$	Set of complex numbers
$\mathbb{C}^d$	Complex coordinate space of d dimensions
$\mathbb{C}^{m\times n}$	Set of complex matrices of order $m \times n$
L^p	Space of pth power integrable functions
$W^{m,p}$	Space of Sobolev functions of order m
$\vert \cdot \vert$	Modulus function
$\Vert \cdot \Vert$	Euclidean vector norm
$\Vert \cdot \Vert_2$	Euclidean (L^2) norm
$\Vert \cdot \Vert_{\ell_1}$	ℓ_1 norm
∇	Gradient operator
$\nabla\cdot$	Divergence operator
$\otimes$	Kronecker product
#dof	Degrees of freedom
span$\{\ldots\}$	Linear span of the elements in $\{\ldots\}$
$\exp(\cdot)$	Exponential of $(\cdot)$
$\mathbf{e}^{(k)}$	kth iteration error
$\mathbf{r}^{(k)}$	kth iteration residual
$\mathbf{e}_k$	kth vector of the cannonical basis
$[\cdot]^T$	Transpose of a vector/matrix
$[\cdot]^*$	Conjugate transpose of a vector/matrix
$\mathtt{diag}(Q)$	Diagonal matrix with diagonal elements of matrix Q

The University of Manchester

Pranjal
Doctor of Philosophy
Optimal iterative solvers for linear systems with stochastic PDE origins: Balanced black-box stopping tests
November 8, 2017

The central theme of this thesis is the design of *optimal balanced black-box* stopping criteria in iterative solvers of symmetric positive-definite, symmetric indefinite, and nonsymmetric linear systems arising from finite element approximation of stochastic (parametric) partial differential equations.

For a given stochastic and spatial approximation, it is known that iteratively solving the corresponding linear(ized) system(s) of equations to too tight algebraic error tolerance results in a wastage of computational resources without decreasing the usually unknown approximation error. In order to stop *optimally*—by avoiding unnecessary computations and premature stopping—algebraic error and a posteriori approximation error estimate must be *balanced* at the optimal stopping iteration. Efficient and reliable a posteriori error estimators do exist for close estimation of the approximation error in a finite element setting. But the algebraic error is generally unknown since the exact algebraic solution is not usually available. Obtaining tractable upper and lower bounds on the algebraic error in terms of a readily computable and monotonically decreasing quantity (if any) of the chosen iterative solver is the distinctive feature of the designed optimal balanced stopping strategy. Moreover, this work states the exact constants, that is, there are no user-defined parameters in the optimal balanced stopping tests. Hence, an iterative solver incorporating the optimal balanced stopping methodology that is presented here will be a *black-box* iterative solver. Typically, employing such a stopping methodology would lead to huge computational savings and in any case would definitely rule out premature stopping.

The constants in the devised optimal balanced black-box stopping tests in MINRES solver for solving symmetric positive-definite and symmetric indefinite linear systems can be estimated cheaply *on-the-fly*. The contribution of this thesis goes one step further for the nonsymmetric case in the sense that it not only provides an optimal balanced black-box stopping test in a memory-expensive Krylov solver like GMRES but it also presents an optimal balanced black-box stopping test in memory-inexpensive Krylov solvers such as BICGSTAB(ℓ), TFQMR etc. Currently, little convergence theory exists for the memory-inexpensive Krylov solvers and hence devising stopping criteria for them is an active field of research. Also, an optimal balanced black-box stopping criterion is proposed for nonlinear (Picard or Newton) iterative method that is used for solving the finite dimensional Navier–Stokes equations.

The optimal balanced black-box stopping methodology presented in this thesis can be generalized for any iterative solver of a linear(ized) system arising from numerical approximation of a partial differential equation. The only prerequisites for this purpose are the existence of a cheap and tight a posteriori error estimator for the approximation error along with cheap and tractable bounds on the algebraic error.

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

Dedication

I dedicate this thesis to the two most admirable beings of my existence: my parents Professor *Balgangadhar Prasad* and Mrs. *Manjula Prasad*. Their unwavering support, loving care, constant encouragement, unbiased and infallible faith in my abilities (that I do not believe myself to possess) have always given me immense confidence, keeping me grounded to aim and dream for the sky. But above all, it is their noble ideals and a myriad of selfless deeds of helping the helpless that continue to inspire and educate me at every instant, always motivating me to evolve into a better sentient being.

I would further dedicate this thesis to my sister *Prerna* and my entire family, especially those who are no more, and in particular to father figure uncles Mr. *Sanjivan Kumar Sinha*, Mr. *Shankar Prasad*, Mr. *Sunil Kumar Sinha*, Mr. *Chandrashekhar Prasad*, and Professor *Vinay Kumar Kanth*. Their vast knowledge ranging from philosophy to science combined with their evergreen simplicity, honesty, and humility will always be an inspiring and blessed benchmark to emulate.

This dedication also belongs rightfully to my very special family of friends who have been wonderful angels of patience, confidence, encouragement, support, care and who mean the universe to me (you know who you are, I need not mention your names!).

Acknowledgements

Like Welkin's poetic portrait on Nature's azured symphony cannot hum the complete melody of Universe's soulful expanse, a few words of acknowledgement cannot fathom my gratitude and respect towards the various well-wishers who have stood by me and encouraged me through every second of this exasperating yet exhilarating endeavour called *Doctor of Philosophy*. However, as per tradition, I would attempt to weave my words into heartfelt expressions.

A special note of gratitude and respect goes to my supervisor Professor *David J. Silvester* for believing in my academic talents and abilities about which I myself had and still have several doubts! Every academic meeting with him was a delight in the sense that at the end of our discussions, self-confidence in my mathematical abilities was always enhanced. His patient, polite, guiding, and encouraging nature never belittled my (at many times appalling) mathematical works. Most importantly, I am grateful to him for making me realize the importance of analytical, logical, and structured thinking. His emphasis on the minutest of details pertaining to scientific thinking, reading, speaking, writing (warning free LaTeX files in particular!), coding, and presentation have finally instilled in me the motivation to strive and develop a scholarly intellect.

I would also like to express my sincere gratitude to all my teachers at the University of Oxford, IIT Guwahati, and Patna University. In this context, I would especially like to thank Dr. *Catherine E. Powell*, Professor *Howard C. Elman*, Professor *Andrew J. Wathen*, Professor *Lloyd N. Trefethen*, Dr. *Kathryn Gillow*, Professor *Ian J. Sobey*, Professor *Anupam Saikia*, Professor *Natesan Srinivasan*, Dr. *Bikash Bhattacharjya*, and Dr. *K.V. Srikanth* for their constant encouragement of my academic interests.

A huge thank you is also due to my flatmates, friends—in particular the wonderful people in my research group—and the School of Mathematics colleagues, faculty, and support staff. All of them have made my Manchester experience very pleasant.

Last but most importantly, I would like to thank that divine energy called God.

A Researcher's Dilemma

To write
Or to let it ripe?
Thoughtless and thoughtful
Both surprise
Ignorance and Knowledge
Liberating to play
Dreams ponder to wonder
To curiously stroll away

A fact to expand
Or a truth to pen and understand
Deeper I seek
To unravel the in-finite
Searching on a mystic verse
I re-wander into Ze's realm divine
Sipping in momentarily peace
The trickling harmony of a truthful piece
Yet is the rationality real
or the sip surreal?
Ever a bemusing note
In Nature's tune
And the heart does sway
Between hope and dismay
Where truths continuously insist
But the dilemma too discretely persists:
To write
Or to let it ripe?

Pranjal

Chapter 1

Introduction

Our universe encompasses innumerable complicated phenomena throughout its being. Understanding its underlying structure, at least up to the level of human consciousness has been the goal from the dawn of human civilization. To this end, the diverse fields of physics, chemistry, biology, engineering, economics etc., have been developed. The tools and the ideas evolved therein are expressed concisely with the aid of the language of mathematics.

Mathematical models of many real-world phenomena are often formulated in the form of partial differential equations (PDEs) with initial/boundary conditions. Most of these arise in fluid flow and transport phenomena [Strauss, 2008, chapter 1]; a few relevant examples of which are given below.

- Heat conduction is modelled by the diffusion equations.
- Transfer and diffusion of materials is modelled by the convection-diffusion equations.
- Low velocity flows/confined flows are modelled by the Stokes equations.
- Flow of an incompressible fluid in general is modelled by the Navier–Stokes equations.

An elementary discussion about fluid flows can be found in [Acheson, 1990]. For a general introduction, motivation, and discussion about PDEs, one can refer to [Strauss, 2008].

1.1 PDE models

Realistic PDE models have the following characteristics.

- Uncertainty/randomness in parameters/coefficients.
- The solution space is infinite dimensional.
- Nonlinearity.

Thus, with appropriate initial/boundary conditions and a given (scalar) source term f, a scalar stochastic[1] PDE model can be represented generally in the following form. Find $u(\vec{x}, \mathbf{y}) : D \times \Gamma \to \mathbb{R}$ such that

$$\mathcal{L}(\vec{x}, \mathbf{y})u(\vec{x}, \mathbf{y}) = f(\vec{x}), \qquad \forall\, (\vec{x}, \mathbf{y}) \in D \times \Gamma, \tag{1.1}$$

where Γ, $D \subset \mathbb{R}^d\, (d = 1, 2, 3, \ldots)$ denote parameter and spatial domain respectively. Here the (nonlinear) PDE operator $\mathcal{L}$ and the solution u depend upon a finite number m of possibly random parameters $\mathbf{y} = [y_1, y_2, \ldots, y_m]^T \in \Gamma$.

Uncertainty in a model of a practical situation is inevitable. It may be *aleatory*, that is, inherent to the phenomenon being modelled or may be *epistemic*, arising due to a lack of knowledge about the modelled phenomenon. Uncertainty in the coefficients of a PDE model translates into randomness in the solution as well and hence a qualification of the uncertainty in a model is essential for obtaining a meaningful solution. This qualification is the central topic of study in the field of uncertainty quantification; see [Smith, 2014]. The nonlinearity and the infinite dimensionality of the solution space make it impossible to find analytical, closed form solutions. Hence, numerical methods like finite difference methods, finite volume methods, finite element methods (FEM) etc., are essential for solving PDE models in practice. A brief survey of the popular numerical methods that are used for solving PDEs can be found in [Sloan et al., 2001, preface, p. ix ff.]. Due to the widespread use and the ever-increasing popularity of FEM in industry and engineering applications, this thesis will focus entirely on FEM for solving the various PDEs encountered in later chapters. One can refer to [Brenner and Scott, 2008] for a detailed discussion on finite element methods.

Errors play an important role in numerical approximations. The next section gives a brief summary of the primary errors associated with numerical approximations.

[1]Such a PDE is called a random PDE [Smith, 2014, p. 97], but the terminology stochastic PDE is used in this thesis in accordance with its wide prevalence in the existing literature on this topic.

1.2 Errors in numerical approximation

Obtaining a *true solution* u to the PDE model (1.1) is fraught with difficulties. Apart from errors arising due to quantification of uncertainties (*stochastic discretization*), solving a PDE numerically results in various other types of errors. The first is the discretization or the *approximation error* which arises due to approximating the infinite dimensional solution u with a finite dimensional solution u_h. Note that h denotes the mesh parameter associated with discretizing (finite number of domain points) the *spatial domain* D. The second source of error arises in solving *iteratively* the discrete linear(ized) system(s) arising from the numerical approximation. This is known as the linear algebra error or the *algebraic error*. Another source of error albeit small are the *roundoff errors* that arise when the PDE (1.1) is solved numerically on a computer. However, roundoff errors can be neglected if *stable* algorithms are used for computing quantities on a computer. A detailed discussion of roundoff errors and stable algorithms can be found in [Higham, 2002]. Typically, errors play a vital role in assessing the accuracy of a numerical method employed to solve (1.1). Thus, errors need to be quantified, which is achieved through *norms*.

Definition 1.2.1 (Norm). [Brenner and Scott, 2008, p. 24]
Let V be a vector space over a field F. Norm is a function $\|\cdot\|_V : V \to \mathbb{R}$ that satisfies the following axioms.

$$
\begin{aligned}
&(i)\ \|v\|_V \geq 0, && \forall\, v \in V,\\
&(ii)\ \|v\|_V = 0 \iff v = 0, && \forall\, v \in V,\\
&(iii)\ \|cv\|_V = |c|\,\|v\|_V, && \forall\, c \in F,\ v \in V,\\
&(iv)\ \|v+w\|_V \leq \|v\|_V + \|w\|_V, && \forall\, v, w \in V, \quad \textbf{(the triangle inequality)}.
\end{aligned}
$$

Norms are innately related to vector spaces, inner products, and orthonormality of vectors; see Appendix A for their respective definitions. These concepts are central to finite element methods. Also, note that for any vector $\mathbf{x} = (x_1, x_2, \ldots, x_d)^T \in \mathbb{R}^d$, its Euclidean norm $\|\mathbf{x}\|_2 := (\sum_{i=1}^{d} x_i^2)^{\frac{1}{2}}$ will be denoted by $\|\mathbf{x}\|$ throughout this thesis. Two other concepts that will be employed throughout this thesis will be that of gradient and divergence.

Definition 1.2.2 (Gradient and Divergence). [Strauss, 2008, p. 178]
Suppose that $F = [F_1, F_2, F_3]^T$ and f are a vector valued and a scalar valued function

respectively of (x, y, z) in $\mathbb{R}^3$. Then the gradient (∇f) of f and the divergence $(\nabla \cdot F)$ of F are defined as follows.

$$\nabla f = \left[\frac{\partial f}{\partial x}, \frac{\partial f}{\partial y}, \frac{\partial f}{\partial z}\right]^T.$$
$$\nabla \cdot F = \frac{\partial F_1}{\partial x} + \frac{\partial F_2}{\partial y} + \frac{\partial F_3}{\partial z}.$$

A brief overview of finite element methods is presented next.

1.3 Finite element methods

Finite element methods compute the solution of a PDE in essentially two steps. Firstly, they relax the regularity restrictions on the original solution by formulating the PDE into a *weak form*. Secondly, the solution of the weak form is approximated in a finite dimensional subspace of the infinite dimensional solution space. Note that the finite dimensional approximation makes sense only if the weak form has a unique solution. This is verified using the Lax–Milgram theorem; see [Brenner and Scott, 2008, p. 62 ff.] for a detailed discussion.

In a nutshell, finite element methods compute $u_h \in U_h \subset U$ such that

$$\int_D \{f_h(\vec{x}) - \mathcal{L}_h(\vec{x}, \mathbf{y}) u_h(\vec{x}, \mathbf{y})\} v_h(\vec{x}, \mathbf{y}) = 0, \qquad \forall\, v_h \in V_h \subset V. \tag{1.2}$$

Here $\mathcal{L}_h$ and f_h are the FEM analogues of $\mathcal{L}$ and f in (1.1) respectively. The space U is the solution space of the weak form while the vector space V is called the space of test functions in the weak form. The spaces U_h and V_h are finite dimensional subspaces of U and V respectively. Some popular choices for U, V are the Sobolev spaces and the L^p spaces.

Definition 1.3.1 (L^p **spaces**)**.** [Oden and Demkowicz, 1996, p. 285]
For a given domain (Lebesgue-measurable and nonempty interior) $D \subset \mathbb{R}^d$

$$L^p(D) := \{f : D \to \mathbb{R} \text{ measurable} \,|\, \|f\|_p < \infty\},$$

where $\|f\|_p := \left(\int_D |f|^p\right)^{\frac{1}{p}}$ for $p \in [1, \infty)$.[2] The L^2 norm $\|\cdot\|_2$, which corresponds to the $L^2(D)$ space will be used frequently in this exposition.

[2]Note that the above definition also holds (here and in the subsequent definitions) for $p = \infty$. The only difference then is in the definition of the norm associated with $p = \infty$.

Definition 1.3.2 (Distributional derivatives). [Oden and Demkowicz, 1996, p. 434] Let $D \subset \mathbb{R}^d$ be an open set, $u \in L^p(D)$ be arbitrary and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_d)$ be a multiindex with $|\boldsymbol{\alpha}| = \sum_{i=1}^{d} \alpha_i$. A function $u^{\boldsymbol{\alpha}}$ defined on D is called distributional derivative of u, denoted by $D^{\boldsymbol{\alpha}}u$, iff

$$\int_D u D^{\boldsymbol{\alpha}}\phi \;=\; (-1)^{|\boldsymbol{\alpha}|} \int_D u^{\boldsymbol{\alpha}}\phi, \qquad \forall\, \phi \in C_0^\infty(D),$$

where $C_0^\infty(D) := \{f \in C^\infty(D) : \operatorname{supp} f \subset D, \operatorname{supp} f \text{ compact}\}$, $C^\infty(D)$ is the set of continuous functions defined on D whose derivatives of all orders exist and are continuous on D, and the support of f is defined as $\operatorname{supp} f := \overline{\{x \in D : f(x) \neq 0\}}$; see [Oden and Demkowicz, 1996, p. 393].

Definition 1.3.3 (Sobolev spaces). [Oden and Demkowicz, 1996, p. 436 ff.] Let $D \subset \mathbb{R}^d$ be an open set, m an integer, and $p \in [1, \infty]$. Then the Sobolev space of order m, denoted by $W^{m,p}$ is

$$W^{m,p}(D) := \{u \in L^p(D) : D^{\boldsymbol{\alpha}}u \in L^p(D), \quad \forall\, |\boldsymbol{\alpha}| \leq m\}.$$

It is a normed space with the norm $\|\cdot\|_{W^{m,p}(D)}$; for any $u \in W^{m,p}(D)$

$$\|u\|_{W^{m,p}(D)} := \left(\sum_{|\boldsymbol{\alpha}| \leq m} \|D^{\boldsymbol{\alpha}}u\|_p^p \right)^{\frac{1}{p}} \text{ for } p \in [1, \infty).$$

The Hilbert space

$$H^m(D) := W^{m,2} = \{u \in L^2(D) : D^{\boldsymbol{\alpha}}u \in L^2(D), \quad \forall\, |\boldsymbol{\alpha}| \leq m\},$$

is quite prevalent in FEM setting. In particular, the space $H^1(D)$ will be used frequently throughout this thesis.

Piecewise bilinear or piecewise linear basis functions that are defined (locally) on a grid composed of rectangles or triangles respectively in two dimensions (or on a grid composed of bricks or tetrahedra respectively in three dimensions) are some of the typical choices for constructing a finite element approximation; see [Brenner and Scott, 2008, chapter 3] for more details. Depending on whether the test space V_h is chosen to be essentially the same or different from the solution space U_h,[3] finite element methods can be classified as the Bubnov–Galerkin approximation or the Petrov–Galerkin approximation respectively.

[3] This statement is made more precise in later chapters.

A clever choice of FEM basis functions ensures that the FEM matrices are sparse. Direct methods for solving linear systems, which employ sparse elimination strategies based on Gaussian elimination do exist. These include reordering strategies, frontal methods etc. A survey of direct methods can be found in [Davies et al., 2016]. Although these direct methods are competitive with iterative methods (in terms of computational memory and time) for solving linear systems with a few thousand degrees of freedom, direct methods become increasingly infeasible for linear systems of higher dimensions. On the other hand, solving a linear system using an iterative method requires storage of only the nonzeros of the coefficient matrix. Thus, iterative methods take complete advantage of the sparsity of a matrix. Moreover, iterative methods specially tailored for specific classes of matrices have been studied in detail; for example see [Axelsson, 1994, chapter 5], [Greenbaum, 1997, chapter 2].

Finite element matrices are usually ill-conditioned with respect to discretization parameters. This implies slow convergence of a chosen iterative method. In order to accelerate convergence, preconditioning is required; see [Wathen, 2015]. Thus, iterative solution strategies like Krylov subspace methods (see [Liesen and Strakoš, 2012]) together with preconditioning can be quite effective for solving linear systems arising from FEM approximation of a PDE. The choice of an iterative solver for solving a linear system depends on the structure of the coefficient matrix. The coefficient matrices that are usually encountered in practice are described in [Saad, 2003, p. 4, 24]. Bubnov–Galerkin FEM approximation often results in a symmetric positive-definite linear system while Petrov–Galerkin FEM approximation always leads to a nonsymmetric linear system. Symmetric indefinite system of linear equations usually arise in mixed finite element approximations, that is, when the solution space is a Cartesian product of two or more approximation spaces; a more detailed discussion on this topic can be found in chapter 3.

Although the conjugate gradient (CG) method [Hestenes and Stiefel, 1952] is popular for solving symmetric positive-definite linear systems, the minimal residual (MINRES) method [Paige and Saunders, 1975] will be employed in this thesis to solve them (the reason for using MINRES instead of CG is explained in chapter 2, section 2.7). This algorithm will also be used for solving symmetric indefinite linear systems. For solving nonsymmetric linear systems, the generalized minimal residual (GMRES)

method [Saad and Schultz, 1986], the biconjugate gradient stabilized (BICGSTAB(ℓ)) method [Sleijpen and Fokkema, 1993], and the transpose free quasi-minimal residual (TFQMR) method [Freund, 1993] will be used.

The next section will provide a general summary of the problem statement, the solution methodology, and the research contribution of the material presented in this thesis.

1.4 Thesis objective

1.4.1 Problem statement

Numerical solution of a PDE with initial/boundary conditions essentially involves two types of errors—approximation error and algebraic error. The approximation error is fixed for chosen stochastic and spatial discretization parameters. Solving iteratively the corresponding discrete linear(ized) system(s) to a very high accuracy is not desirable. This is because a highly accurate iterative solution may require too many iterations and simply waste computational resources without decreasing the approximation error. On the other hand, if the iterations are stopped too early the iterative solution will not be a good approximation to the exact solution. This thesis attempts to handle these issues by presenting *optimal balanced black-box* stopping tests in Krylov solvers for solving linear systems with (stochastic) PDE origins.

1.4.2 Solution methodology

In order to stop *optimally*, that is, by avoiding premature stopping and unnecessary computations, it is important to use the fundamental relation between the algebraic error and the approximation error (for a given approximation): the total error at any iteration step is essentially the sum of the approximation error and the algebraic error; all the errors are measured in some *natural* norm (this issue is addressed in detail in later chapters). By balancing the algebraic error and the total error (which is the approximation error obtained from the solution computed at that iteration step), a *balanced* stopping test is obtained.

The approximation error can be measured a priori or/and a posteriori. A priori

approximation error estimation usually requires the solution to satisfy some regularity conditions which may not hold or/and may not be easily verifiable a priori. On the other hand, robust a posteriori approximation error estimation techniques are generally readily available. Moreover, a posteriori error estimation can be used for driving the FEM procedure adaptively. Hence, a posteriori approximation error estimation approach is used throughout this thesis.

Generally, the algebraic error is unknown since the exact algebraic solution is not usually available. Obtaining tractable upper and lower bounds on the algebraic error in terms of a readily computable and monotonically decreasing quantity (if any) of the chosen iterative solver is the novel feature of the devised stopping strategy. Moreover, there are no user-defined constants in the optimal balanced stopping tests presented in this thesis. Thus, iterative solvers incorporating such optimal balanced stopping strategies will be *black-box* solvers.

This thesis investigates the design of optimal balanced black-box stopping tests in iterative solvers for solving symmetric positive-definite, symmetric indefinite, and nonsymmetric linear systems arising from FEM approximation of (stochastic) PDEs. This is an active research field; see [Jiránek et al., 2010; Pietro et al., 2014a,b]. For the sake of brevity, the term balanced stopping test will usually be used in place of optimal balanced black-box stopping test throughout this thesis.

1.4.3 Contribution to existing literature

The contribution of this thesis to the existing literature lies in the fact that it states the exact constants, that is, there are no user-defined parameters in the balanced stopping tests. Hence, a solver based on the balanced stopping methodology will be a black-box solver. Moreover, for the symmetric positive-definite and symmetric indefinite linear systems, the constants in the stopping test for MINRES can be estimated cheaply *on-the-fly*. In the nonsymmetric case, the contribution of this thesis goes one step further in the sense that it presents a balanced stopping test for suboptimal Krylov solvers such as BICGSTAB(ℓ), TFQMR etc. Currently, little convergence theory exists for such solvers. The work on nonsymmetric systems and symmetric indefinite linear systems will soon be submitted for publication. The work on symmetric positive-definite linear systems has already been published [Silvester and Pranjal, 2016].

Parameterized partial differential equations are ubiquitous; see [Butler et al., 2012; Constantine et al., 2010]. For both symmetric and nonsymmetric linear systems arising from FEM discretization of a PDE, parametric PDEs have been considered here as the underlying PDEs in order to demonstrate the effectiveness of the balanced stopping tests. Stochastic Galerkin FEM [Babuška et al., 2004] and stochastic collocation methods [Babuška et al., 2007] are the popular choices for solving parametric PDEs. At the linear algebra level, these methods involve solving a single huge linear system and many smaller linear systems respectively. Since the existing storage requirements and computational flops increase with the size and the number of linear systems, a balanced stopping test will help to save significant computational work of an iterative solver. Note that the balanced stopping methodology remains applicable for solving the corresponding deterministic PDE too.

1.5 Thesis organization

This thesis has 7 chapters. Chapter 2 contains a balanced stopping test in MINRES for solving symmetric positive-definite linear systems arising from FEM approximation of the (stochastic) diffusion equations. In chapter 3, MINRES incorporating a balanced stopping criterion is used for solving symmetric indefinite linear systems that arise from mixed FEM approximation of the Stokes equations. Chapters 4 and 5 use a balanced stopping test in GMRES, BICGSTAB(ℓ), and TFQMR for solving nonsymmetric linear systems. The linear systems considered therein arise from FEM approximation of the convection-diffusion equations and the Navier–Stokes equations respectively. Chapter 6 contains open research questions arising from the material presented in this thesis. Research undertaken apart from designing balanced stopping tests is presented in chapter 7. Appendix A contains the basic definitions, concepts, and theorems that are used frequently in this thesis, while the appendix B contains sample runs of iterative solvers (MINRES, GMRES, BICGSTAB(2), TFQMR) with optimal balanced black-box stopping tests for some of the test problems presented in this thesis.

All computational results presented in this work have been produced using the software MATLAB and this thesis has been typeset in LaTeX 2_ε; see [Higham and Higham, 2017] for MATLAB fundamentals and [Griffiths and Higham, 2016] for a basic introduction to LaTeX 2_ε.

Chapter 2

Balanced MINRES stopping for symmetric positive-definite systems

Publication

- The material presented in this chapter is an expanded discussion based on the paper: David Silvester and Pranjal. An optimal solver for linear systems arising from stochastic FEM approximation of diffusion equations with random coefficients. SIAM/ASA J. Uncertainty Quantification, 4(1):298–311, 2016. `https://doi.org/10.1137/15M1017740`

- The devised balanced stopping test in MINRES solver for solving symmetric positive-definite linear systems arising from stochastic FEM approximation of parametric diffusion equations has resulted in the function `SPD_MINRES` in the toolbox S-IFISS [Silvester et al., 2015] in MATLAB. Note that this function was called `EST-MINRES` in [Silvester and Pranjal, 2016].

An optimal balanced black-box stopping test in preconditioned MINRES for solving symmetric positive-definite linear systems will be developed in this chapter. The underlying diffusion PDE with parametric coefficients will be discretized using the stochastic Galerkin finite element method, which is a combination of Galerkin FEM discretization in the spatial dimension and a discretization of the stochastic dimension. The corresponding linear system will be solved by a preconditioned MINRES solver with a balanced stopping test. This modified MINRES solver is an extension of the EST_MINRES solver of [Silvester and Simoncini, 2011], which has been developed for solving discrete saddle point problems. The algorithm has two main ingredients. First, a block preconditioner is used to ensure fast MINRES convergence independent of the problem parameters. Second, an optimal balanced black-box stopping test is incorporated to maximize efficiency. As mentioned in section 1.4, a balanced stopping test requires tractable bounds on the unobservable algebraic error and an a posteriori measure of the approximation error and the total errors. The a posteriori error estimator devised by [Bespalov et al., 2014] for the parametric diffusion problem will be employed here. Alternatively, a posteriori error estimation techniques such as residual based a posteriori error estimation strategies of [Eigel et al., 2014] can also be used, but these will not be considered here. Also, tractable bounds on the algebraic error will be obtained in terms of the norm of the iteration residual involving the preconditioner. This readily computable residual norm is monotonically decreasing with iteration index in preconditioned MINRES, which makes it an ideal candidate for estimating the usually unknown algebraic error.

This chapter is organized as follows. A discussion about parametric PDEs in general and parameterized diffusion PDE in particular is done in sections 2.1 and 2.2 respectively. An overview of MINRES is presented in section 2.3. The description of the block preconditioner that is used for accelerating MINRES convergence is presented in section 2.4. Section 2.5 has the balanced stopping test. Some computational results with discussions illustrating the effectiveness of the balanced stopping test in MINRES are presented in section 2.6. These results are compared with those obtained from the CG method with a balanced stopping test in section 2.7. The reason for using MINRES instead of CG is presented therein. A summary of the chapter is in section 2.8.

2.1 Parameter dependent PDEs

Popular numerical methods for solving a PDE dependent on a finite number (m) of parameters (say equation (1.1)) are: stochastic Galerkin FEM, Monte-Carlo methods, and stochastic collocation methods. Stochastic collocation methods have been studied in some detail over the past ten years; see [Babuška et al., 2007] and [Gordon, 2013]. These methods compute an approximation of the exact solution using interpolation techniques. It is difficult to implement such methods efficiently on a computer in general owing to the *curse of dimensionality* that is typically associated with them. Monte-Carlo methods are the most popular method among the practitioners in the field of uncertainty quantification. These involve sampling using large number of independent realizations of the parameter input. In case of a PDE this implies solving a large number of small deterministic linear systems. Although Monte-Carlo methods are robust and easily parallelizable, these become inefficient and infeasible especially for large-scale models where single deterministic linear system (and hence many linear systems) solve is computationally expensive. Instead of solving millions of small deterministic linear systems, an efficient alternative is to use a stochastic Galerkin finite element method which results in a single-coupled and huge deterministic linear system. Although this linear system is orders of magnitude larger than the subproblems of the Monte-Carlo methods, it is highly structured which can be utilized in the construction of fast and efficient solvers. Introduced by Ghanem (see [Ghanem and Spanos, 1991], [Ghanem and Kruger, 1996]) in the early 1990s, efficient and fast linear algebra for such systems have been studied extensively in the last two decades; for example see [Deb et al., 2001], [Babuška et al., 2004], and [Eiermann et al., 2007]. For a more detailed review of the research done in the field of stochastic finite elements, one can refer to [Gunzburger et al., 2014].

The stochastic Galerkin finite element method is a combination of discretization of the stochastic dimension coupled with a Galerkin finite element discretization of the spatial dimension. There are two popular strategies for the approximation of the stochastic dimension using global multivariate polynomials in the m parameters $y_1, y_2, \ldots, y_m$. The parameter approximation space $S_p = \text{span}\{\xi_j\}_{j=1}^{n_\xi}$ might be the span of one of the following two classes of global multivariate polynomials:

a) global multivariate polynomials of total degree $\leq p$.

b) global multivariate polynomials of total degree $\leq p$ in each of the m parameters.

Legendre polynomials are suitable candidates when uniform random variables are used for modelling the parameters, while Hermite polynomials are used when Gaussian random variables are employed. The importance of the particular choice of polynomial based on the chosen random variables for modelling the parameters will be discussed later. The dimension n_ξ of the space S_p is greater when it is spanned by global multivariate polynomials of total degree $\leq p$ in each of the m parameters rather when it is spanned by global multivariate polynomials of total degree $\leq p$. For example, suppose $m = 2$ and $p = 2$. If Legendre polynomials are used, then

- $S_p =: S_p^a := \text{span}\{1, y_1, y_2, y_1^2, y_2^2, y_1 y_2\}$.
- $S_p =: S_p^b := \text{span}\{y_1^2 y_2, y_1 y_2^2, y_1^2 y_2^2\} \cup S_p^a$.

For large m, the dimension of S_p^a grows algebraically as $\dfrac{(m+p)!}{m!\,p!}$ while that of S_p^b grows exponentially as $(p+1)^m$. Hence from the point of optimizing storage constraints in a computer's memory, global multivariate polynomials of total degree less than or equal to p for the parameter approximation space will be used. Each coefficient u_i of the solution u is then written as a linear combination of the polynomial basis functions

$$u_i = u_i^1 \xi_1 + u_i^2 \xi_2 + \ldots + u_i^{n_\xi} \xi_{n_\xi}, \tag{2.1}$$

and the system is projected (in a least-squares sense) to give the best approximation to the solution from the finite-dimensional subspace $S_p = \text{span}\{\xi_j\}_{j=1}^{n_\xi}$. This (Galerkin) projection leads to the linear system (see [Powell et al., 2017, p. A143] for derivation)

$$A_0 X G_0^T + \sigma \sum_{k=1}^{m} A_k X G_k^T = F, \tag{2.2}$$

where X is the $n \times n_\xi$ matrix of the unknown coefficients u_i^j and G_k is the weighted symmetric Gram matrix associated with the kth parameter. Here n is the dimension of the spatial approximation space, $\sigma > 0$ might represent the standard deviation of the parameter variation, and A_k are matrices associated with the spatial discretization.

Notice that writing $\mathbf{x} = \texttt{vec}(X)$ results in an equivalent high dimensional linear system of dimension $n \cdot n_\xi$ with a characteristic Kronecker product ($\otimes$) structure

$$\mathcal{A}\mathbf{x} = \mathbf{f} \iff \big(G_0 \otimes A_0 + \sigma \textstyle\sum_{k=1}^{m} G_k \otimes A_k\big)\mathbf{x} = \mathbf{f}. \tag{2.3}$$

In this chapter, $G_0 \otimes A_0$ will be a positive-definite matrix. The matrices $G_k \otimes A_k$ may be indefinite and $\mathcal{A}$ will be guaranteed to be invertible only when σ is *sufficiently small*. This issue will be addressed later in this chapter. More details on the structure of these matrices will be presented later.

2.2 Stochastic steady-state diffusion PDE

Stochastic diffusion process, which is used for modelling groundwater flow, conduction of heat in a medium etc., is a model that falls under the above framework. Suppose that the diffusion PDE is defined on a spatial domain $D \subset \mathbb{R}^d$ with an isotropic permeability tensor[1] $K = \kappa I$ where $\kappa : D \times \Gamma \to \mathbb{R}$ is a random field (that is parameterized by m independent and identically distributed random variables). Also, assume that κ can be written as

$$\kappa(\vec{x}, y_1, \ldots, y_m) := \mu(\vec{x}) + \sum_{k=1}^{m} \psi_k(\vec{x})\, y_k, \tag{2.4}$$

where $\mu(\vec{x})$ is the mean value of the permeability coefficient at $\vec{x} \in D$. Here $y_k \in \Gamma_k$ is the image of the kth random variable, $\Gamma := \Gamma_1 \times \cdots \times \Gamma_m$, $\Gamma_k \subset \mathbb{R}$, and $\{\psi_k\}_{k=1}^m$ are given functions defined on D. The expression (2.4) might be a Karhunen–Loève expansion [Lord et al., 2014, p. 201] or a polynomial chaos expansion [Wiener, 1938] of κ.

The stochastic steady-state diffusion problem requires solving for a random field $u(\vec{x}, \mathbf{y}) : D \times \Gamma \to \mathbb{R}$ that satisfies

$$-\nabla \cdot K(\vec{x}, \mathbf{y})\, \nabla u(\vec{x}, \mathbf{y}) = f(\vec{x}), \qquad \forall\, \vec{x} \in D \subset \mathbb{R}^d,\ (d = 2, 3),\ \mathbf{y} \in \Gamma, \tag{2.5a}$$

$$u(\vec{x}, \mathbf{y}) = g(\vec{x}), \qquad \forall\, \vec{x} \in \partial D_D,\ \mathbf{y} \in \Gamma, \tag{2.5b}$$

$$K(\vec{x}, \mathbf{y})\, \nabla u(\vec{x}, \mathbf{y}) \cdot \vec{n} = 0, \qquad \forall\, \vec{x} \in \partial D_N = \partial D \backslash \partial D_D,\ \mathbf{y} \in \Gamma, \tag{2.5c}$$

almost surely. Here ∂D_D, ∂D_N are the Dirichlet and the Neumann parts respectively of the boundary ∂D with an outward normal vector $\vec{n}$. The source term f and the boundary data g are given deterministic functions. The treatment of uncertainty in the source term f can be done easily; see [Deb et al., 2001] and [Elman et al., 2005].

[1] The isotropic permeability tensor is known by alternative names depending on the phenomenon being modelled by the diffusion PDE. It is known as the conductivity coefficient in a heat flow model, diffusion coefficient in a fluid flow model etc.

However, here it will be assumed to be a given deterministic function. Also, without any loss of generality zero Neumann boundary conditions are assumed.

In order to cast (2.5) in the weak form it is essential that $K(\vec{x}, \mathbf{y})$ is strictly positive and bounded, that is, there exists $k_1, k_2 \in \mathbb{R}$ such that

$$0 < k_1 \leq K(\vec{x}, \mathbf{y}) \leq k_2 < \infty, \qquad \text{almost everywhere in } D \times \Gamma,$$

so that the existence and uniqueness of the weak form solution is guaranteed in the separable solution space $W := \{u : \|u\|_E < \infty,\, u|_{\partial D_D \times \Gamma} = 0\} = H_0^1(D) \otimes L^2(\Gamma)$ by the Lax–Milgram lemma. The weak form of (2.5) is to find u such that $u - \hat{g} \in W$ satisfies

$$\left\langle \int_D K(\vec{x}, \mathbf{y})\, \nabla u(\vec{x}, \mathbf{y}) \cdot \nabla w(\vec{x}, \mathbf{y})\, \mathrm{d}\vec{x} \right\rangle = \left\langle \int_D f(\vec{x}) w(\vec{x}, \mathbf{y})\, \mathrm{d}\vec{x} \right\rangle, \quad \forall w \in W. \tag{2.6}$$

The function $\hat{g}$ is a smooth extension of g into the domain. The symbol $\langle \cdot \rangle$ denotes the expected value of a multivariate random variable that is defined on a probability space $(\Gamma, \mathcal{B}(\Gamma), \pi)$ with joint probability density function[2] $\rho(\mathbf{y})$ defined on the product set Γ. In terms of the joint probability density function $\rho(\mathbf{y})$, (2.6) can be rewritten as, find $u - \hat{g} \in W$ such that

$$\int_\Gamma \rho(\mathbf{y}) \int_D K(\vec{x}, \mathbf{y})\, \nabla u(\vec{x}, \mathbf{y}) \cdot \nabla w(\vec{x}, \mathbf{y})\, \mathrm{d}\vec{x}\, \mathrm{d}\mathbf{y} = \int_\Gamma \rho(\mathbf{y}) \int_D f(\vec{x}) w(\vec{x}, \mathbf{y})\, \mathrm{d}\vec{x}\, \mathrm{d}\mathbf{y}, \tag{2.7}$$

$\forall w \in W$. The natural *energy* norm $\|\cdot\|_E$ from (2.7) can be defined as

$$\|w\|_E^2 := \int_\Gamma \rho(\mathbf{y}) \int_D K(\vec{x}, \mathbf{y})\, |\nabla w(\vec{x}, \mathbf{y})|^2\, \mathrm{d}\vec{x}\, \mathrm{d}\mathbf{y}. \tag{2.8}$$

For a posteriori error estimate computations, another equally important norm $\|\cdot\|_{E_0}$ (called the mean energy norm henceforth) based on the mean field $\mu(\vec{x})$ of the permeability coefficient is required

$$\|w\|_{E_0}^2 := \int_\Gamma \rho(\mathbf{y}) \int_D \mu(\vec{x})\, |\nabla w(\vec{x}, \mathbf{y})|^2\, \mathrm{d}\vec{x}\, \mathrm{d}\mathbf{y}. \tag{2.9}$$

A more detailed discussion about these norms is done in the section on a posteriori error estimation. A crucial point here is that the two norms are equivalent whenever the formulation (2.6) is well-posed [Bespalov et al., 2014], that is, there exist positive constants λ and Λ such that

$$\lambda \|w\|_{E_0}^2 \leq \|w\|_E^2 \leq \Lambda \|w\|_{E_0}^2, \qquad \forall w \in W. \tag{2.10}$$

[2]Note that the conventional assumption of using identical and independent random variables makes the evaluation of $\rho(\mathbf{y})$ easier.

Galerkin finite dimensional approximation of (2.7) is associated with choosing a finite dimensional subspace $W_{h,p}$ of W. This is achieved by choosing subspaces of the component spaces, that is, $X_h \subset H_0^1(D)$, $S_p \subset L^2(\Gamma)$ and setting $W_{h,p} := X_h \otimes S_p$; see [Lord et al., 2014, section 9.5].

For the test problems generated using the S-IFISS toolbox [Silvester et al., 2015], the spatial domain D is two-dimensional and piecewise bilinear ($\boldsymbol{Q}_1$) or biquadratic ($\boldsymbol{Q}_2$) finite elements on a uniform rectangular grid are employed. This results in sparse *stiffness matrices* A_0 and A_k in (2.3). For the parameter approximation space S_p, choosing a basis set $\{\xi_j\}_{j=1}^{n_\xi}$ of global multivariate polynomials that is orthonormal with respect to the probability measure π will result in sparse *stochastic matrices* G_k ($G_0 = I$ and at most two nonzeros in any row otherwise). This orthogonality correspondence is the precise reason for choosing Legendre polynomials with uniform random variables and Hermite polynomials with Gaussian random variables; see [Gautschi, 2004] for a discussion on orthogonal polynomials.

The huge matrix $\mathcal{A}$ is never assembled in a practical implementation. Only the entries of G_k and the $(m+1)$ stiffness matrices each of size $n \times n$ are stored; see [Ghanem and Kruger, 1996]. The sparsity of the stiffness and the stochastic matrices implies that the matrix-vector products with the coefficient matrix $\mathcal{A}$ in (2.3) are cheap to compute—an essential ingredient for a computationally effective iterative solver. If (2.6) is well-posed, then $\mathcal{A}$ is a symmetric positive-definite matrix. However, it is ill-conditioned with respect to the discretization parameters. Thus, preconditioning is required with MINRES to solve the huge linear system (2.3) with coefficient matrix $\mathcal{A}$. An overview of MINRES is presented in the next section.

2.3 An overview of MINRES

2.3.1 MINRES strategy

Iteratively solving $\mathcal{A}\mathbf{x} = \mathbf{f}$ using MINRES [Elman et al., 2014a, chapter 4] involves constructing a sequence of iterates $\mathbf{x}^{(k)}$ ($k = 1, 2, \ldots$) from the shifted Krylov space

$$\mathbf{x}^{(0)} + \text{span}\,\{\mathbf{r}^{(0)}, \mathcal{A}\mathbf{r}^{(0)}, \ldots, \mathcal{A}^{k-1}\mathbf{r}^{(0)}\}, \tag{2.11}$$

where $\mathbf{x}^{(0)}$ is the initial solution vector, $\mathbf{r}^{(0)} = \mathbf{f} - \mathcal{A}\mathbf{x}^{(0)}$ is the initial residual and the spanning space $\mathrm{K}_k(\mathcal{A}, \mathbf{r}^{(0)}) := \text{span}\,\{\mathbf{r}^{(0)}, \mathcal{A}\mathbf{r}^{(0)}, \ldots, \mathcal{A}^{k-1}\mathbf{r}^{(0)}\}$ is the Krylov subspace

of order k generated by the matrix $\mathcal{A}$ and the vector $\mathbf{r}^{(0)}$. The residual $\mathbf{r}^{(k)}$ at the kth iterative step is

$$\begin{aligned}\mathbf{r}^{(k)} &= \mathbf{f} - \mathcal{A}\mathbf{x}^{(k)} \\ &= \mathbf{f} - \mathcal{A}(\mathbf{x}^{(0)} + \operatorname{span}\{\mathbf{r}^{(0)}, \mathcal{A}\mathbf{r}^{(0)}, \ldots, \mathcal{A}^{k-1}\mathbf{r}^{(0)}\}) \\ &= \mathbf{r}^{(0)} + \operatorname{span}\{\mathcal{A}\mathbf{r}^{(0)}, \mathcal{A}^2\mathbf{r}^{(0)}, \ldots, \mathcal{A}^k\mathbf{r}^{(0)}\}.\end{aligned} \tag{2.12}$$

The MINRES method chooses the iterate $\mathbf{x}^{(k)}$ from the space (2.11) such that it minimizes the Euclidean norm $\|\cdot\|$ of the corresponding residual $\mathbf{r}^{(k)}$ over the shifted space in the right-hand-side of (2.12).

A basis of orthonormal vectors $\{\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(k)}\}$ is constructed for the k-dimensional Krylov space, where $\mathbf{w}^{(1)} := \mathbf{f}/\|\mathbf{f}\|$. This construction process is known as the Lanczos method [Lanczos, 1950] where the basis vectors are generated iteratively using the recurrence

$$\mathcal{A}W_k = W_k T_k + t_{k+1,k}\, \mathbf{w}^{(k+1)}\mathbf{e}_k^T =: W_{k+1}\, \underline{T}_k, \tag{2.13}$$

where $W_k := [\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(k)}]$ and $\mathbf{e}_k$ is the kth vector of the canonical basis. The tridiagonal symmetric matrix T_k contains the orthogonalization coefficients and $\underline{T}_k$ is the tridiagonal matrix T_k with an additional final row $[0, \ldots, 0, t_{k+1,k}]$; for complete details see [Greenbaum, 1997, section 2.5]. The constant $t_{k+1,k}$ is chosen such that $\|\mathbf{w}^{(k+1)}\| = 1$. The Lanczos step (2.13) provides the following characterization of the iterate $\mathbf{x}^{(k)}$ and the residual $\mathbf{r}^{(k)}$

$$\mathbf{x}^{(k)} = \mathbf{x}^{(0)} + W_k \mathbf{y}^{(k)}, \tag{2.14a}$$

$$\mathbf{r}^{(k)} = \mathbf{f} - \mathcal{A}\mathbf{x}^{(k)} = W_{k+1}\left(\mathbf{e}_1 \|\mathbf{r}^{(0)}\| - \underline{T}_k \mathbf{y}^{(k)}\right). \tag{2.14b}$$

By solving the least squares problem $\min_{\mathbf{y}}(\mathbf{e}_1\|\mathbf{r}^{(0)}\| - \underline{T}_k\mathbf{y})$, the minimizing solution $\mathbf{x}^{(k)}$ is computed. Here $\mathbf{e}_1$ is the first canonical basis vector in $(k+1)$ dimensions. In order to solve the least squares problem, a QR factorization (see [Golub and Van Loan, 2013, p. 246]) of $\underline{T}_k$ is performed using k Givens rotations. The advantage of this approach is that only one new rotation is needed to update the QR factorization from the previous iteration; see [Fischer, 2011, p. 179].

The eigenvalues of $T_k = W_k^T \mathcal{A} W_k$ are known as the *Ritz values*; see [Golub and Van Loan, 2013, p. 551]. These can be computed cheaply and readily in the Lanczos

method at each iterative step of the MINRES solver. As the iteration progresses, the extremal Ritz values provide an increasingly better approximation to the corresponding extremal eigenvalues of $\mathcal{A}$ or of $\mathcal{M}^{-1}\mathcal{A}$ if the matrix is preconditioned with matrix $\mathcal{M}$. *This holds true for even small iteration index* k, and has been discussed extensively in [Parlett, 1998, chapter 13]. This aspect will be crucial for devising the balanced stopping criterion in section 2.4 since one may stop prematurely if the devised stopping test is applied before the relevant (here extremal) Ritz values have converged.

2.3.2 Convergence estimate for MINRES

In order to obtain a convergence estimate for MINRES [Elman et al., 2014a, chapter 4], (2.12) can be rewritten as

$$\mathbf{r}^{(k)} \;=\; p_k(\mathcal{A})\mathbf{r}^{(0)}, \tag{2.15}$$

where $p_k(\mathcal{A}) \in \Pi_k$—set of real polynomials of degree less than or equal to k—and by construction $p_k(0) = 1$. Since $\mathcal{A}$ is a symmetric matrix, therefore, there exists a basis of mutually orthogonal eigenvectors of $\mathcal{A}$ for the corresponding solution space (2.11). Let $\{\mathbf{w}_j\}_{j=1}^{n\cdot n_\xi}$ be an orthogonal basis (for the solution space) of eigenvectors of $\mathcal{A}$ with corresponding eigenvalues $\{\lambda_j\}$ and

$$\mathbf{r}^{(0)} \;=\; \sum_j \alpha_j\, \mathbf{w}_j, \quad \mathcal{A}\mathbf{w}_j \;=\; \lambda_j \mathbf{w}_j, \tag{2.16}$$

where α_j are scalars. Using (2.15)

$$\begin{aligned} \mathbf{r}^{(k)} \;&=\; p_k(\mathcal{A}) \sum_j \alpha_j\, \mathbf{w}_j \\ &=\; \sum_j \alpha_j\, p_k(\lambda_j)\, \mathbf{w}_j. \end{aligned}$$

In the last equality the following fact has been used, that is, if a linear operator $\mathcal{A}$ has eigenvalue λ, then any polynomial $p(\mathcal{A})$ has the same eigenvector with eigenvalue $p(\lambda)$. From the minimal residual criterion, it follows that

$$\begin{aligned} \|\mathbf{r}^{(k)}\| \;&=\; \min_{p_k \in \Pi_k,\, p_k(0)=1} \|\sum_j \alpha_j\, p_k(\lambda_j)\, \mathbf{w}_j\| \\ &\leq\; \min_{p_k \in \Pi_k,\, p_k(0)=1} \max_j |p_k(\lambda_j)|\; \|\mathbf{r}^{(0)}\|. \end{aligned} \tag{2.17}$$

The convergence estimate (2.17) for the unpreconditioned MINRES implies that it converges in exact arithmetic in a finite number of iterations $k \leq n \cdot n_\xi$.

Let $\mathcal{M}$ be a preconditioner for (2.3). Then the corresponding preconditioned linear system is

$$\mathcal{M}^{-1}\,\mathcal{A}\,\mathbf{x} = \mathcal{M}^{-1}\,\mathbf{f}. \tag{2.18}$$

In order to apply the MINRES method to the preconditioned system it is essential that the preconditioned coefficient matrix is symmetric. This is achieved by using a symmetric positive-definite preconditioner $\mathcal{M} = HH^T$. From (2.18) it follows that

$$(H^T)^{-1}\,H^{-1}\,\mathcal{A}\,\mathbf{x} \;= (H^T)^{-1}\,H^{-1}\,\mathbf{f} \iff H^{-1}\,\mathcal{A}\,\mathbf{x} \;= H^{-1}\,\mathbf{f},$$

which leads to the symmetric system

$$H^{-1}\,\mathcal{A}\,H^{-T}\mathbf{y} = H^{-1}\,\mathbf{f}, \qquad \mathbf{y} = H^T\mathbf{x}. \tag{2.19}$$

Any solution to (2.19) is also a solution to (2.18). The residual at iteration k is

$$H^{-1}\,(\mathbf{f} - \mathcal{A}\,\mathbf{x}^{(k)}) = H^{-1}\mathbf{r}^{(k)}.$$

If MINRES is applied to (2.19), then at iteration k the Euclidean norm $\|H^{-1}\mathbf{r}^{(k)}\|$ is minimized over

$$\begin{aligned} H^{-1}\,\mathbf{r}^{(0)} + \operatorname{span}\{&H^{-1}\,\mathcal{A}\,H^{-T}(H^{-1}\mathbf{r}^{(0)}),\,(H^{-1}\,\mathcal{A}\,H^{-T})^2(H^{-1}\mathbf{r}^{(0)}),\ldots,\\ &(H^{-1}\,\mathcal{A}\,H^{-T})^k(H^{-1}\mathbf{r}^{(0)})\}\\ = H^{-1}\,(\mathbf{r}^{(0)} + \operatorname{span}\{&\mathcal{A}\,\mathcal{M}^{-1}\mathbf{r}^{(0)},\,(\mathcal{A}\,\mathcal{M}^{-1})^2\mathbf{r}^{(0)},\ldots,\,(\mathcal{A}\,\mathcal{M}^{-1})^k\mathbf{r}^{(0)}\}). \end{aligned}$$

In fact, it follows that

$$\|H^{-1}\mathbf{r}^{(k)}\|^2 = (\mathbf{r}^{(k)})^T\,(H^T)^{-1}H^{-1}\,\mathbf{r}^{(k)} = \|\mathbf{r}^{(k)}\|^2_{\mathcal{M}^{-1}} := (\mathbf{r}^{(k)})^T\,\mathcal{M}^{-1}\,\mathbf{r}^{(k)}. \tag{2.20}$$

Thus, for preconditioned MINRES the convergence estimate analogous to (2.17) is

$$\frac{\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}}{\|\mathbf{r}^{(0)}\|_{\mathcal{M}^{-1}}} \;\le\; \min_{p_k\in\Pi_k,\,p_k(0)=1}\;\max_j|p_k(\lambda_j)|, \tag{2.21}$$

where λ_j are the eigenvalues of the matrix $H^{-1}\,\mathcal{A}\,H^{-T}$. However, because of the similarity transformation, $\mathcal{M}^{-1}\,\mathcal{A} = H^{-T}\,(H^{-1}\,\mathcal{A}\,H^{-T})\,H^T$ it follows that λ_j are also the eigenvalues of $\mathcal{M}^{-1}\,\mathcal{A}$. Since $H^{-1}\,\mathcal{A}\,H^{-T}$ is a symmetric matrix, therefore all the eigenvalues of $\mathcal{M}^{-1}\,\mathcal{A}$ are real. But $H^{-1}\,\mathcal{A}\,H^{-T}$ is also a positive-definite matrix if (2.6) is well-posed. Hence, all the eigenvalues of $\mathcal{M}^{-1}\,\mathcal{A}$ are positive. This aspect

is important since it ensures that the relevant (smallest and/or largest) eigenvalue involved in the balanced stopping test is (are) real and greater than zero.

In actual computations one needs to know only the action of $\mathcal{M}^{-1}$ on a vector and the matrix H is not needed. Also, the above analysis holds only when the matrix $\mathcal{M}^{-1}$ is positive-definite because only then $\|\cdot\|_{\mathcal{M}^{-1}}$ defines a norm. Since, the residual reduction for the preconditioned MINRES is in a norm based on the preconditioner (see (2.21)), the choice of preconditioner is crucial for accelerating the convergence of the MINRES method.

2.4 A fast iterative solver

From the structure of the matrix $\mathcal{A}$ in (2.3) it follows that when σ is small relative to $\|A_0\|$, the matrix $\mathcal{M} := I \otimes A_0$ will be a close approximation to $\mathcal{A}$. Also, since $I \otimes A_0$ is a block diagonal (with the symmetric stiffness matrix A_0 as the diagonal blocks) positive-definite matrix, it is cheaply and readily invertible. The action of the inverse of $I \otimes A_0$ can be computed through a single sparse factorization (of A_0) followed by n_ξ forward and backward substitutions. This preconditioning is known as the mean-based preconditioning and a detailed discussion of the spectral properties of it can be found in [Powell and Elman, 2009]. Besides accelerating convergence of MINRES, the preconditioner $I \otimes A_0$ is also spectrally equivalent to $\mathcal{A}$. Mathematically, this leads to Rayleigh quotient (see [Golub and Van Loan, 2013, p. 453]) bounds θ, Θ—the smallest and the largest eigenvalues of $\mathcal{M}^{-1}\mathcal{A}$ respectively—independent of the discretization parameters, that is

$$\theta \leq \frac{\mathbf{x}^T \mathcal{A}\, \mathbf{x}}{\mathbf{x}^T \mathcal{M} \mathbf{x}} \leq \Theta, \quad \forall\, \mathbf{x} \in \mathbb{R}^{n \cdot n_\xi}. \tag{2.22}$$

The expression (2.22) is equivalent to computing the extremal eigenvalues of the generalized eigenvalue problem for $\mathcal{A}$ and $\mathcal{M}$. The optimal convergence bound (2.21) can be weakened to hold over the finite interval $[\theta, \Theta]$. Thus, the following convergence estimate is obtained

$$\frac{\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}}{\|\mathbf{r}^{(0)}\|_{\mathcal{M}^{-1}}} \leq \min_{p_k \in \Pi_k,\, p_k(0)=1} \max_{z \in [\theta, \Theta]} |p_k(z)|. \tag{2.23}$$

The independence of θ, Θ from the discretization parameters implies that the number of iterations for the convergence of (preconditioned) MINRES is bounded independently of the discretization parameters.

2.5 A balanced stopping test

2.5.1 Error equation

For devising a balanced stopping criterion, it is stipulated on the premise that the algebraic error at a given iteration step cannot be worse than the approximation error at that step. To describe this in detail, suppose that $u_{hp} - \hat{g} \in W_{h,p}$ be the stochastic Galerkin finite element approximation to the true solution u. Then $\mathbf{x}$ (the true algebraic solution) is the coordinate vector of u_{hp} with respect to a chosen ordered basis. Let $\mathbf{x}^{(k)}$ be the coordinate vector at the iteration step k of the linear solver. Corresponding to this iterate, the approximation $u_{hp}^{(k)}$ can be formed. The Galerkin orthogonality [Elman et al., 2014a, p. 36] at iteration k decomposes the total error as the sum of the Galerkin approximation error and the algebraic error

$$\underbrace{\|u - u_{hp}^{(k)}\|_E^2}_{\text{total error}} = \underbrace{\|u - u_{hp}\|_E^2}_{\text{approximation error}} + \underbrace{\|u_{hp} - u_{hp}^{(k)}\|_E^2}_{\text{algebraic error}}, \qquad k = 0, 1, 2, \ldots. \tag{2.24}$$

where $\|u_{hp} - u_{hp}^{(k)}\|_E^2 = \|\mathbf{e}^{(k)}\|_{\mathcal{A}}^2 := (\mathbf{e}^{(k)})^T \mathcal{A}\, \mathbf{e}^{(k)}$. Here $\mathbf{e}^{(k)} := \mathbf{x} - \mathbf{x}^{(k)}$ denotes the kth iteration error. Also, note that the total error at iteration k is the approximation error at that iteration.

For obtaining a posteriori error estimate η to the approximation error $\|u - u_{hp}\|_E$, the mean-based local error estimation strategy of [Bespalov et al., 2014] is used. In this strategy, a posteriori error estimator is constructed in the mean-based $\|\cdot\|_{E_0}$ energy norm [Bespalov et al., 2014, Lemma 4.1][3]

$$\frac{1}{\sqrt{5}}\,\eta \;\le\; \|u - u_{hp}\|_{E_0} \;\le\; \frac{1}{\sqrt{1-\gamma^2}}\eta, \quad \gamma \in [0, 1). \tag{2.25}$$

Using (2.10), the equation (2.25) can be expressed as

$$c_1\,\eta \;\le\; \|u - u_{hp}\|_E \;\le\; C_1\eta, \quad \text{with } \frac{C_1}{c_1} \sim O(1). \tag{2.26}$$

Assuming that the a posteriori error estimators η and $\eta^{(k)}$ are close estimates of the approximation error and the total error (at the kth iteration step) respectively, then from (2.24) it follows that

$$(\eta^{(k)})^2 \;\simeq\; \eta^2 + \|\mathbf{e}^{(k)}\|_{\mathcal{A}}^2, \qquad k = 0, 1, 2, \ldots. \tag{2.27}$$

[3] A tighter a posteriori error bound has been derived in [Bespalov and Silvester, 2016]. But the computations were carried out more than a year before this new result was published. Moreover, the balanced stopping methodology remains applicable for the improved error bounds as well.

The relation $\simeq$ follows directly from (2.26). In practice for chosen (fixed) stochastic and spatial parameters, the approximation error (and hence η) is fixed but unknown. Thus, the iterative strategy can be looked upon as constructing a sequence $\{\eta^{(k)}\}$ which converges to η and hence the MINRES iteration should be stopped when the contribution of the algebraic error in (2.27) is insignificant, that is, stop at the first iteration k^* such that

$$\|\mathbf{e}^{(k^*)}\|_{\mathcal{A}} \leq \eta^{(k^*)}, \tag{2.28}$$

which implies that the total error cannot be reduced significantly any further and $\{\eta^{(k)}\}$ has converged with some accuracy to the approximation error estimate η.

2.5.2 Tractable bounds on algebraic error

Computing $\|\mathbf{e}^{(k)}\|_{\mathcal{A}}$ is anything but straightforward; see [Arioli, 2004]. The alternative is to obtain tractable bounds on $\|\mathbf{e}^{(k)}\|_{\mathcal{A}}$. At iteration k

$$\mathbf{r}^{(k)} = \mathbf{f} - \mathcal{A}\,\mathbf{x}^{(k)} \iff \mathbf{r}^{(k)} = \mathcal{A}\,\mathbf{e}^{(k)}, \tag{2.29}$$

Thus, from (2.29)

$$\|\mathbf{e}^{(k)}\|_{\mathcal{A}}^2 = (\mathbf{e}^{(k)})^T\,\mathcal{A}\,\mathbf{e}^{(k)} \iff \|\mathbf{e}^{(k)}\|_{\mathcal{A}}^2 = (\mathbf{r}^{(k)})^T\,\mathcal{A}^{-1}\,\mathbf{r}^{(k)} =: \|\mathbf{r}^{(k)}\|_{\mathcal{A}^{-1}}^2. \tag{2.30}$$

Equation (2.30) expresses the usually unknown energy error in terms of the iteration residual in the $\mathcal{A}^{-1}$ norm. Thus, in order to bound the algebraic error by the norm $\|\cdot\|_{\mathcal{M}^{-1}}$ of the iteration residual, computation of scalars δ and Δ is required such that

$$\delta \leq \frac{\mathbf{x}^T\mathcal{A}^{-1}\,\mathbf{x}}{\mathbf{x}^T\mathcal{M}^{-1}\,\mathbf{x}} \leq \Delta, \quad \forall\,\mathbf{x} \in \mathbb{R}^{n\cdot n_\xi}. \tag{2.31}$$

The expression (2.31) is equivalent to computing the extremal eigenvalues of the generalized eigenvalue problem for $\mathcal{A}^{-1}$ and $\mathcal{M}^{-1}$. Thus, the extremal generalized eigenvalue problem is to find $(\delta_*, \Delta_*) \in \mathbb{R}^2$ such that

$$\mathcal{A}^{-1}\mathbf{x_1} = \delta_*\mathcal{M}^{-1}\mathbf{x_1}, \qquad \mathcal{A}^{-1}\mathbf{x_2} = \Delta_*\mathcal{M}^{-1}\mathbf{x_2}, \tag{2.32}$$

where δ_*, Δ_* are the smallest and the largest eigenvalue respectively and $\mathbf{x_1}, \mathbf{x_2} \in \mathbb{R}^{n\cdot n_\xi}$ are the corresponding eigenvectors. Let $\mathcal{A}^{-1}\mathbf{x_1} = \mathbf{y_1}$ and $\mathcal{A}^{-1}\mathbf{x_2} = \mathbf{y_2}$. Then (2.32) can be rewritten as

$$\mathcal{A}\mathbf{y_1} = \frac{1}{\delta_*}\mathcal{M}\mathbf{y_1}, \qquad \mathcal{A}\mathbf{y_2} = \frac{1}{\Delta_*}\mathcal{M}\mathbf{y_2}, \tag{2.33}$$

Equation (2.33) shows that the extremal generalized eigenvalues of $\mathcal{A}^{-1}$ and $\mathcal{M}^{-1}$ are the reciprocal of the extremal generalized eigenvalues of $\mathcal{A}$ and $\mathcal{M}$ (that is the extremal eigenvalues of $\mathcal{M}^{-1}\mathcal{A}$). Thus, from (2.22) it follows that in (2.31), $\delta = \frac{1}{\Theta}$, $\Delta = \frac{1}{\theta}$, which together with (2.30) implies that for $k = 0, 1, \ldots$

$$\frac{1}{\Theta} \leq \frac{(\mathbf{r}^{(0)})^T \mathcal{A}^{-1}\,\mathbf{r}^{(0)}}{(\mathbf{r}^{(0)})^T \mathcal{M}^{-1}\,\mathbf{r}^{(0)}} = \frac{\|\mathbf{e}^{(0)}\|_{\mathcal{A}}^2}{\|\mathbf{r}^{(0)}\|_{\mathcal{M}^{-1}}^2}, \quad \frac{\|\mathbf{e}^{(k)}\|_{\mathcal{A}}^2}{\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}^2} = \frac{(\mathbf{r}^{(k)})^T \mathcal{A}^{-1}\,\mathbf{r}^{(k)}}{(\mathbf{r}^{(k)})^T \mathcal{M}^{-1}\,\mathbf{r}^{(k)}} \leq \frac{1}{\theta}. \tag{2.34}$$

Equation (2.34) leads to the following upper bounds on $\|\mathbf{e}^{(k)}\|_{\mathcal{A}}$, that is

$$\frac{\|\mathbf{e}^{(k)}\|_{\mathcal{A}}}{\|\mathbf{e}^{(0)}\|_{\mathcal{A}}} \leq \sqrt{\frac{\Theta}{\theta}}\,\frac{\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}}{\|\mathbf{r}^{(0)}\|_{\mathcal{M}^{-1}}} \Longleftrightarrow \|\mathbf{e}^{(k)}\|_{\mathcal{A}} \leq \sqrt{\frac{\Theta}{\theta}}\,\frac{\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}}{\|\mathbf{r}^{(0)}\|_{\mathcal{M}^{-1}}}\,\|\mathbf{e}^{(0)}\|_{\mathcal{A}}$$

$$\Longleftrightarrow \|\mathbf{e}^{(k)}\|_{\mathcal{A}} \leq \frac{\sqrt{\Theta}}{\theta}\,\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}, \tag{2.35a}$$

$$\|\mathbf{e}^{(k)}\|_{\mathcal{A}} \leq \frac{1}{\sqrt{\theta}}\,\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}. \tag{2.35b}$$

The tighter bound (2.35b) will be used in the presence of *tight* a posteriori error estimators (discussion in the next subsection) and the quantity $\frac{1}{\sqrt{\theta}}\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}$ will be called the *algebraic error bound* in the rest of the chapter. Note that since (2.22) is the finite dimensional analogue of (2.10), so $\lambda \leq \theta$ and $\Theta \leq \Lambda$. But a priori estimates of λ and Λ are pessimistic and/or difficult to find. So, θ is estimated *on-the-fly* as the smallest Ritz value in the Lanczos process of MINRES.[4]

2.5.3 Stopping criterion

In light of (2.35b) an optimal balanced black-box stopping criterion in MINRES is

$$\frac{1}{\sqrt{\theta}}\,\|\mathbf{r}^{(k^*)}\|_{\mathcal{M}^{-1}} \leq \eta^{(k^*)}. \tag{2.36}$$

Here k^* is the smallest value of iteration index k such that (2.36) is satisfied.

The resulting MINRES algorithm with the balanced stopping test (2.36) is called `SPD_MINRES` and is given in Figure 2.1. The external functions `matvecA`, `precM` compute the action of the matrices $\mathcal{A}$ and $\mathcal{M}^{-1}$ on a vector respectively. The function `error_est` computes the a posteriori error estimate.

2.5.4 A posteriori error estimation

The computation in S-IFISS of the a posteriori error estimate $\eta^{(k)}$ at the iteration step k requires the solution of two local (element level) problems having a block diagonal

[4] [Silvester and Simoncini, 2011] exploited the Lanczos connection for saddle point problems. Since estimates of interior eigenvalues were required, harmonic Ritz value estimates were computed there.

Algorithm: `SPD_MINRES`
given vectors $\mathbf{f}$, $\mathbf{x}^{(0)}$ and functions `matvecA`, `precM`, `param_est`, `error_est`

set $\mathbf{r}^{(0)} = \mathbf{f} - \texttt{matvecA}\,(\mathbf{x}^{(0)})$, $\hat{\mathbf{r}}^{(0)} = \texttt{precM}\,(\mathbf{r}^{(0)})$, $\rho_0 = \sqrt{(\mathbf{r}^{(0)})^T\hat{\mathbf{r}}^{(0)}}$
initialize basis vectors: $\mathbf{w} = \hat{\mathbf{r}}^{(0)}/\rho_0$, $\mathbf{p}^{(-1)} = \mathbf{0}$, $\mathbf{p}^{(0)} = \mathbf{r}^{(0)}/\rho_0$
initialize auxiliary vectors: $\mathbf{d}^{(-1)} = \mathbf{0}$, $\mathbf{d}^{(0)} = \mathbf{0}$
initialize projected right-hand side: $f = \rho_0$

`for` $k = 1, 2, \ldots$ `until` convergence `do`
  generate new basis and auxiliary vectors: $\mathbf{p}^{(k)} = \texttt{matvecA}\,(\mathbf{w})$, $\mathbf{d}^{(k)} = \mathbf{w}$
    `if` k>1, $t_{k-1,k} = t_{k,k-1}$, $\mathbf{p}^{(k)} = \mathbf{p}^{(k)} - \mathbf{p}^{(k-1)}t_{k-1,k}$
  $t_{k,k} = \mathbf{w}^T\mathbf{p}^{(k)}$, $\mathbf{p}^{(k)} = \mathbf{p}^{(k)} - \mathbf{p}^{(k-1)}t_{k,k}$
  compute preconditioned basis vector: $\mathbf{w} = \texttt{precM}\,(\mathbf{p}^{(k)})$
  $t_{k+1,k} = \sqrt{\mathbf{w}^T\mathbf{p}^{(k)}}$, $\mathbf{p}^{(k)} = \mathbf{p}^{(k)}/t_{k+1,k}$, $\mathbf{w} = \mathbf{w}/t_{k+1,k}$
  compute parameter for stopping test: `coef` = `param_est` (T_k)
  apply previous rotations:
    `if` k>2, $\rho_{1:2} = S_{k-2}t_{k-2:k-1,k}$, $\rho_{2:3} = S_{k-1}[\rho_2; t_{k,k}]$
    `elseif` k=2, $\rho_{2:3} = S_{k-1}t_{1:2,2}$
    `elseif` $k = 1$, $\rho_3 = t_{1,1}$
  compute new rotations:
    $\delta = \sqrt{\rho_3^2 + t_{k+1,k}^2}$, $c = |\rho_3|/\delta$, $s = \mathrm{sign}(\rho_3)t_{k+1,k}/\delta$
  apply new rotations: $\rho_3 = c\rho_3 + st_{k+1,k}$, $\hat{f} = -sf$, $f = cf$, $S_k = [c\ s; -s\ c]$
  update auxiliary vector: $\mathbf{d}^{(k)} = (\mathbf{d}^{(k)} - \mathbf{d}^{(k-1)}\rho_1 - \mathbf{d}^{(k-2)}\rho_2)/\rho_3$
  update solution: $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \mathbf{d}^{(k)}\hat{f}$
  compute discretization error estimate : $\eta^{(k)} = \texttt{error\_est}\,(\mathbf{x}^{(k)})$
  stopping test: `if coef`$\cdot|\hat{f}| \le \eta^{(k)}$, convergence
  update residual norm: $f = \hat{f}$
`enddo`

function `coef` = `param_est` (T_k)
  compute the smallest eigenvalue λ_1 of T_k
  set $\theta^{(k)} = \lambda_1$
    `if` $\theta^{(k)} \le 0$ output 'indefinite system' error message
    `else` set `coef` $= 1/\sqrt{\theta^{(k)}}$
`endfunction`

Figure 2.1: The `SPD_MINRES` algorithm expressed in pseudo-code.

(essentially with 5×5 stiffness matrices as the blocks) structure which differ overall only in the total number of blocks. It also requires the solution of one nonlocal sparse block diagonal system with each block representing the (already assembled) sparse stiffness matrix corresponding to the mean permeability coefficient. Thus, the computation of $\eta^{(k)}$ is relatively cheap and fast.

Computational results strongly suggest that the a posteriori error estimator that is employed here is a close estimate of the approximation error. To demonstrate this, some results are presented for the test problem 1; see section 2.6. The a posteriori error estimate η and the corresponding actual energy error $\|u_{ref} - u_{hp}\|_E$ using a reference solution are tabulated in Tables 2.1, 2.2, 2.3, and 2.4. Since the exact solution to the model problem is not available, a reference solution u_{ref} is computed on a fine spatial mesh $h = 2^{-7}$ with $m = 3$ random variables of total polynomial degree $p = 7$.

Table 2.1: Energy errors, a posteriori errors, and effectivity indices for diffusion test problem 1 with $m = 3$, $p = 2$, and $\sigma = 0.2$.

h	η	$\|u_{ref} - u_{hp}\|_E$	β_{eff}
1/4	1.8411e-2	1.8873e-2	0.98
1/8	8.7125e-3	9.4331e-3	0.92
1/16	4.3394e-3	4.7411e-3	0.92
1/32	2.3500e-3	2.4229e-3	0.97
1/64	1.5192e-3	1.3114e-3	1.16

Table 2.2: Energy errors, a posteriori errors, and effectivity indices for diffusion test problem 1 with $m = 3$, $h = 1/32$, and $\sigma = 0.2$.

p	η	$\|u_{ref} - u_{hp}\|_E$	β_{eff}
1	6.2228e-3	4.8648e-3	1.28
2	2.3500e-3	2.4229e-3	0.97
3	2.0770e-3	2.2840e-3	0.91
4	2.0650e-3	2.2781e-3	0.91
5	2.0645e-3	2.2778e-3	0.91

Table 2.3: Energy errors, a posteriori errors, and effectivity indices for diffusion test problem 1 with $m = 3$, $p = 2$, and $\sigma = 0.4$.

h	η	$\|u_{ref} - u_{hp}\|_E$	β_{eff}
1/4	2.2711e-2	2.2000e-2	1.03
1/8	1.4637e-2	1.4018e-2	1.04
1/16	1.2047e-2	1.1179e-2	1.08
1/32	1.1341e-2	1.0349e-2	1.09
1/64	1.1161e-2	1.01314e-2	1.10

This reference solution is then compared with the computed stochastic Galerkin FEM solution u_{hp}, which is linearly interpolated using MATLAB `interp2` function

Table 2.4: Energy errors, a posteriori errors, and effectivity indices for diffusion test problem 1 with $m = 3$, $h = 1/32$, and $\sigma = 0.4$.

p	η	$\|u_{ref} - u_{hp}\|_E$	β_{eff}
1	2.5555e-2	2.2042e-2	1.16
2	1.1341e-2	1.0349e-2	1.10
3	5.8487e-3	5.6422e-3	1.07
4	3.6488e-3	3.6656e-3	1.00
5	2.8008e-3	2.8459e-3	0.98

for compatible comparison with the reference solution. This was done for varying problem parameters. The remaining problem logistics are the same as in [Silvester and Pranjal, 2016]. The corresponding effectivity index $\beta_{\text{eff}} = \frac{\eta}{\|u_{ref} - u_{hp}\|_E}$ is also presented. From Tables 2.1, 2.2, 2.3, and 2.4, it follows that the effectivity index is very close to 1 thereby indicating that the a posteriori error estimate is a close estimate of the approximation error.

2.5.5 Computational logistics

The computational cost of the `SPD_MINRES` algorithm is similar to the computational cost of the MINRES algorithm apart from the computation of the quantities in the balanced stopping test (2.36). The $\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}$ norm of the iteration residual is readily available as a byproduct at each preconditioned MINRES step. As mentioned earlier, the smallest eigenvalue θ of $\mathcal{M}^{-1}\mathcal{A}$ can be estimated cheaply and readily by the smallest Ritz value $\theta^{(k)}$ at iteration k. In the case when the call to the a posteriori error estimator function is computationally expensive, one could compute $\eta^{(k)}$ periodically at every 4–5 iterations (say) to have a minor impact on the overall algorithmic cost.

2.6 Computational results

To provide a proof-of-concept, results of some computational experiments are presented here when the (preconditioned) MINRES stopping test (2.36) is applied to solve symmetric positive-definite linear systems arising from the stochastic Galerkin FEM approximation of (2.5).

2.6.1 Test problem 1

Following [Deb et al., 2001], the PDE (2.5) is defined on $D = (-1,1)\times(-1,1)$ with zero Dirichlet boundary conditions everywhere on the boundary and the source function $f(x_1,x_2) = \frac{1}{8}(2 - x_1^2 - x_2^2), \forall\,(x_1,x_2) \in D$. Rectangular $\boldsymbol{Q}_1$ (piecewise bilinear) finite elements are used on a uniform grid with mesh step size h. Uniform random variables defined on $\Gamma_k = [-1,1]$ are used for parameterizing the diffusion coefficient κ in (2.4). Multivariate Legendre polynomials of total polynomial degree $p = 3$ are used as basis for the parameter approximation space S_p and the mean field $\mu(\vec{x}) = 1, \forall\,\vec{x} \in D$ in the expansion (2.4). The spatial functions $\psi_k = \sqrt{3\lambda_k}\,\varphi_k$ in (2.4) are associated with eigenpairs $\{(\lambda_k,\varphi_k)\}_{k=1}^m$ of the covariance operator $C(\vec{x},\vec{x}') = \sigma^2 \exp\left(-\frac{1}{2}\|\vec{x}-\vec{x}'\|_{\ell_1}\right)$, $\forall\,\vec{x},\vec{x}' \in D$. The correlation length is set to two in each coordinate direction. This problem can be generated in S-IFISS by choosing example `2` when running the driver `stoch_diff_testproblem`.

Results are presented for different number of random variables m, different standard deviation σ, and various mesh parameter h in order to show the robustness of the balanced stopping test. A reference algebraic solution $\mathbf{x}$ can be computed in each case by solving the preconditioned discrete system with an absolute residual ($\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}$) reduction tolerance of `1e-14`. Corresponding to this reference solution $\mathbf{x}$, a reference a posteriori error estimate η can also be generated. The initial vector $\mathbf{x}^{(0)}$ for the solver is generated using MATLAB function `rand`. The same initial guess is used for a given discrete system to generate the reference solution and the algebraic solution based on the stopping test (2.36).

Representative results are presented in Figure 2.2 and Figure 2.3. Each subplot shows the evolution of $\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}$, a posterior error estimator $\eta^{(k)}$, and the algebraic error bound $\frac{1}{\sqrt{\theta}}\,\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}$ at each iteration step k; with θ estimated on-the-fly as the smallest Ritz value $\theta^{(k)}$ of the tridiagonal Lanczos matrix in the Lanczos process of preconditioned MINRES. It can be seen that the sequence $\{\eta^{(k)}\}$ converges to the reference a posteriori error η on each plot. Note that on each plot there are 9 more extra iterations after convergence to exhibit stopping at the correct place, that is, $\{\eta^{(k)}\}$ converges with some accuracy to η. Here $\eta^{(k)}$ is computed at each iterative step to illustrate the stopping methodology; in practice it should be computed periodically.

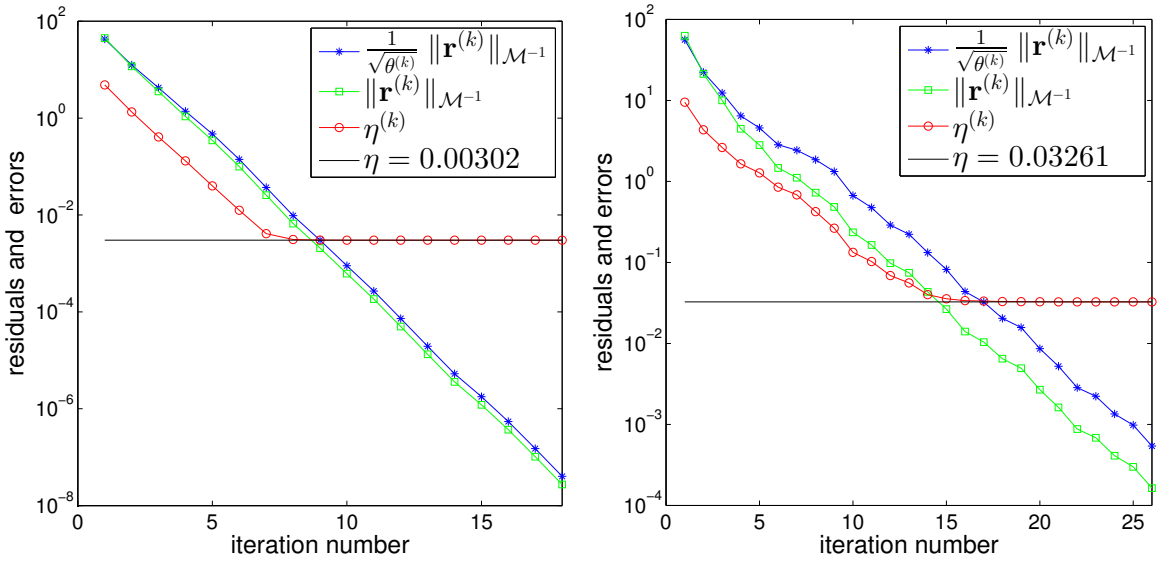

Figure 2.2: Errors vs iteration number for preconditioned MINRES for diffusion test problem 1 with $h = 1/32$, $m = 5$, $p = 3$ | $\sigma = 0.3$ (left), $\sigma = 0.5$ (right).

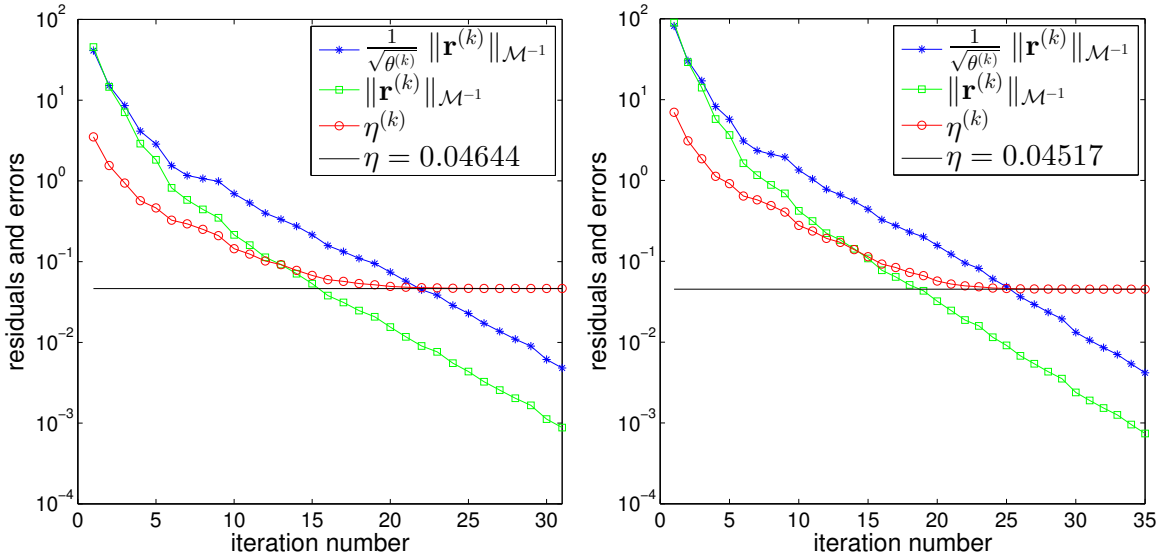

Figure 2.3: Errors vs iteration number for preconditioned MINRES for diffusion test problem 1 with $m = 7$, $p = 3$, $\sigma = 0.5$ | $h = 1/16$ (left), $h = 1/32$ (right).

Notice that the curves for $\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}$ and $\frac{1}{\sqrt{\theta^{(k)}}}\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}$ are not parallel initially but soon become parallel as $\theta^{(k)}$ converges to θ (which also eliminates the possibility of premature stopping). The plots in Figure 2.4 further confirm this convergence. In fact the computational experiments suggest no sign of the problematic *ghost* (spurious) Ritz values [Golub and Van Loan, 2013, p. 566] in any of the computations; for example see Figure 2.4. Note that on these plots, the actual extremal eigenvalues of the preconditioned matrix were computed using MATLAB `eigs`; this approach would be impractical in general owing to the huge size of the coefficient matrix.

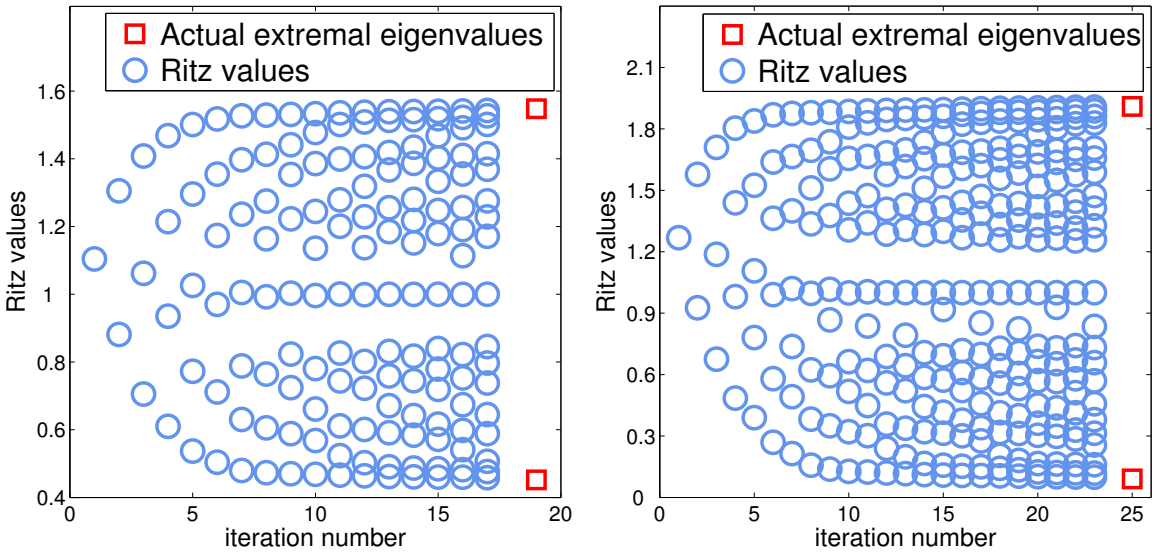

Figure 2.4: Computed Ritz values for diffusion test problem 1 with $m = 5$, $p = 3$ | $h = 1/16$ and $\sigma = 0.3$ (left), $h = 1/8$ and $\sigma = 0.5$ (right).

Table 2.5: Iteration counts and Rayleigh quotients estimates for diffusion test problem 1 with $\sigma = 0.3$, $m = 5$, and $p = 3$.

h	k_{tol1}	k_{tol2}	k^*	e_η^*	θ^*	Θ^*	#dof
1/4	14	19	6	7.2e-5	0.5276	1.5044	2744
1/8	14	20	7	2.1e-5	0.4833	1.5257	12600
1/16	15	20	8	1.5e-5	0.4734	1.5283	53816
1/32	16	21	9	7.2e-6	0.4708	1.5311	222264

Table 2.6: Iteration counts and Rayleigh quotients estimates for diffusion test problem 1 with $\sigma = 0.5$, $m = 5$, and $p = 3$.

h	k_{tol1}	k_{tol2}	k^*	e_η^*	θ^*	Θ^*	#dof
1/4	30	43	11	2.2e-3	0.1358	1.8789	2744
1/8	34	49	14	5.5e-5	0.1110	1.8941	12600
1/16	36	52	16	2.5e-4	0.1042	1.9032	53816
1/32	38	53	17	7.3e-4	0.1029	1.9045	222264

To show the effectiveness of the balanced stopping test for various parameters, the iteration counts k^* needed to satisfy the stopping test (2.36) have been compared in Tables 2.5, 2.6, and 2.7 with iteration counts $k_{\text{tol1}}, k_{\text{tol2}}$ needed to satisfy a fixed absolute residual $\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}$ reduction tolerance of `1e-6` and `1e-9` respectively. These tolerance values are a realistic user-input tolerance choices in the absence of a balanced stopping test (2.36). The user will not know in general the stopping point k^* a priori

Table 2.7: Iteration counts and Rayleigh quotients estimates for diffusion test problem 1 with $\sigma = 0.5$, $m = 7$, and $p = 3$.

h	k_{tol1}	k_{tol2}	k^*	e^*_η	θ^*	Θ^*	#dof
1/4	38	53	13	7.8e-4	0.0908	1.9315	5880
1/8	50	74	18	8.9e-4	0.0558	1.9590	27000
1/16	58	85	22	1.2e-3	0.0413	1.9649	115320
1/32	62	89	26	2.5e-4	0.0353	1.9678	476280

and is likely to provide a tighter tolerance than actually required. This would lead to wastage of computational work and time. The Tables 2.5, 2.6, and 2.7 indicate that for chosen (fixed) stochastic parameters the number of iterations for convergence remains bounded even as the spatial grid is refined. This is clearly indicated by the tabulated extremal Ritz value estimates (θ^*, Θ^*) corresponding to iteration k^*. This reconfirms that the mean-based preconditioner $\mathcal{M}$ is spectrally equivalent to $\mathcal{A}$. Also when σ is increased, θ becomes increasingly smaller (close to zero).[5] This slower convergence with increasing σ can be also gauged from Figure 2.3 too. The number of iterations for convergence based on (2.36) in the Table 2.5 is at least twice less as compared to those in Tables 2.6 and 2.7. In fact from heuristics it has been observed that for $\sigma > 0.5$, the problem is not well-posed.

Let $\eta^{(k^*)}$ be the corresponding a posteriori error estimate at the optimal stopping iteration k and $e^*_\eta := |\eta - \eta^{(k^*)}|$. The e^*_η columns in Tables 2.5, 2.6, and 2.7 show that $\{\eta^{(k)}\}$ has converged with an acceptable accuracy to the reference a posteriori error estimate η at the balanced stopping iteration k^*. A comparison of the corresponding entries for iteration counts $k_{\text{tol1}}, k_{\text{tol2}}$, and k^* shows that for the same approximation error, a significant number of iterations is saved by using the balanced stopping test. This would result in significant savings in computational work of the solver (as compared to using fixed absolute residual $\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}$ reduction tolerance of `1e-6` or tighter) if one were to solve the (preconditioned) linear systems arising from (spatial) adaptive finite element for the chosen stochastic parameters. The number of (internal nodes) degrees of freedom (#dof) of the resulting finite dimensional space which is equal to $\dfrac{(m+p)!}{m!\,p!}(2h^{-1}-1)^2$ is also tabulated. The savings in the computational work

[5] Sharp bounds $[1-\tau, 1+\tau]$ for the Rayleigh quotient (2.22) are established in [Powell and Elman, 2009, theorem 3.8], where the factor τ is the sum of the norms $\|\psi_k\|_\infty$ of the functions in (2.4). These bounds suggest that convergence will also be affected if m is increased with σ kept fixed.

of the iterative solver becomes further significant in light of the huge size of these linear systems.

2.6.2 Test problem 2

The second test problem is taken from [Eigel et al., 2014, section 11]. The PDE problem (2.5) is defined on a square domain $D = (0,1) \times (0,1)$ with zero Dirichlet boundary conditions and a constant source function $f = 1$. The diffusion coefficient κ in (2.4) is parameterized by uniform random variables that are defined on $\Gamma_k = [-1,1]$. The parameter approximation space S_p is spanned by complete polynomials of degree p. The mean field $\mu(\vec{x}) = 1, \forall\, \vec{x} \in D$ and the spatial functions ψ_k denote planar Fourier modes of increasing total order, so that

$$\psi_k(\vec{x}) := \alpha_k \cos(2\pi\beta_1(k)\, x_1) \cos(2\pi\beta_2(k)\, x_2), \quad \vec{x} = (x_1, x_2) \in D, \tag{2.37}$$

where for any $k \in \mathbb{N}$, $\beta_1(k) = k - \ell(k)(\ell(k)+1)/2$ and $\beta_2(k) = \ell(k) - \beta_1(k)$ with $\ell(k) = \lfloor -1/2 + \sqrt{1/4 + 2k} \rfloor$. The amplitude coefficients in (2.37) satisfy $\alpha_k = \bar{\alpha} k^{-\tilde{\sigma}}$ with some $\tilde{\sigma} > 1$ and $0 < \bar{\alpha} < 1/\zeta(\tilde{\sigma})$, where ζ is the Riemann zeta function. This example can be generated in S-IFISS by choosing example `5` when running the driver `stoch_diff_testproblem`. As in [Eigel et al., 2014], expansion (2.4) with slow ($\tilde{\sigma} = 2$) and fast ($\tilde{\sigma} = 4$) decay of the amplitudes α_k in (2.37) are considered. In each case, $\bar{\alpha}$ is chosen such that $\tau = \bar{\alpha}\zeta(\tilde{\sigma}) = 0.9$, which results in $\bar{\alpha} \approx 0.547$ for $\tilde{\sigma} = 2$ and $\bar{\alpha} \approx 0.832$ for $\tilde{\sigma} = 4$. Piecewise bilinear approximation on uniform rectangular grids is employed.

Representative results similar to test problem 1 are presented in Figure 2.5 for this test problem. The balanced MINRES solver takes more iterations (12 as opposed to 9) when the rate of decay of the coefficients is increased, all other parameters being kept constant. This occurs because the ratio Λ/λ is bigger in the case of the fast decay problem.

The convergence of Ritz values at balanced stopping iteration for this test problem for $h = 1/8$ is illustrated in Figure 2.6. Note that at the balanced stopping iteration k^* the extremal Ritz values have converged to the corresponding actual extremal eigenvalues. Thus, there is no danger of premature stopping here because of the nonconvergence of the Ritz estimate(s) at the balanced stopping iteration.

Results analogous to Tables 2.5, 2.6, and 2.7 are presented in Tables 2.8 and 2.9.

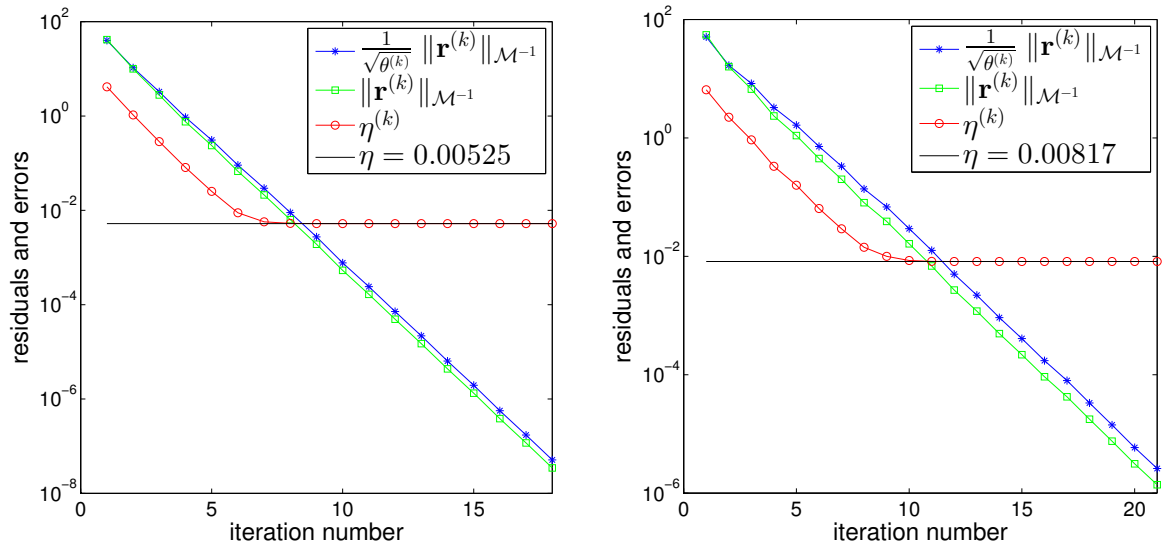

Figure 2.5: Errors vs iteration number for preconditioned MINRES for diffusion test problem 2 with $m = 5$, $p = 3$, $h = 1/32$ | slow decay (left), fast decay (right).

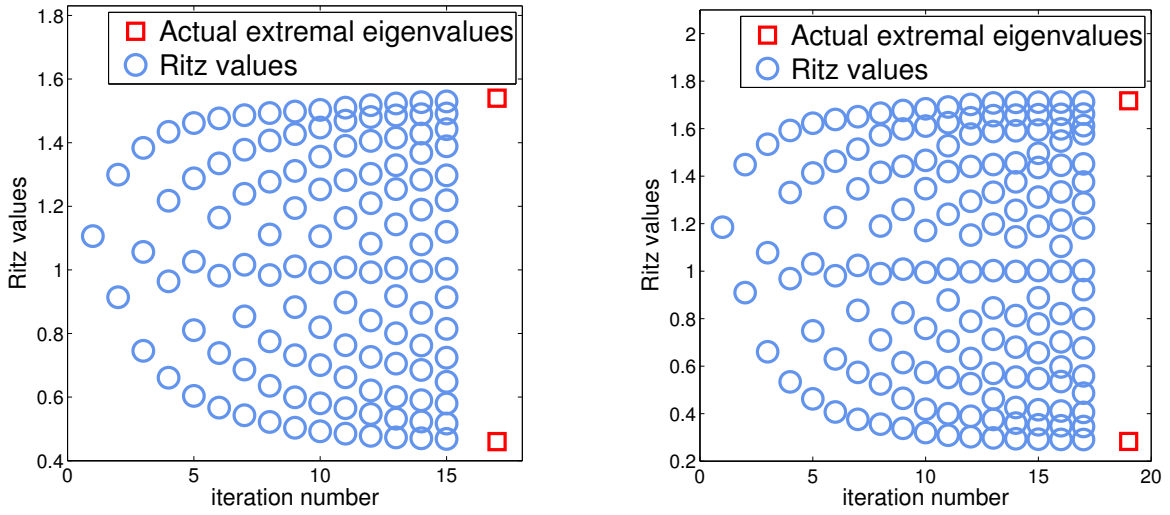

Figure 2.6: Computed Ritz values for diffusion test problem 2 with $m = 5$, $p = 3$, $h = 1/8$ | slow decay (left), fast decay (right).

It can be seen from these tables that the balanced stopping test is more efficient than using a fixed absolute residual $\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}$ reduction tolerance of `1e-6` or tighter. The tabulated extremal Ritz value estimates (θ^*, Θ^*) corresponding to iteration k^* once again reconfirm that the mean-based preconditioner is spectrally equivalent to $\mathcal{A}$. Further insights from these tables are also similar to those from test problem 1.

Table 2.8: Iteration counts and Rayleigh quotients estimates for diffusion test problem 2 with slow decay, $m = 5$, and $p = 3$.

h	k_{tol1}	k_{tol2}	k^*	e^*_η	θ^*	Θ^*	#dof
1/4	13	18	5	1.5e-5	0.6382	1.3905	2744
1/8	14	20	6	3.5e-5	0.5670	1.4762	12600
1/16	15	21	8	4.9e-6	0.5029	1.5215	53816
1/32	16	21	9	3.7e-6	0.4857	1.5320	222264

Table 2.9: Iteration counts and Rayleigh quotients estimates for diffusion test problem 2 with fast decay, $m = 5$, and $p = 3$.

h	k_{tol1}	k_{tol2}	k^*	e^*_η	θ^*	Θ^*	#dof
1/4	17	25	6	1.7e-4	0.4211	1.5932	2744
1/8	20	27	8	1.1e-4	0.3558	1.6649	12600
1/16	21	29	10	4.0e-5	0.3118	1.7064	53816
1/32	22	30	12	8.6e-6	0.2922	1.7157	222264

2.7 Balanced stopping in CG

A balanced stopping test in CG for solving symmetric positive-definite systems arising from the FEM approximation of diffusion equations is proposed in [Arioli, 2004]. However, this stopping test is based on a priori approximation error bounds, and hence it is difficult to employ it as a black-box solver. So, instead an optimal balanced black-box stopping test in preconditioned MINRES for solving symmetric positive-definite linear systems is devised here.

In the same paper, a method for computation of $\|\mathbf{e}^{(k)}\|_{\mathcal{A}}$ at any iteration step k using quantities available at iteration $k + d_e$ ($d_e \geq 1$ is the delay in iterations in computing $\|\mathbf{e}^{(k)}\|_{\mathcal{A}}$) of preconditioned CG has been proposed; this approach has been studied further in [Strakoš and Tichý, 2005]. The balanced stopping test

$$\|\mathbf{e}^{(k^*)}\|_{\mathcal{A}} \leq \eta^{(k^*)}, \tag{2.38}$$

based on a posteriori approximation error estimator can then be easily be utilized for preconditioned CG. But an optimal choice of the *delay* iteration count d_e for a generic problem is still an open question, which further makes CG with the balanced stopping (2.38) a non black-box solver.

For the sake of completeness, a comparison of the iteration counts of CG (with stopping test (2.38)) and MINRES (with stopping test (2.36)) for the test problem 1 is presented in Table 2.10.

Table 2.10: Comparison of iteration counts for preconditioned MINRES and CG for diffusion test problem 1 for $p = 3$.

h	$\sigma = 0.3$, $m = 3$		$\sigma = 0.5$, $m = 5$		$\sigma = 0.5$, $m = 7$	
	k^*_{MINRES}	k^*_{CG}	k^*_{MINRES}	k^*_{CG}	k^*_{MINRES}	k^*_{CG}
1/4	6	7	11	11	13	11
1/8	7	8	14	13	18	17
1/16	8	9	16	15	22	21
1/32	9	10	17	17	26	26

Table 2.11: Convergence of a posteriori approximation error at stopping point k^*_{CG} in preconditioned CG for diffusion test problem 1 for $p = 3$.

h	$\sigma = 0.3$, $m = 3$	$\sigma = 0.5$, $m = 5$	$\sigma = 0.5$, $m = 7$
	$e^*_{\eta_{\text{CG}}}$	$e^*_{\eta_{\text{CG}}}$	$e^*_{\eta_{\text{CG}}}$
1/4	2.2e-7	8.7e-4	2.1e-3
1/8	1.7e-6	6.2e-4	5.6e-4
1/16	1.3e-8	3.9e-4	8.1e-4
1/32	6.8e-7	4.1e-4	9.4e-4

It follows that iteration counts k^*_{CG} (with corresponding a posteriori error η^*_{CG}) of CG are similar to iteration counts k^*_{MINRES} of MINRES. A comparison of $e^*_{\eta_{\text{CG}}} := |\eta - \eta^*_{\text{CG}}|$ values in Table 2.11 with corresponding e^*_η values in Tables 2.5, 2.6, and 2.7 also indicate similar results. This is expected since CG is the ideal solver for solving a symmetric positive-definite linear system, so it should not perform any worse than MINRES. Note that iteration count k^*_{CG} takes into account the delay iteration count d_e in its counting. The minimum possible value of $d_e = 1$ is chosen here. However, as mentioned above, an optimal choice of d_e for a generic problem is an open research question. Thus, it will be better to employ preconditioned MINRES with an optimal balanced stopping test as a black-box solver for solving symmetric positive-definite linear systems arising from FEM approximation of PDEs.

2.8 Summary

A new algorithm for solving symmetric positive-definite linear systems arising from the (stochastic) Galerkin finite element approximation of PDEs with random coefficients has been devised. The PDE origins of these systems have to be taken into account when devising an optimal balanced black-box stopping test. In the presence of a 'good' preconditioner and an efficient a posteriori (energy-) error estimation routine, a balanced stopping criterion can be constructed. An on-the-fly cheap method for computing the constants involved in the stopping test has been also presented. Using this balanced stopping test in preconditioned MINRES for solving symmetric positive-definite linear systems will result in optimal use of computational resources. This aspect becomes further significant when solving a PDE adaptively using FEM.

Chapter 3

Balanced MINRES stopping for symmetric indefinite systems

Publication

- The material presented in this chapter will soon be submitted for publication.
- The devised balanced stopping test in MINRES for solving symmetric indefinite linear systems arising from FEM approximation of (parametric) Stokes equations has resulted in the function `SADDLE_MINRES` in the software IFISS [Elman et al., 2014b].

Large linear systems in saddle point form are ubiquitous. They frequently arise in optimization and in mixed finite element approximation of problems arising in fluid and solid mechanics. In matrix form such systems usually have a 2×2 block form

$$\begin{bmatrix} A & B^T \\ B & -C \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix}, \tag{3.1}$$

where $A \in \mathbb{R}^{n \times n}$ is symmetric positive-definite, $B \in \mathbb{R}^{m \times n}$, $C \in \mathbb{R}^{m \times m}$ is symmetric positive semi-definite, $\mathbf{u}, \mathbf{f} \in \mathbb{R}^n$ and $\mathbf{p}, \mathbf{g} \in \mathbb{R}^m$ with $n \geq m$. The coefficient matrix in (3.1) is always symmetric indefinite and so (preconditioned) MINRES is used for solving (3.1). An introduction about discrete saddle point systems and a detailed discussion on numerical methods for solving them can be found in [Benzi et al., 2005].

An optimal balanced black-box stopping test in preconditioned MINRES for solving (3.1) arising from mixed FEM approximation of PDEs has been devised by [Silvester and Simoncini, 2011]. In their analysis the matrix C is taken to be the zero matrix. An extension of their algorithm `EST_MINRES` henceforth called `SADDLE_MINRES` is presented in this chapter. The solver `SADDLE_MINRES` has essentially the same ingredients as the `EST_MINRES` solver: it employs a block preconditioner to accelerate MINRES convergence with a rate that is independent of problem parameters and incorporates a balanced stopping strategy to maximize efficiency. The balanced stopping test is obtained by balancing the a posteriori approximation error estimate with the iteration error in the *natural* norm associated with the underlying PDE. Similar to chapter 2 and [Silvester and Simoncini, 2011], tractable bounds on the usually unobservable (natural) norm of the iteration error are obtained in terms of the monotonically decreasing preconditioner norm of MINRES iteration residual.

Unlike `EST_MINRES`, the balanced stopping test in `SADDLE_MINRES` is catered for the coefficient matrix in (3.1) with a nonzero matrix C. Moreover, the constant in the balanced stopping test of [Silvester and Simoncini, 2011] has been 'improved' in this chapter. The improvement is in the sense that one now stops optimally a 'bit' earlier than using the balanced stopping test of [Silvester and Simoncini, 2011].

Balanced stopping criterion for symmetric indefinite linear systems arising from mixed FEM approximation of PDEs have been studied in detail by [Arioli and Loghin, 2008]. Their stopping criterion is based on a priori approximation error bounds and the constants involved in the balanced stopping test are also estimated a priori. This

is in contrast to the material presented here and in [Silvester and Simoncini, 2011] where the approximation error is estimated a posteriori and the constants involved in the balanced stopping test are estimated *on-the-fly*.

The stochastic Stokes equations is the underlying PDE considered here whose mixed FEM discretization gives rise to (3.1). 'Tight' a posteriori approximation error estimators for stochastic Stokes equations have not yet been developed. Since a posteriori error estimators play an important role in devising a balanced stopping test, stochastic collocation followed by FEM will be used to solve the Stokes equations. Thus, it is sufficient to focus on devising a balanced stopping test in MINRES for solving the symmetric indefinite linear system arising from mixed FEM approximation of deterministic Stokes equations.

This chapter has 6 sections. The weak form and the mixed FEM set up of the Stokes equations is done in section 3.1 and the target linear system is formulated in section 3.2. A discussion about block preconditioning for accelerating MINRES convergence in solving the target linear system is presented in section 3.3. The balanced stopping methodology is presented in section 3.4. Computational results that are produced using the IFISS toolbox are discussed in section 3.5. A summary of the chapter is presented in section 3.6.

3.1 Deterministic steady-state Stokes equations

Stokes equations are used for modelling flows at 'low speed'. Examples include highly viscous and confined flows such as flow of blood etc.; see [Elman et al., 2014a, p. 119]. Following the notation in [Elman et al., 2014a, p. 119], the steady-state Stokes solution $(\vec{u}, p)$ is defined on a spatial domain $D \subset \mathbb{R}^d$, $(d = 2, 3)$, where the vector valued function $\vec{u}(\vec{x}) : D \to \mathbb{R}^d$ and the scalar valued function $p(\vec{x}) : D \to \mathbb{R}$ satisfy

$$-\nabla \cdot \nabla \vec{u}(\vec{x}) + \nabla p(\vec{x}) = \vec{0}, \qquad \forall\, \vec{x} \in D, \tag{3.2a}$$

$$\nabla \cdot \vec{u}(\vec{x}) = 0, \qquad \forall\, \vec{x} \in D, \tag{3.2b}$$

$$\vec{u}(\vec{x}) = \vec{w}(\vec{x}), \qquad \forall\, \vec{x} \in \partial D_D, \tag{3.2c}$$

$$\nabla \vec{u}(\vec{x}) \cdot \vec{n} - \vec{n} p(\vec{x}) = \vec{s}(\vec{x}), \qquad \forall\, \vec{x} \in \partial D_N. \tag{3.2d}$$

Here ∂D_D and ∂D_N are the Dirichlet and Neumann parts respectively of the spatial boundary ∂D. The functions $\vec{w}, \vec{s}$ are given and $\vec{n}$ denotes the outward normal to ∂D.

3.1.1 Weak formulation

The weak formulation of (3.2) is to find $\vec{u} \in \mathbf{H}^1_E(D)$ and $p \in L^2(D)$ such that

$$\begin{aligned} a(\vec{u},\vec{v}) + b(\vec{v},p) &= f(\vec{v}), \qquad \forall\, \vec{v} \in \mathbf{H}^1_{E_0}(D), \\ b(\vec{u},q) &= 0, \qquad \forall\, q \in L^2(D), \end{aligned} \tag{3.3}$$

where

$$\begin{aligned} a(\vec{u},\vec{v}) &:= \int_D \nabla\vec{u} : \nabla\vec{v} - \int_D p\,(\nabla\cdot\vec{v}), \\ &\nabla\vec{u} : \nabla\vec{v} \text{ denotes componentwise dot product,} \\ b(\vec{u},q) &:= \int_D q\,(\nabla\cdot\vec{u}), \qquad f(\vec{v}) := \int_{\partial D_N} \vec{s}\cdot\vec{v}, \\ \mathbf{H}^1_E(D) &:= \{\vec{v} \in H^1(D)^d \,|\, \vec{v} = \vec{w} \text{ on } \partial D_D\}, \\ \mathbf{H}^1_{E_0}(D) &:= \{\vec{v} \in H^1(D)^d \,|\, \vec{v} = \vec{0} \text{ on } \partial D_D\}. \end{aligned}$$

Here $H^1(D)^d$ is the d-fold Cartesian product of the $H^1(D)$ space.

3.1.2 Mixed FEM formulation

Choosing finite dimensional subspaces $\mathbf{X}^h_E \subset \mathbf{H}^1_E(D)$, $\mathbf{X}^h_{E_0} \subset \mathbf{H}^1_{E_0}(D)$, $M^h \subset L^2(D)$ leads to a mixed FEM formulation from (3.3); find $\vec{u}_h \in \mathbf{X}^h_E$, $p_h \in M^h$ such that

$$\begin{aligned} a(\vec{u}_h,\vec{v}_h) + b(\vec{v}_h,p_h) &= f(\vec{v}_h), \qquad \forall\, \vec{v}_h \in \mathbf{X}^h_{E_0}, \\ b(\vec{u}_h,q_h) &= 0, \qquad \forall\, q_h \in M^h. \end{aligned} \tag{3.4}$$

Let $\{\vec{\phi}_j\}_{j=1}^{n_u}$ be a basis for the finite dimensional space $\mathbf{X}^h_{E_0}$. It can be extended (loosely speaking)[1] to form a basis $\{\vec{\phi}_j\}_{j=1}^{n_u+n_\partial}$ for $\mathbf{X}^h_E$, so that any $\vec{u}_h \in \mathbf{X}^h_E$ can be written as

$$\vec{u}_h = \sum_{j=1}^{n_u+n_\partial} u_j\vec{\phi}_j, \qquad u_j \in \mathbb{R}, \tag{3.5}$$

where the known term $\sum_{j=n_u+1}^{n_u+n_\partial} u_j\vec{\phi}_j$ interpolates the boundary data on ∂D_D.

Similarly, if $\{\psi_k\}_{k=1}^{n_p}$ be a basis for M^h, then any $p_h \in M^h$ has an expansion

$$p_h = \sum_{k=1}^{n_p} p_k\psi_k, \qquad p_k \in \mathbb{R}. \tag{3.6}$$

[1] The space $\mathbf{X}^1_E$ is not a vector space unless $\vec{w} = \vec{0}$.

3.2 Block matrix form

Plugging the basis expansions from equations (3.5) and (3.6) in (3.4) results in the following block matrix formulation[2]

$$\begin{bmatrix} \mathbf{A} & B^T \\ B & O \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix}. \tag{3.7}$$

The symmetric positive-definite matrix $\mathbf{A}$ (henceforth called *vector-Laplacian matrix*) is a block diagonal matrix with the usual FEM stiffness matrix on its diagonals and the matrix B is called the *divergence matrix*. Solution vectors $\mathbf{u} = [u_1, \ldots, u_{n_u}]^T \in \mathbb{R}^{n_u}$, $\mathbf{p} = [p_1, \ldots, p_{n_p}]^T \in \mathbb{R}^{n_p}$, and the entries of $\mathbf{A}$, B, $\mathbf{f}$, and $\mathbf{g}$ are given by [Elman et al., 2014a, p. 130]

$$\begin{aligned}
\mathbf{A} &= [a_{ij}] \in \mathbb{R}^{n_u \times n_u}, & a_{ij} &:= \int_D \nabla\vec{\phi}_i : \nabla\vec{\phi}_j, \\
B &= [b_{kj}] \in \mathbb{R}^{n_p \times n_u}, & b_{kj} &:= -\int_D \psi_k \, (\nabla \cdot \vec{\phi}_j), \\
\mathbf{f} &= [f_i] \in \mathbb{R}^{n_u}, & f_i &:= \int_{\partial D_N} \vec{s} \cdot \vec{\phi}_i - \sum_{j=n_u+1}^{n_u+n_\partial} u_j \int_D \nabla\vec{\phi}_i : \nabla\vec{\phi}_j, \\
\mathbf{g} &= [g_k] \in \mathbb{R}^{n_p}, & g_k &:= \sum_{j=n_u+1}^{n_u+n_\partial} u_j \int_D \psi_k \, (\nabla \cdot \vec{\phi}_j).
\end{aligned} \tag{3.8}$$

For the Stokes equations (continuous and discrete) to be well-posed, a compatibility condition needs to be satisfied at the inflow and outflow boundaries (if any). Moreover, if the discrete system (3.7)–(3.8) is to be a faithful representation of the continuous problem (3.2), then the mixed FEM velocity and pressure spaces need to be chosen carefully such that they satisfy an *inf-sup* (or correspondingly a (discrete) uniform inf-sup) stability condition; see [Elman et al., 2014a, p. 133 ff.] for more details. Typically, choosing more pressure basis functions than velocity basis functions necessarily results in a singular linear system.

Using the popular (piecewise quadratic) $\boldsymbol{Q}_2$-$\boldsymbol{P}_{-1}$ (piecewise linear, discontinuous across elemental boundaries) finite elements or the (Taylor–Hood) $\boldsymbol{Q}_2$-$\boldsymbol{Q}_1$ (piecewise bilinear pressure) finite elements for velocity and pressure space combination leads to inf-sup stable approximations on a rectangular grid. However, the use of higher order

[2]Some discretizations of the Stokes equations can lead to nonsymmetric linear systems. But such discretizations are not considered here.

finite elements might not always provide more accurate FEM solutions, especially if the true solution is not very regular. Because of this reason and from the ease of programming and computational efficiency, $\boldsymbol{Q}_1$-$\boldsymbol{P}_0$ (piecewise constant pressure) finite elements or $\boldsymbol{Q}_1$-$\boldsymbol{Q}_1$ finite elements are attractive choices for velocity-pressure FEM basis. But these approximations are not inf-sup stable on a rectangular grid. In order to make these finite element methods stable, a symmetric positive semi-definite *stabilization matrix* C is introduced in place of the zero block of the coefficient matrix in (3.7). A detailed discussion about the stabilization rationale and strategy for the discrete Stokes system can be found in [Elman et al., 2014a, pp. 139–149].

The symmetric coefficient matrix $K := \begin{bmatrix} \mathbf{A} & B^T \\ B & -C \end{bmatrix}$ of stabilized discrete Stokes system is always indefinite; this follows by applying Sylvester's law of inertia on the matrix K [Elman et al., 2014a, p. 189]. Moreover, it will be assumed that K is nonsingular, that is, it has no zero eigenvalue. Since K is symmetric indefinite, MINRES is the popular and robust iterative method of choice for solving discrete linear systems with coefficient matrix K.

3.3 Block preconditioning

Typically, matrices arising from FEM approximation are ill-conditioned with respect to discretization parameters. Thus, preconditioning is required to accelerate convergence. [Mardal and Winther, 2011] advocate that block diagonal preconditioners are intrinsic choices for symmetric linear systems (in saddle point problems), which arise from numerical approximation of PDEs. Proceeding in this flavour, [Elman et al., 2014a, p. 194 ff.] argue that the symmetric matrix $\begin{bmatrix} \mathbf{A} & O \\ O & B\mathbf{A}^{-1}B^T + C \end{bmatrix}$ is the *desired* but an impractical preconditioner (since $B\mathbf{A}^{-1}B^T + C$ is a dense matrix and hence computing its inverse and also of vector-Laplacian matrix is not cheap) for solving the preconditioned linear system

$$M^{-1}K \begin{bmatrix} \mathbf{u} \\ \mathbf{p} \end{bmatrix} = M^{-1} \begin{bmatrix} \mathbf{f} \\ \mathbf{g} \end{bmatrix}, \tag{3.9}$$

where $M := \begin{bmatrix} \mathbf{P} & O \\ O & S \end{bmatrix}$ is a preconditioner. A practical choice of $\mathbf{P}$ is a block diagonal matrix with each block a preconditioner for the scalar Laplacian matrix (which is on

the diagonal of $\mathbf{A}$). It would be ideal to have $\mathbf{P}$ to be spectrally equivalent to $\mathbf{A}$, that is, there exist positive constants δ_1 and Δ_1 that are independent of discretization parameters such that

$$\delta_1 \leq \frac{\mathbf{u}^T\mathbf{A}\mathbf{u}}{\mathbf{u}^T\mathbf{P}\mathbf{u}} \leq \Delta_1, \qquad \forall\, \mathbf{u} \in \mathbb{R}^{n_u}. \tag{3.10}$$

Indeed this is the case when a Laplacian multigrid preconditioner is used; see [Elman et al., 2014a, lemma 4.2, p. 197]. For the block S of the preconditioner, a good choice is the pressure mass matrix $Q = [q_{kl}]$, $q_{kl} := \int_D \psi_k \psi_l$, $\forall\, k, l = 1, \ldots, n_p$ [Elman et al., 2014a, p. 172]. The matrix Q is spectrally equivalent to the matrix $B\mathbf{A}^{-1}B^T + C$, that is, there exist positive constants γ and Γ that are independent of discretization parameters [Elman et al., 2014a, p. 193–194] such that

$$\gamma^2 \leq \frac{\mathbf{q}^T(B\mathbf{A}^{-1}B^T + C)\mathbf{q}}{\mathbf{q}^T Q\mathbf{q}} \leq \Gamma^2 \leq d, \qquad \forall\, \mathbf{q} \in \mathbb{R}^{n_p} \text{ and } \mathbf{q} \neq \mathbf{1}, \tag{3.11}$$

where d is the dimension of the domain D. In fact the particular choice of $S = \texttt{diag}(Q)$ for continuous ($\boldsymbol{P}_1$ or $\boldsymbol{Q}_1$)[3] makes S spectrally equivalent to Q, that is, there exist positive constants δ_2 and Δ_2 that are independent of discretization parameters [Elman et al., 2014a, pp. 198–199] such that

$$\delta_2^2 \leq \frac{\mathbf{q}^T Q\mathbf{q}}{\mathbf{q}^T S\mathbf{q}} \leq \Delta_2^2, \qquad \forall\, \mathbf{q} \in \mathbb{R}^{n_p}. \tag{3.12}$$

Note that the constant γ in (3.11) is the uniform inf-sup constant when $C = 0$ and $\delta^2 = 2\gamma^2$, where δ is the uniform inf-sup constant for the case when $C \neq 0$.

Having formulated a mixed FEM matrix formulation of (3.2) and discussed briefly about MINRES preconditioners to be used for solving the corresponding discrete linear system, the balanced stopping strategy is presented in the next section.

3.4 A balanced stopping test

According to [Wathen, 2007], a natural norm for a function in the space of square integrable functions is its L^2 norm while the L^2 norm of the gradient of the function is a natural choice if the function is in $H^1_{E_0}$. So, a natural choice of norm $(\|\cdot\|_\mathcal{E})$ for any $(\vec{u}, p) \in \mathbf{H}^1_{E_0}(D) \times L^2(D)$ is[4]

$$\|(\vec{u}, p)\|_\mathcal{E} := \|\nabla\vec{u}\|_2 + \|p\|_2. \tag{3.13}$$

[3]For $\boldsymbol{P}_0$ pressure approximation, Q is in fact diagonal.
[4]Note that $\nabla\vec{u}$ is to be interpreted componentwise.

In terms of vectors, $\|\cdot\|_{\mathcal{E}}$ translates into the norm $\|\cdot\|_E$

$$\|\mathbf{e}\|_E := \sqrt{\mathbf{e}^T E \mathbf{e}} = \sqrt{\mathbf{e}_1^T \mathbf{A} \mathbf{e}_1 + \mathbf{e}_2^T Q \mathbf{e}_2}, \qquad \forall\, \mathbf{e} = [\mathbf{e}_1^T,\, \mathbf{e}_2^T]^T \in \mathbb{R}^{n_u + n_p}, \tag{3.14}$$

where $\mathbf{e}_1 \in \mathbb{R}^{n_u}, \mathbf{e}_2 \in \mathbb{R}^{n_p}$, and $E := \begin{bmatrix} \mathbf{A} & O \\ O & Q \end{bmatrix}$. Since the vector-Laplacian matrix $\mathbf{A}$ and the pressure mass matrix Q are both symmetric positive-definite, the matrix E is also symmetric positive-definite and hence $\|\cdot\|_E$ is indeed a norm on $\mathbb{R}^{n_u + n_p}$.

3.4.1 Error equation

For a given approximation, by the triangle inequality at iteration k

$$\underbrace{\|(\vec{u} - \vec{u}_h^{(k)}, p - p_h^{(k)})\|_{\mathcal{E}}}_{\text{total error}} \;\leq\; \underbrace{\|(\vec{u} - \vec{u}_h, p - p_h)\|_{\mathcal{E}}}_{\text{approximation error}} + \underbrace{\|(\vec{u}_h - \vec{u}_h^{(k)}, p_h - p_h^{(k)})\|_{\mathcal{E}}}_{\text{algebraic error}}, \tag{3.15}$$

where $(\vec{u}, p)$ is the true solution, $(\vec{u}_h, p_h)$ is the true mixed FEM solution, and $(\vec{u}_h^{(k)}, p_h^{(k)})$ is the FEM solution formed from the kth iterate of the chosen iterative solver. It follows from the definition of $\|\cdot\|_{\mathcal{E}}$ from (3.13) that

$$\|(\vec{u}_h - \vec{u}_h^{(k)}, p_h - p_h^{(k)})\|_{\mathcal{E}} = \|\nabla(\vec{u}_h - \vec{u}_h^{(k)})\|_2 + \|p_h - p_h^{(k)}\|_2. \tag{3.16}$$

Note that

$$\begin{aligned} \|\nabla(\vec{u}_h - \vec{u}_h^{(k)})\|_2 &= \sqrt{(\mathbf{e}_1^{(k)})^T \mathbf{A} \mathbf{e}_1^{(k)}}, \qquad \mathbf{e}_1^{(k)} = [u_{1_h} - u_{1_h}^{(k)}, \ldots, u_{n_{uh}} - u_{n_{uh}}^{(k)}]^T, \\ \|p_h - p_h^{(k)}\|_2 &= \sqrt{(\mathbf{e}_2^{(k)})^T Q \mathbf{e}_2^{(k)}}, \qquad \mathbf{e}_2^{(k)} = [p_{1_h} - p_{1_h}^{(k)}, \ldots, p_{n_{ph}} - p_{n_{ph}}^{(k)}]^T, \end{aligned} \tag{3.17}$$

where $\vec{u}_h - \vec{u}_h^{(k)} = \sum_{i=1}^{n_u} (u_{i_h} - u_{i_h}^{(k)}) \vec{\phi}_i$, $p_h - p_h^{(k)} = \sum_{j=1}^{n_p} (p_{j_h} - p_{j_h}^{(k)}) \psi_j$. Also, for any two nonnegative real numbers a and b [Elman et al., 2014a, p. 213]

$$\sqrt{a+b} \;\leq\; \sqrt{a} + \sqrt{b} \;\leq\; \sqrt{2}\sqrt{a+b}. \tag{3.18}$$

Putting $a = \|\nabla(\vec{u}_h - \vec{u}_h^{(k)})\|_2^2$, $b = \|p_h - p_h^{(k)}\|_2^2$ in (3.18) and using (3.17), (3.14) gives

$$\|\mathbf{e}^{(k)}\|_E \;\leq\; \|\nabla(\vec{u}_h - \vec{u}_h^{(k)})\|_2 + \|p_h - p_h^{(k)}\|_2 \;\leq\; \sqrt{2}\|\mathbf{e}^{(k)}\|_E. \tag{3.19}$$

For enclosed flow problems, a slight variant of the L^2 norm known as the *quotient space norm* $\|\cdot\|_{0,D}$ is used for measuring pressure. Here $\|q_h\|_{0,D} = \|q_h - \frac{1}{|D|}\int_D q_h\|_2$,

$|D| = \int_D$ for any $q_h \in M^h$ [Elman et al., 2014a, p. 128]. Note that

$$\begin{aligned} \|q_h - \frac{1}{|D|}\int_D q_h\|_2^2 &= \int_D \left(q_h - \frac{1}{|D|}\int_D q_h\right)^2 \\ &= \int_D q_h\, q_h + \int_D \left(\frac{1}{|D|}\int_D q_h\right)^2 - \int_D 2q_h \left(\frac{1}{|D|}\int_D q_h\right) \\ &= \|q_h\|_2^2 + \frac{1}{|D|}\left(\int_D q_h\right)^2 - 2\frac{1}{|D|}\left(\int_D q_h\right)^2 \\ &= \|q_h\|_2^2 - \frac{1}{|D|}\left(\int_D q_h\right)^2 \leq \|q_h\|_2^2, \end{aligned} \tag{3.20}$$

since $\frac{1}{|D|}\left(\int_D q_h\right)^2 \geq 0$. So, $\|q_h\|_{0,D} \leq \|q_h\|_2$, $\forall\, q_h \in M^h$. Thus, the (quotient space norm) algebraic error $\|(\vec{u}_h - \vec{u}_h^{(k)}, p_h - p_h^{(k)})\|_\varepsilon = \|\nabla(\vec{u}_h - \vec{u}_h^{(k)})\|_2 + \|p_h - p_h^{(k)}\|_{0,D}$ can be bounded from above by the usual L^2 norm of the algebraic error, that is

$$\|\nabla(\vec{u}_h - \vec{u}_h^{(k)})\|_2 + \|p_h - p_h^{(k)}\|_{0,D} \leq \|\nabla(\vec{u}_h - \vec{u}_h^{(k)})\|_2 + \|p_h - p_h^{(k)}\|_2. \tag{3.21}$$

Using (3.21) one can obtain the same bound (3.19) for the enclosed flow algebraic error at kth iterative step in terms of $\|\mathbf{e}^{(k)}\|_E$ norm of the kth iteration error.

A handle on the approximation error and the total error (approximation error at the kth iteration) is obtained with a posteriori error estimators η and $\eta^{(k)}$ respectively. The a posteriori error estimator $\eta^{(k)}$ is equivalent to the total error in the sense that

$$c_1\, \eta^{(k)} \leq \|\nabla(\vec{u} - \vec{u}_h^{(k)})\|_2 + \|p - p_h^{(k)}\|_2 \leq C_1\, \eta^{(k)}, \quad \text{with } \frac{C_1}{c_1} \sim O(1), \tag{3.22}$$

If the a posteriori error estimators η and $\eta^{(k)}$ are assumed to be 'close' estimates of the approximation error and total error (at kth iteration step) respectively, then the error equation (3.15) can be rewritten as

$$\eta^{(k)} \simeq \eta + \|\mathbf{e}^{(k)}\|_E, \qquad k = 0, 1, 2, \ldots. \tag{3.23}$$

The relation $\simeq$ is a result of (3.22) and (3.19). In fact it follows from (3.23) that when the norm $\|\mathbf{e}^{(k)}\|_E$ of the iteration error $\mathbf{e}^{(k)}$ is 'small', then $\{\eta^{(k)}\}$ converges to η. Thus, one would stop optimally when $\|\mathbf{e}^{(k)}\|_E$ and the a posteriori error estimate $\eta^{(k)}$ of the total error are balanced, that is, stop at the first iteration k^* such that

$$\|\mathbf{e}^{(k^*)}\|_E \leq \eta^{(k^*)}. \tag{3.24}$$

In the subsequent subsections, a brief discussion on the a posteriori error estimation for the Stokes equations is done and tractable bounds on difficult to compute $\|\mathbf{e}^{(k)}\|_E$ are derived.

3.4.2 Tractable bounds on algebraic error

In preconditioned MINRES with symmetric positive-definite preconditioner M, the norm $\|\mathbf{r}^{(k)}\|_{M^{-1}} := \sqrt{\mathbf{r}^{(k)^T} M^{-1}\mathbf{r}^{(k)}}$ is monotonically decreasing with iteration count k and hence a suitable surrogate norm for computations in place of $\|\mathbf{e}^{(k)}\|_E$. Here $\mathbf{r}^{(k)} := K\mathbf{e}^{(k)}$ is the residual at iteration k. Thus, one obtains an expression for the algebraic error at kth iterative step in terms of the iteration residual $\mathbf{r}^{(k)}$, that is

$$\|\mathbf{e}^{(k)}\|_E^2 \;=\; (\mathbf{e}^{(k)})^T E\mathbf{e}^{(k)} \;=\; (\mathbf{r}^{(k)})^T K^{-T}EK^{-1}\mathbf{r}^{(k)}. \tag{3.25}$$

It follows from (3.25) that bounding $\|\mathbf{e}^{(k)}\|_E$ by $\|\mathbf{r}^{(k)}\|_{M^{-1}}$ requires computing constants c_2 and C_2 such that

$$c_2 \;\leq\; \frac{(\mathbf{r}^{(k)})^T K^{-T}EK^{-1}\mathbf{r}^{(k)}}{(\mathbf{r}^{(k)})^T M^{-1}\mathbf{r}^{(k)}} \;\leq\; C_2, \tag{3.26}$$

This leads to computing extremal Rayleigh quotient bounds of $K^{-T}EK^{-1}$ and M^{-1}, that is, find $\lambda_{\min}$, $\lambda_{\max} \in \mathbb{R}$ such that

$$\lambda_{\min} \;\leq\; \frac{\mathbf{v}^T K^{-T}EK^{-1}\mathbf{v}}{\mathbf{v}^T M^{-1}\mathbf{v}} \;\leq\; \lambda_{\max}, \qquad \forall\, \mathbf{v} \in \mathbb{R}^{n_u+n_p}. \tag{3.27}$$

Equation (3.27) implies that one needs to compute generalized extremal eigenvalues for $K^{-T}EK^{-1}$ and M^{-1}, that is, find the extremal eigenvalues λ such that

$$K^{-T}EK^{-1}\mathbf{y} \;=\; \lambda M^{-1}\mathbf{y}, \qquad \mathbf{y} \in \mathbb{R}^{n_u+n_p} \text{ is an eigenvector.} \tag{3.28}$$

Note that the matrices K, E are symmetric so the matrix $K^{-T}EK^{-1}$ is also symmetric. Also, since M is symmetric positive-definite, its inverse M^{-1} is also symmetric positive-definite. So, the generalized eigenvalue problem (3.28) can be converted (theoretically) into a symmetric algebraic eigenvalue problem through a Cholesky factorization of M^{-1}. Hence all λ's in (3.28) are real. Let $\mathbf{z} = K^{-1}\mathbf{y}$, then (3.28) becomes

$$K^{-T}E\mathbf{z} \;=\; \lambda M^{-1}K\mathbf{z}, \qquad \mathbf{z} \in \mathbb{R}^{n_u+n_p}. \tag{3.29}$$

It is clear from the discussions in section 3.3 that an ideal but an impractical choice for the preconditioner M is the matrix E. A more practical choice is where the matrices $\mathbf{P}$ and S satisfy (3.10) and (3.12) respectively and hence M is spectrally equivalent to E. Thus, for 'good' choices of $\mathbf{P}$ and S, M will 'behave like' E after a 'few'

iterations. Therefore, the analysis presented here will be for the ideal preconditioner E. Substituting E for M in (3.29), and using that K is symmetric gives

$$(E^{-1}K)^{-1}\mathbf{z} \;=\; \lambda E^{-1}K\mathbf{z}, \qquad \mathbf{z}\in\mathbb{R}^{n_u+n_p}. \tag{3.30}$$

Let $W := E^{-1}K$, then (3.30) can be rearranged as the following eigenvalue problem

$$W^2\mathbf{z} \;=\; \mu\mathbf{z}, \qquad \mathbf{z}\in\mathbb{R}^{n_u+n_p}, \tag{3.31}$$

where $\mu = 1/\lambda$. Note that since $W = E^{-1}K$ is symmetric and nonsingular, all its eigenvalues are real and nonzero. So, the eigenvalues μ's of W^2 (which are the squares of eigenvalues of W) are all real and greater than zero. So, any λ cannot be zero; in fact all λ's are greater than zero.

In light of (3.29), (3.30), and (3.31) the eigenvalue problem (3.28) is transformed into finding the largest ($\mu_{\max}$) and smallest ($\mu_{\min}$) eigenvalues of W^2 such that

$$W^2\mathbf{z} \;=\; \mu\mathbf{z}, \qquad \mathbf{z}\in\mathbb{R}^{n_u+n_p} \text{ is an eigenvector.} \tag{3.32}$$

Since the eigenvalues of W^2 are just the square of the eigenvalues of W, it is sufficient to compute the eigenvalues of W. In fact, one obtains

$$\mu_{\max} \;=\; \max\{|\theta^+_{\max}|^2, |\theta^-_{\min}|^2\}, \tag{3.33a}$$

$$\mu_{\min} \;=\; \min\{|\theta^+_{\min}|^2, |\theta^-_{\max}|^2\}, \tag{3.33b}$$

where θ's are eigenvalues of W such that

$\theta^+_{\max}$ – maximum positive eigenvalue, $\theta^+_{\min}$ – minimum positive eigenvalue,

$\theta^-_{\max}$ – maximum negative eigenvalue, $\theta^-_{\min}$ – minimum negative eigenvalue.

3.4.3 Stopping criterion

Using $\lambda_{\min} = \frac{1}{\mu_{\max}}, \lambda_{\max} = \frac{1}{\mu_{\min}}$; (3.26), (3.27), and (3.33) can be combined into

$$\frac{1}{\max\{|\theta^+_{\max}|^2, |\theta^-_{\min}|^2\}} \;\le\; \frac{(\mathbf{r}^{(k)})^T K^{-T}EK^{-1}\mathbf{r}^{(k)}}{(\mathbf{r}^{(k)})^T M^{-1}\mathbf{r}^{(k)}} \;\le\; \frac{1}{\min\{|\theta^+_{\min}|^2, |\theta^-_{\max}|^2\}}. \tag{3.34}$$

It follows from (3.34) that

$$\frac{1}{\sqrt{\max\{|\theta^+_{\max}|^2, |\theta^-_{\min}|^2\}}} \;\le\; \frac{\|\mathbf{e}^{(0)}\|_E}{\|\mathbf{r}^{(0)}\|_{M^{-1}}}, \qquad \frac{\|\mathbf{e}^{(k)}\|_E}{\|\mathbf{r}^{(k)}\|_{M^{-1}}} \;\le\; \frac{1}{\sqrt{\min\{|\theta^+_{\min}|^2, |\theta^-_{\max}|^2\}}}. \tag{3.35}$$

Equation (3.35) leads to the following upper bounds on $\|\mathbf{e}^{(k)}\|_E$, that is

$$\|\mathbf{e}^{(k)}\|_E \;\leq\; \frac{1}{\sqrt{\min\{|\theta^-_{\max}|^2, |\theta^+_{\min}|^2\}}}\;\|\mathbf{r}^{(k)}\|_{M^{-1}},$$

$$\begin{aligned} \frac{\|\mathbf{e}^{(k)}\|_E}{\|\mathbf{e}^{(0)}\|_E} \;&\leq\; \sqrt{\frac{\max\{|\theta^+_{\max}|^2, |\theta^-_{\min}|^2\}}{\min\{|\theta^-_{\max}|^2, |\theta^+_{\min}|^2\}}}\;\frac{\|\mathbf{r}^{(k)}\|_{M^{-1}}}{\|\mathbf{r}^{(0)}\|_{M^{-1}}} \\ \Longleftrightarrow \|\mathbf{e}^{(k)}\|_E \;&\leq\; \frac{\sqrt{\max\{|\theta^+_{\max}|^2, |\theta^-_{\min}|^2\}}}{\min\{|\theta^-_{\max}|^2, |\theta^+_{\min}|^2\}}\|\mathbf{r}^{(k)}\|_{M^{-1}}. \end{aligned} \tag{3.36}$$

Thus, from (3.24) it follows that an optimal stopping point is the first iteration k^* at which one of the following tests is satisfied

$$\frac{\sqrt{\max\{|\theta^+_{\max}|^2, |\theta^-_{\min}|^2\}}}{\min\{|\theta^-_{\max}|^2, |\theta^+_{\min}|^2\}}\|\mathbf{r}^{(k^*)}\|_{M^{-1}} \;\leq\; \eta^{(k^*)}. \tag{3.37}$$

$$\frac{1}{\sqrt{\min\{|\theta^-_{\max}|^2, |\theta^+_{\min}|^2\}}}\;\|\mathbf{r}^{(k^*)}\|_{M^{-1}} \;\leq\; \eta^{(k^*)}. \tag{3.38}$$

Henceforth, the stopping test (3.37) will be called the stronger stopping test while the stopping test (3.38) will be called the weaker stopping test.

3.4.4 A posteriori error estimation

The a posteriori error estimation technique used in the software IFISS for Stokes equations is due to [Ainsworth and Oden, 1997] and it essentially involves solving a local Poisson problem for each velocity component; see [Elman et al., 2014a, section 3.4.2]. The a poseriori error estimator based on this strategy provides 'acceptable' close estimates of the true total (approximation) errors. In fact for $\boldsymbol{Q}_1$-$\boldsymbol{P}_0$ rectangular finite elements, this a posteriori error estimator is both a global upper bound, (that is, it is reliable) and a local elementwise bound (that is, it is efficient) on the actual error; see [Kay and Silvester, 1999] for full details. A comparison of η [Elman et al., 2014a, table 3.4, p. 169] and 'actual' approximation error $\|\nabla(\vec{u} - \vec{u}_h)\|_2 + \|p - p_h\|_{0,2}$ [Elman et al., 2014a, table 3.3, p. 166] are tabulated in Table 3.1. The results presented therein are for the Stokes test problem 1 [Elman et al., 2014a, p. 126] in section 3.5 with $\boldsymbol{Q}_1$-$\boldsymbol{P}_0$ rectangular finite elements on a uniform grid and mesh step size h.

The entries for corresponding effectivity index $\beta_{\text{eff}} = \dfrac{\eta}{\|\nabla(\vec{u} - \vec{u}_h)\|_2 + \|p - p_h\|_{0,2}}$ in Table 3.1 show that a posteriori approximation error estimator employed here is an 'acceptable close' estimate of the true error.

Table 3.1: Actual approximation errors, a posteriori errors, and effectivity indices for $\boldsymbol{Q}_1$-$\boldsymbol{P}_0$ rectangular finite elements on uniform grids for Stokes test problem 1.

h	η	$\|\nabla(\vec{u}-\vec{u}_h)\|_2 + \|p-p_h\|_{0,2}$	β_{eff}
1/4	9.501	18.729	0.51
1/8	5.307	8.853	0.59
1/16	2.761	4.290	0.64
1/32	1.399	2.116	0.66

3.4.5 Computational logistics

The M^{-1} norm of the iteration residual, that is, $\|\mathbf{r}^{(k)}\|_{M^{-1}}$ is readily available in preconditioned MINRES. Also, it is advisable in general to compute $\eta^{(k)}$ periodically to minimize the overall algorithmic cost. The eigenvalues involved in the stopping test (3.37) and (3.38) can be estimated cheaply on-the-fly, the strategy for which is described in the next subsection.

3.4.6 Cheap estimation of eigenvalues in stopping test

Note that the extremal eigenvalues $\theta^{+}_{\max}$, $\theta^{-}_{\min}$ of the preconditioned matrix can cheaply be estimated by the corresponding extremal Ritz values $\theta^{k+}_{\max}$, $\theta^{k-}_{\min}$ (the maximum positive Ritz value and the minimum negative Ritz value respectively) of the Lanczos matrix T_k in preconditioned MINRES; see section 2.3.1. But for the interior most eigenvalues $\theta^{+}_{\min}$ and $\theta^{-}_{\max}$, the Ritz values usually provide a poor estimation. So, the interior most eigenvalues are estimated here by computing the corresponding interior most eigenvalues $\theta^{k+}_{\min}$, $\theta^{k-}_{\max}$ of the following generalized eigenvalue problem

$$\underline{T}_k^T \underline{T}_k \mathbf{y} \;=\; \theta_{\text{har}} T_k \mathbf{y}, \qquad \mathbf{y} \text{ is an eigenvector.} \tag{3.39}$$

where $\underline{T}_k$ is the $\mathbb{R}^{(k+1)\times k}$ Lanczos matrix; see section 2.3.1 for more details. The eigenvalues θ_{har} in (3.39) are known as *harmonic Ritz values*; see [Bai et al., 2000, section 3.2, p. 41–43]. Here $\theta^{k+}_{\min}$ and $\theta^{k-}_{\max}$ denote the minimum positive harmonic Ritz value and the maximum negative harmonic Ritz value respectively. Unlike the Ritz values, which approximate first the extremal eigenvalues of the preconditioned matrix, the harmonic Ritz values approximate first the interior most eigenvalues of the preconditioned matrix. This is better than using Ritz values to estimate the actual interior most eigenvalues since the interior most Ritz values might take a long time to

provide a good approximation (if at all) to the interior most eigenvalues.

Further insight into the eigenvalues of the preconditioned matrix is obtained from the following result in [Elman et al., 2014a, theorem 4.7, p. 201].

Theorem 3.4.1. *The eigenvalues of $M^{-1}K$ satisfy*

$$\begin{aligned} -\Delta_2^2\left(\Gamma^2 + \Upsilon\right) \leq \theta_{\min}^- \leq \theta_{\max}^- \leq \frac{1}{2}\left(\delta_1 - \sqrt{\delta_1^2 + 4\delta_1\gamma^2\delta_2^2}\right), \\ \delta_1 \leq \theta_{\min}^+ \leq \theta_{\max}^+ \leq \Delta_1 + \Gamma^2\Delta_2^2, \end{aligned} \tag{3.40}$$

where $\delta_1, \Delta_1, \delta_2, \Delta_2, \gamma,$ and Γ are the same as in (3.10), (3.12), and (3.11) respectively. The constant Υ satisfies

$$\frac{\mathbf{q}^T C\mathbf{q}}{\mathbf{q}^T Q\mathbf{q}} \leq \Upsilon, \qquad \forall\, \mathbf{q} \in \mathbb{R}^{n_p}. \tag{3.41}$$

Proceeding along the lines of [Silvester and Simoncini, 2011] note that for $M = E$, $\delta_1 = \Delta_1 = 1$. Also, for $\boldsymbol{P}_0$ pressure approximation $\delta_2 = \Delta_2 = 1$. In any case if preconditioner blocks $\mathbf{P}$ and S 'closely' approximate $\mathbf{A}$ and Q respectively, then

$$\delta_1 \simeq 1, \Delta_1 \simeq 1, \delta_2 \simeq 1, \text{ and } \Delta_2 \simeq 1. \tag{3.42}$$

Also, the asymptotic simplification $(1+x)^{\frac{1}{2}} = 1 + \frac{1}{2}x$ gives

$$\frac{1}{2}\left(\delta_1 - \sqrt{\delta_1^2 + 4\delta_1\gamma^2\delta_2^2}\right) \simeq \frac{1}{2}\left(1 - \sqrt{1 + 4\gamma^2}\right) \simeq -\gamma^2. \tag{3.43}$$

Combining (3.40), (3.42), and (3.43) leads to

$$\theta_{\max}^- \simeq -\gamma^2 \leq 1 \simeq \theta_{\min}^+. \tag{3.44}$$

The validity of the equivalence $\theta_{\max}^- \simeq -\gamma^2$ for $C = 0$ case can be further confirmed from the discussions in [Elman et al., 2014a, pp. 196–197]. If $\gamma^2 \leq 1$, which is usually the case[5] then from (3.44) it follows $\dfrac{1}{\sqrt{\min\{|\theta_{\max}^-|^2, |\theta_{\min}^+|^2\}}} = \dfrac{1}{\sqrt{|\theta_{\max}^-|^2}} \simeq \dfrac{1}{\sqrt{\gamma^4}} = \dfrac{1}{\gamma^2}$. In light of this analysis, the weaker stopping test (3.38) can be transformed into

$$\frac{1}{\gamma^2}\,\|\mathbf{r}^{(k^*)}\|_{M^{-1}} \leq \eta^{(k^*)} \quad \Longleftrightarrow \quad \|\mathbf{r}^{(k^*)}\|_{M^{-1}} \leq \gamma^2\,\eta^{(k^*)}. \tag{3.45}$$

An equivalence similar to (3.44) holds for maximum positive eigenvalue and minimum negative eigenvalue

$$\theta_{\min}^- \simeq -\left(\Gamma^2 + \Upsilon\right) \leq (1 + \Gamma^2) \simeq \theta_{\max}^+. \tag{3.46}$$

[5] In fact $\gamma^2 \leq \Gamma^2 \leq d$, where d (equal to 2 or 3 here) denotes the dimensionality of the domain D. But for $C = 0$ with Dirichlet boundary conditions in $\mathbb{R}^2, \gamma^2 \leq 1$; see [Elman et al., 2014a, theorem 3.22, p. 174]

Since $0 \leq \Upsilon \leq 1$ [Elman et al., 2014a, p. 200], it follows from equation (3.46) that $\sqrt{\max\{|\theta^{+}_{\max}|^2, |\theta^{-}_{\min}|^2\}} = \sqrt{|\theta^{+}_{\max}|^2} \simeq (1 + \Gamma^2) \leq 1 + d$. Combining this with $\min\{|\theta^{-}_{\max}|^2, |\theta^{+}_{\min}|^2\} = |\theta^{-}_{\max}|^2 \simeq \gamma^4$, the stronger stopping test (3.37) becomes

$$\frac{1+d}{\gamma^4}\,\|\mathbf{r}^{(k^*)}\|_{M^{-1}} \leq \eta^{(k^*)} \quad \Longleftrightarrow \quad \|\mathbf{r}^{(k^*)}\|_{M^{-1}} \leq \frac{\gamma^4}{1+d}\,\eta^{(k^*)}. \tag{3.47}$$

In presence of 'tight' a posteriori error estimators and 'good' preconditioner blocks, the stopping test (3.45) or (3.47) can be used and they hold for both $C = 0$ and $C \neq 0$.

Table 3.2: Comparison of literature and improved stopping tests for $\boldsymbol{Q}_2$-$\boldsymbol{P}_1$ finite elements on rectangular uniform grids for Stokes test problem 1.

h	k^*_{lit}	e^*_{lit}	k^*_{imp}	e^*_{imp}
1/8	10	6.0e-2	8	5.4e-2
1/16	17	1.0e-5	15	5.6e-3
1/32	21	3.1e-4	19	1.7e-3
1/64	24	1.1e-4	24	1.1e-4

In fact for the case $C = 0$, using (3.45) one stops optimally a 'bit' earlier than using the stopping test $\|\mathbf{r}^{(k^*)}\|_{M^{-1}} \leq \dfrac{\gamma^2}{\sqrt{2}}\,\eta^{(k^*)}$ of [Silvester and Simoncini, 2011]. This is because $\dfrac{\gamma^2}{\sqrt{2}} < \gamma^2$ and hence an 'improvement' of constants (over those in the existing literature) involved in the stopping test for the case $C = 0$ has been obtained here. Note that this improvement is only a theoretical result. Since $\sqrt{2} \approx 1.41$, in practice a gain of only 'very few' (if any) iterations is obtained by using γ^2 over $\dfrac{\gamma^2}{\sqrt{2}}$ in balanced stopping (3.45); see Table 3.2. The stopping iteration k^*_{lit} and k^*_{imp} corresponding to the stopping test in [Silvester and Simoncini, 2011] and (3.45) respectively are tabulated in Table 3.2. Also, at each grid level tabulated are $e^*_{\text{lit}} := |\eta - \eta^{(k^*_{\text{lit}})}|$ and $e^*_{\text{imp}} := |\eta - \eta^{(k^*_{\text{imp}})}|$. These denote the corresponding absolute differences in a posteriori error estimates from the actual a posteriori estimate η obtained using the 'true' solution (MATLAB backslash solution). It follows from Table 3.2 that savings of only a few iterations is obtained on using the stopping test (3.45) over that in [Silvester and Simoncini, 2011]. Also, at the stopping iteration for both these stopping tests, the sequence $\{\eta^{(k)}\}$ has converged with some accuracy to the true η; see columns for e^*_{lit} and e^*_{imp}. These numbers have been obtained by running `itsolve_stokes` with default options in IFISS toolbox of MATLAB after setting up the Stokes test problem 1 that

is described in section 3.5. The constants involved in the stopping test can be modified suitably in the function `param_est` in IFISS.

3.4.7 Choice of stopping test

A drawback of using the stopping test (3.37) or (3.38) is that they might lead to premature stopping because one or more of the computed extremal Ritz values or the interior most harmonic Ritz values would have not yet converged to their corresponding (discrete system) actual eigenvalue respectively. Although this convergence is usually quite fast, it is generally difficult to determine beforehand the iteration count at which they will converge. Hence, it is proposed here to store the required Ritz and harmonic Ritz values of previous 4-5 consecutive iterations and apply the stopping test (3.37) or (3.38) only when the absolute successive differences of these values for each of the required quantities is below a prescribed tolerance of 10^{-2} (say).[6]

Substituting (3.45) for (3.38) and (3.47) for (3.37) overcomes this drawback. This is because the constants in the stopping tests now depend on d, which is trivially known and the discrete inf-sup constant γ, which in many practical applications is known beforehand and depends only on the topology of the spatial domain; see [Chizhonkov and Olshanskii, 2000]. However, the stopping tests (3.45) and (3.47) were derived using many equivalences ($\simeq$) which may not be tight in general. Hence, in presence of a preconditioner M which is spectrally equivalent to E, it will be better to employ the weaker or the stronger stopping test based on interior most harmonic Ritz values and extremal Ritz values.

The resulting algorithm known as `SADDLE_MINRES` in the software IFISS is given in the form of pseudo-code in Figure 3.1. Similar to chapter 2, the external functions `matvecK`, `precM` compute the action of the matrices K and M^{-1} on a vector respectively while the function `Stokes_error_est` computes the a posteriori error estimate. Also, $\mathbf{b} = [\mathbf{f}^T, \mathbf{g}^T]^T$ denotes the right-hand-side vector in Figure (3.1). This algorithm can easily be modified for the weaker stopping test. A practical implementation of this algorithm should incorporate periodic computations of the a posteriori error estimate. Also, it should involve storage of previous 4-5 values from consecutive iterations for each of the Ritz and the harmonic Ritz values involved in the balanced stopping test.

[6]For the weaker stopping test this procedure has to be done for only the harmonic Ritz values.

Algorithm: `SADDLE_MINRES`
given vectors $\mathbf{b}$, $\mathbf{x}^{(0)}$ and functions `matvecK`, `precM`, `param_intest`, `param_extest` `Stokes_error_est`

set $\mathbf{r}^{(0)} = \mathbf{b} - \texttt{matvecK}\,(\mathbf{x}^{(0)})$, $\hat{\mathbf{r}}^{(0)} = \texttt{precM}\,(\mathbf{r}^{(0)})$, $\rho_0 = \sqrt{(\mathbf{r}^{(0)})^T \hat{\mathbf{r}}^{(0)}}$
initialize basis vectors: $\mathbf{w} = \hat{\mathbf{r}}^{(0)}/\rho_0$, $\mathbf{p}^{(-1)} = \mathbf{0}$, $\mathbf{p}^{(0)} = \mathbf{r}^{(0)}/\rho_0$
initialize auxiliary vectors: $\mathbf{d}^{(-1)} = \mathbf{0}$, $\mathbf{d}^{(0)} = \mathbf{0}$
initialize projected right-hand side: $f = \rho_0$

`for` $k = 1, 2, \ldots$ `until` convergence `do`
  generate new basis and auxiliary vectors: $\mathbf{p}^{(k)} = \texttt{matvecK}\,(\mathbf{w})$, $\mathbf{d}^{(k)} = \mathbf{w}$
    `if` k>1, $t_{k-1,k} = t_{k,k-1}$, $\mathbf{p}^{(k)} = \mathbf{p}^{(k)} - \mathbf{p}^{(k-1)} t_{k-1,k}$
  $t_{k,k} = \mathbf{w}^T \mathbf{p}^{(k)}$, $\mathbf{p}^{(k)} = \mathbf{p}^{(k)} - \mathbf{p}^{(k-1)} t_{k,k}$
  compute preconditioned basis vector: $\mathbf{w} = \texttt{precM}\,(\mathbf{p}^{(k)})$
  $t_{k+1,k} = \sqrt{\mathbf{w}^T \mathbf{p}^{(k)}}$, $\mathbf{p}^{(k)} = \mathbf{p}^{(k)}/t_{k+1,k}$, $\mathbf{w} = \mathbf{w}/t_{k+1,k}$
  compute parameters for stopping test:
  `coefext` = `param_extest` (T_k)
  `coefint` = `param_intest` $(T_k,\, t_{k+1,k})$
  `coef` = `coefext`/(`coefint`)^2
  apply previous rotations:
    `if` k>2, $\rho_{1:2} = S_{k-2} t_{k-2:k-1,k}$, $\rho_{2:3} = S_{k-1}[\rho_2; t_{k,k}]$
    `elseif` k=2, $\rho_{2:3} = S_{k-1} t_{1:2,2}$
    `elseif` $k = 1$, $\rho_3 = t_{1,1}$
  compute new rotations:
  $\hat{\delta} = \sqrt{\rho_3^2 + t_{k+1,k}^2}$, $c = |\rho_3|/\hat{\delta}$, $s = \mathrm{sign}(\rho_3) t_{k+1,k}/\hat{\delta}$
  apply new rotations: $\rho_3 = c\rho_3 + s t_{k+1,k}$, $\hat{f} = -sf$, $f = cf$, $S_k = [c\ s; -s\ c]$
  update auxiliary vector: $\mathbf{d}^{(k)} = (\mathbf{d}^{(k)} - \mathbf{d}^{(k-1)}\rho_1 - \mathbf{d}^{(k-2)}\rho_2)/\rho_3$
  update solution: $\mathbf{x}^{(k)} = \mathbf{x}^{(k-1)} + \mathbf{d}^{(k)} \hat{f}$
  compute discretization error estimate : $\eta^{(k)} = \texttt{Stokes\_error\_est}\,(\mathbf{x}^{(k)})$
  stopping test: `if` `coef`$\cdot|\hat{f}| \leq \eta^{(k)}$, convergence
  update residual norm: $f = \hat{f}$
`enddo`

function `coefext` = `param_extest` (T_k)
  compute the smallest negative eigenvalue $\theta_{\min}^{k-}$ and the largest positive eigenvalue $\theta_{\max}^{k+}$ of T_k
    `if` $|\theta_{\min}^{k-}|^2 \leq |\theta_{\max}^{k+}|^2$ set `coefext` $= \theta_{\max}^{k+}$
    `else` set `coefext` $= |\theta_{\min}^{k-}|$
`endfunction`

function `coefint` = `param_intest` $(T_k,\, t_{k+1,k})$
  compute the smallest positive eigenvalue $\theta_{\min}^{k+}$ and the largest negative eigenvalue $\theta_{\max}^{k-}$ of generalized eigenvalue problem $\underline{T}_k^T \underline{T}_k$ and T_k
    `if` $|\theta_{\max}^{k-}|^2 \leq |\theta_{\min}^{k+}|^2$ set `coefint` $= |\theta_{\max}^{k-}|$
    `else` set `coefint` $= |\theta_{\min}^{k+}|$
`endfunction`

Figure 3.1: The `SADDLE_MINRES` algorithm expressed in pseudo-code.

3.5 Computational results

To provide a proof-of-concept, some computational results are presented in this section for two test problems in IFISS. The stronger stopping test (3.37) is employed for both the test problems in order to exhibit the nuances associated with using a stopping test based on both interior most and exterior most eigenvalues of the preconditioned matrix. It has also been observed from computations for the test problems considered here that the relevant extremal Ritz values and the interior most harmonic Ritz values have converged with some accuracy before optimal stopping has been reached. So, one does not need to store previous 4-5 values from consecutive iterations for these quantities. Also, instead of computing the a posteriori error estimator periodically, it is computed here at each iteration to illustrate the balanced stopping methodology.

There are four preconditioners built in IFISS for the discrete Stokes problem. They are: diagonal (DIAG) preconditioner—the diagonal matrix formed from the diagonal elements of $\mathbf{A}$ and the diagonal entries of Q—the block ideal preconditioner E, block geometric multigrid (GMG), and block algebraic multigrid (AMG) [Elman et al., 2014a, chapter 4] preconditioners. Results are presented here for block ideal and block AMG preconditioners for both the test problems. Note that the block AMG preconditioner is employed with its specified default settings in IFISS.

Piecewise bilinear ($\boldsymbol{Q}_1$) finite elements are used for FEM velocity space and $\boldsymbol{P}_0$ finite elements are employed for FEM pressure space on rectangular grids. The uniform mesh step size h is used for the test problem 1 while $2^l \times (2^l \times 3)$ grids are employed for the test problem 2. The inbuilt *stabilization* parameter value in IFISS is used for setting up the matrix block C in K for both the test problems.

3.5.1 Test Problem 1

The Stokes PDE (3.2) is defined on a square domain $D = (-1, 1) \times (-1, 1)$ with Dirichlet boundary condition specified everywhere on the boundary. This (enclosed flow) problem [Elman et al., 2014a, p. 126] can be generated by choosing example `4` when running the driver `stokes_testproblem` in IFISS. On a given grid, the 'true' algebraic solution $\mathbf{x}$ is obtained from (block ideal/block AMG) preconditioned MINRES with a tight relative residual $\frac{\|\mathbf{r}^{(k)}\|_{M^{-1}}}{\|\mathbf{r}^{(0)}\|_{M^{-1}}}$ reduction tolerance of `1e-14`. From $\mathbf{x}$, the 'exact' a posteriori error estimate η is computed. The starting vector $\mathbf{x}^{(0)}$ is

Table 3.3: MINRES iteration counts and errors along with extremal Ritz values and interior most harmonic Ritz values for block ideal preconditioning on uniform grids for Stokes test problem 1.

h	k_{tol1}	k_{tol2}	k^*	e_η^*	$\theta_{\min}^{k_-^*}$	$\theta_{\max}^{k_-^*}$	$\theta_{\min}^{k_+^*}$	$\theta_{\max}^{k_+^*}$	#dof
1/16	33	48	15	1.3e-2	-1.2994	-0.2911	1.000	1.6152	3202
1/32	33	48	24	5.3e-4	-1.3173	-0.1949	1.000	1.6170	12546
1/64	33	50	27	1.2e-4	-1.3184	-0.1841	1.000	1.6175	49666
1/128	33	50	30	2.8e-5	-1.3192	-0.1781	1.000	1.6177	197634

generated using the MATLAB function `rand`. Also, let $\eta^{(k^*)}$ denote the a posteriori error estimate at the optimal stopping iteration k^* and $e_\eta^* := |\eta - \eta^{(k^*)}|$. These values are tabulated in Tables 3.3 and 3.4 for block ideal and block AMG preconditioner respectively on various grids.

Table 3.4: MINRES iteration counts and errors along with extremal Ritz values and interior most harmonic Ritz values for block AMG preconditioning on uniform grids for Stokes test problem 1.

h	k_{tol1}	k_{tol2}	k^*	e_η^*	$\theta_{\min}^{k_-^*}$	$\theta_{\max}^{k_-^*}$	$\theta_{\min}^{k_+^*}$	$\theta_{\max}^{k_+^*}$	#dof
1/16	37	54	18	1.1e-5	-1.3010	-0.2815	0.8676	1.5989	3202
1/32	39	55	27	5.2e-6	-1.3069	-0.2017	0.8375	1.6093	12546
1/64	41	58	31	8.4e-7	-1.3088	-0.1816	0.8159	1.6119	49666
1/128	41	58	35	1.7e-6	-1.3095	-0.1756	0.8070	1.6134	197634

The e_η^* columns show that $\{\eta^{(k)}\}$ has converged with a good accuracy to the true a posteriori error estimate η at the balanced stopping iteration. The effectiveness of the balanced stopping test can be gauged by comparing the iteration counts k^* needed to satisfy the balanced stopping test with the iteration counts $k_{\text{tol1}}, k_{\text{tol2}}$ needed to satisfy a fixed relative residual $\dfrac{\|\mathbf{r}^{(k)}\|_{M^{-1}}}{\|\mathbf{r}^{(0)}\|_{M^{-1}}}$ reduction tolerance of `1e-6` (which is the default tolerance in MATLAB solvers) and `1e-9` respectively. In the absence of a balanced stopping test, these are realistic choices for algebraic error tolerance. It is unlikely that the user will know the stopping point k^* a priori and is likely to provide a tighter tolerance than actually required. This would result in needless computations. A quick glance at the columns for optimal iteration counts k^* and those of k_{tol2} shows that a significant number of iterations is wasted (without decreasing the approximation error) by not using the balanced stopping test. Typically, employing the balanced stopping test (3.37) or (3.38) would result in significant savings in computational

work of the solver, especially if one were to solve the underlying PDE adaptively using FEM. These computational savings are further significant in light of huge size of some of these linear systems; see the last (#dof) column in Tables 3.3 and 3.4.

The stopping tests (3.47) and (3.45) suggest that the relevant eigenvalues involved in the stopping tests (3.37) and (3.38) are independent of the discretization parameters. Indeed this is the case, which can be seen from the column entries at balanced stopping iteration for extremal Ritz values $\theta^{k^*_-}_{\min}, \theta^{k^*_+}_{\max}$ and interior most harmonic Ritz values $\theta^{k^*_-}_{\max}, \theta^{k^*_+}_{\min}$ estimates of the corresponding eigenvalues of the discrete system. Also, a comparison of the corresponding eigenvalue (Ritz and harmonic Ritz) estimates for block AMG and block ideal preconditioners shows that block AMG approximates the block ideal preconditioner quite closely.

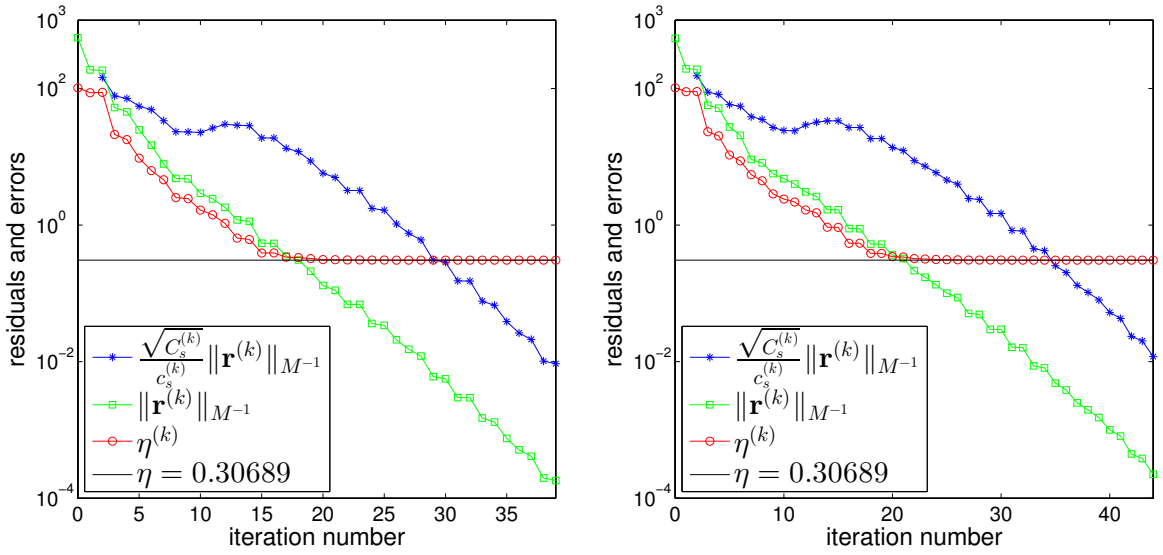

Figure 3.2: Errors vs iteration number for block ideal (left) and block AMG (right) preconditioned MINRES on a uniform grid $h = 1/128$ for Stokes test problem 1.

Further insight into the intricacies associated with applying the stronger stopping test (3.37) is provided by Figures 3.2, 3.3, and 3.4 for both block ideal and block AMG preconditioning on a uniform grid with $h = \dfrac{1}{128}$. On both plots of Figure 3.2, note that at the optimal stopping iteration k^*—the iteration where the red curve for $\eta^{(k)}$ is first above the blue curve for $\|\mathbf{r}^{(k)}\|_{M^{-1}}$—$\{\eta^{(k)}\}$ has converged with some accuracy to the exact a posteriori error estimate η. The convergence is further illustrated by continuing for 9 more iterations after balanced stopping where the red curve for $\eta^{(k)}$ always 'stays' on the black line for η. Note that on these plots $C_s^{(k)} := \max\{|\theta^{k_+}_{\max}|^2, |\theta^{k_-}_{\min}|^2\}$ and $c_s^{(k)} := \min\{|\theta^{k_-}_{\max}|^2, |\theta^{k_+}_{\min}|^2\}$.

The convergence of extremal Ritz values and interior most harmonic Ritz values at the balanced stopping iteration to the corresponding eigenvalues of the discrete problem can be seen from Figures 3.3 and 3.4 respectively. The actual extremal and interior most eigenvalues of the preconditioned (block ideal and block AMG) matrix on these plots are estimated as the corresponding Ritz and harmonic Ritz values respectively. Preconditioned MINRES is run 'long enough' here to ensure that these estimates have 'converged' (this was ascertained by looking at the values of these estimates). Note that the data plotted in Figures 3.3 and 3.4 corresponds to the entries in the last row for block ideal and block AMG preconditioner respectively in Tables 3.3 and 3.4. Also, the plots continue for 9 more iterations after balanced stopping to illustrate that the converged extremal Ritz values and interior most harmonic Ritz values stay convergent to the corresponding discrete system eigenvalues.[7]

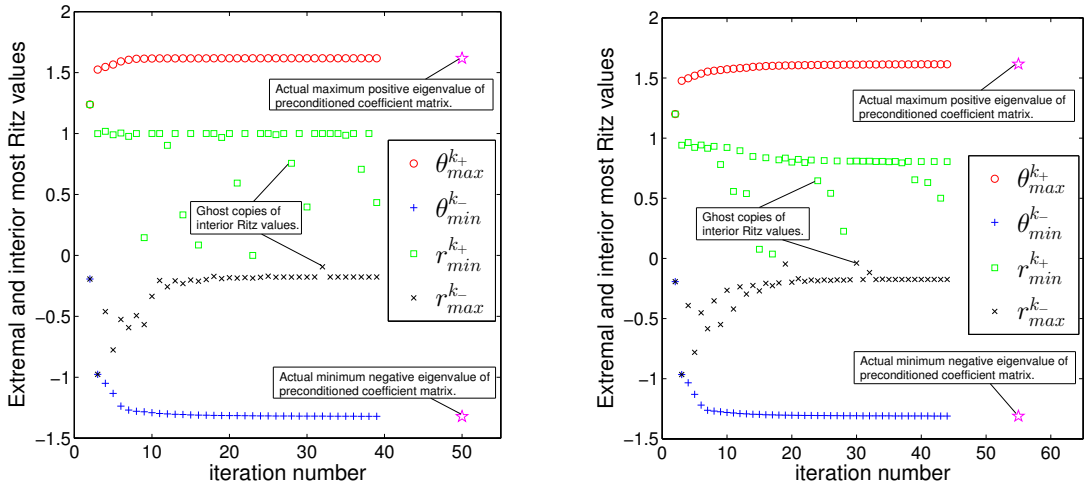

Figure 3.3: Computed Ritz values for block ideal (left) and block AMG (right) MINRES on a uniform grid $h = 1/128$ for Stokes test problem 1.

The Ritz value plots in Figure 3.3 further suggest that there are no *ghost (spurious copies)* of extremal Ritz values. The same is suggested for interior most harmonic Ritz values in Figure 3.4. In contrast there are ghost Ritz values for interior most Ritz values $r_{\max}^{k-}$ (the maximum negative Ritz value at the kth step) and $r_{\min}^{k+}$ (the minimum positive Ritz value at the kth step); see Figure 3.3. This is also the case for the extremal harmonic Ritz values $h_{\max}^{k+}$ (the maximum positive harmonic Ritz value

[7]Lanczos method can lose orthogonalization after convergence of Ritz vectors and hence these estimates might not remain converged. But implementing the Lanczos procedure in MINRES with reorthogonalization solves this issue.

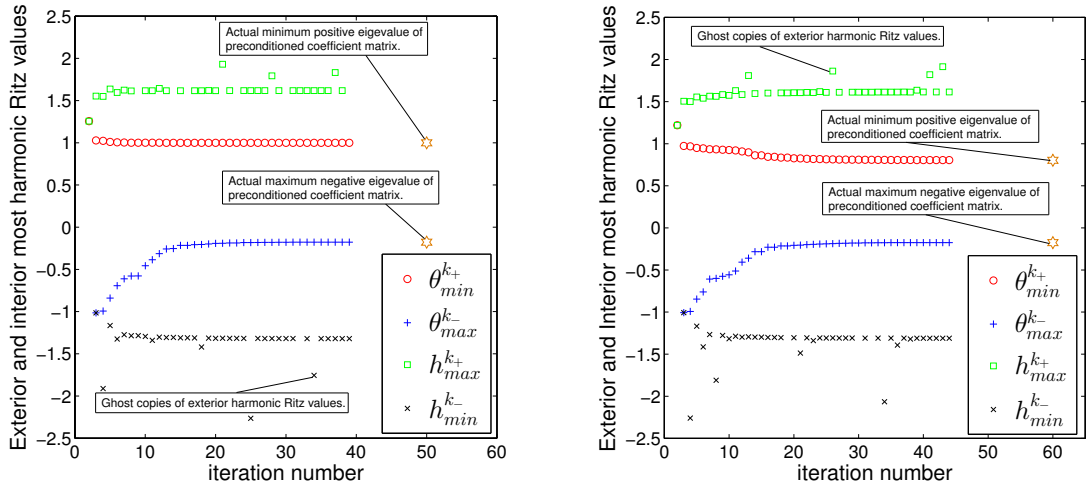

Figure 3.4: Computed harmonic Ritz values for block ideal (left) and block AMG (right) MINRES on a uniform grid $h = 1/128$ for Stokes test problem 1.

at the kth step) and $h_{\min}^{k_-}$ (the minimum negative harmonic Ritz value at the kth step); see Figure 3.4. Thus, $\theta_{\max}^{+}$ and $\theta_{\min}^{-}$ should be estimated by the corresponding extremal Ritz values while $\theta_{\min}^{+}$ and $\theta_{\max}^{-}$ should be estimated by the corresponding interior most harmonic Ritz values. This is consistent with the discussion in section 3.4.6.

The discrete inf-sup constant can also be estimated on-the-fly as suggested in the work of [Silvester and Simoncini, 2011]. It follows from Theorem 3.4.1 that if the bounds in (3.40) are tight then

$$\gamma^2 = \frac{(\theta_{\max}^{-})^2 - \theta_{\max}^{-}\theta_{\min}^{+}}{\theta_{\min}^{+}}. \tag{3.48}$$

In light of the Lanczos estimates for the extremal and interior most eigenvalues of the preconditioned matrix, (3.48) can be rewritten as

$$(\gamma^{(k)})^2 = \frac{(\theta_{\max}^{k_-})^2 - \theta_{\max}^{k_-}\theta_{\min}^{k_+}}{\theta_{\min}^{k_+}}. \tag{3.49}$$

Thus, the balanced stopping strategy also provides a cheap estimate for γ on-the-fly; see Figure 3.5. The 'true' γ in Figure 3.5 is computed by running (block ideal and block AMG) preconditioned MINRES 'long enough' to ensure convergence (from inspection of the estimate values).

A closer examination of Figure 3.2 shows that $\{\eta^{(k)}\}$ has converged to η much before balanced stopping iteration on each plot. In fact if one were to apply the weaker stopping test (3.38) then this is the iteration at which one would stop optimally. However, there is always the pitfall of premature stopping due to nonconvergence of

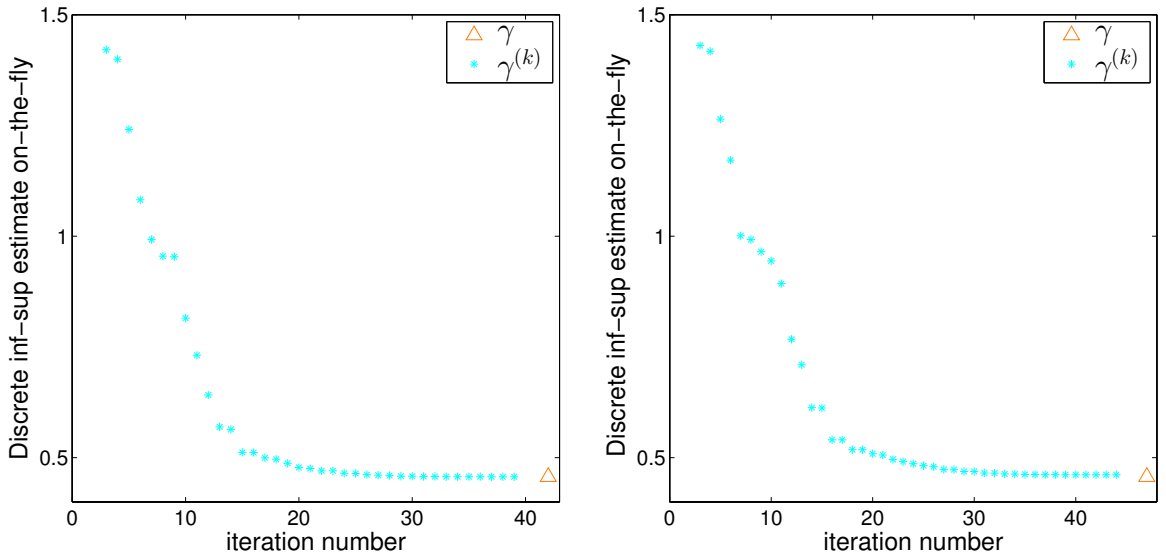

Figure 3.5: Computed discrete inf-sup constant for block ideal (left) and block AMG (right) preconditioned MINRES on a uniform grid $h = 1/128$ for Stokes test problem 1.

the interior most harmonic Ritz values. A way to overcome this issue of premature stopping has been discussed in section 3.4.7. To reiterate, the results are presented here for the stronger stopping test (3.37) only to illustrate the nuances associated with optimal stopping for symmetric indefinite systems. In general, the weaker stopping test should be used in practice.

Note that the plots in Figure 3.2 merit a further investigation in devising an optimal balanced black-box stopping test that is independent of the extremal and interior most eigenvalues of the preconditioned matrix. If $\sqrt{\min\{|\theta^-_{\max}|^2, |\theta^+_{\min}|^2\}} \geq 1$, then

$$\frac{1}{\sqrt{\min\{|\theta^-_{\max}|^2, |\theta^+_{\min}|^2\}}} \, \|\mathbf{r}^{(k)}\|_{M^{-1}} \leq \|\mathbf{r}^{(k)}\|_{M^{-1}}. \tag{3.50}$$

The weaker stopping test (3.38) in light of (3.50) can be transformed into the following. Stop at the first iteration k^* such that

$$\|\mathbf{r}^{(k^*)}\|_{M^{-1}} \leq \eta^{(k^*)}. \tag{3.51}$$

However, application of the stopping test (3.51) depends on the assumption that $\sqrt{\min\{|\theta^-_{\max}|^2, |\theta^+_{\min}|^2\}} \geq 1$, which is not always true; see the corresponding entries for $\theta^+_{\min}$ and $\theta^-_{\max}$ in Tables 3.3 and 3.4.

3.5.2 Test Problem 2

The Stokes PDE (3.2) is defined on a L-shaped ('flow over a backward-facing step')

domain $D = (-1, 5) \times (-1, 1) \setminus (-1, 0] \times (-1, 0]$. Poiseuille flow profile is imposed on the inflow boundary ($x_1 = -1, 0 \leq x_2 \leq 1$) for $\vec{x} = (x_1, x_2) \in D$, and zero velocity condition is imposed on the walls. Neumann boundary conditions are defined everywhere on the outflow boundary ($x_1 = 5, -1 < x_2 < 1$) [Elman et al., 2014a, p. 124]. This problem can be generated IFISS by choosing example `2` when running the driver `stokes_testproblem`.

Table 3.5: MINRES iteration counts and errors along with extremal Ritz values and interior most harmonic Ritz values for block ideal preconditioning on $2^l \times (2^l \times 3)$ grids for Stokes test problem 2.

l	k_{tol1}	k_{tol2}	k^*	e^*_η	$\theta^{k^*_-}_{\min}$	$\theta^{k^*_-}_{\max}$	$\theta^{k^*_+}_{\min}$	$\theta^{k^*_+}_{\max}$	#dof
4	53	73	51	6.2e-6	-1.3632	-0.0242	1.000	1.7909	2242
5	55	73	54	3.5e-6	-1.3638	-0.0242	1.000	1.8109	8706
6	53	76	58	1.4e-6	-1.3669	-0.0242	1.000	1.8184	34306
7	53	77	61	3.1e-7	-1.3671	-0.0241	1.000	1.8214	136194

Table 3.6: MINRES iteration counts and errors along with extremal Ritz values and interior most harmonic Ritz values for block AMG preconditioning on $2^l \times (2^l \times 3)$ grids for Stokes test problem 2.

l	k_{tol1}	k_{tol2}	k^*	e^*_η	$\theta^{k^*_-}_{\min}$	$\theta^{k^*_-}_{\max}$	$\theta^{k^*_+}_{\min}$	$\theta^{k^*_+}_{\max}$	#dof
4	59	80	55	1.1e-5	-1.3540	-0.0241	0.8025	1.7191	2242
5	63	84	61	5.2e-6	-1.3571	-0.0241	0.7865	1.7294	8706
6	63	86	65	8.4e-7	-1.3576	-0.0240	0.7606	1.7334	34306
7	63	88	69	1.7e-6	-1.3577	-0.0241	0.7290	1.7361	136194

Results are tabulated for block ideal and block AMG preconditioned MINRES in Tables 3.5 and 3.6 for this test problem on various $2^l \times (2^l \times 3)$ grids. The quantities in these tables are defined exactly in the same way as for the test problem 1.[8]

The insights from the results here is essentially similar to those for test problem 1. As compared to the test problem 1, the slower convergence is due to a singularity in the problem near the 'step' which is reflected in the largest negative eigenvalue estimate (see the $\theta^{k^*_-}_{\max}$ column in Tables 3.5 and 3.6) of the preconditioned matrix, which is more closer to zero than $\theta^{k^*_-}_{\max}$ of test problem 1 (where there was no singularity in the problem).

[8]However, here the 'true' algebraic solution `x` is obtained from preconditioned MINRES with a tight relative residual reduction tolerance of `1e-12` instead of `1e-14` since (preconditioned) MINRES gives a warning that latter 'input tolerance may not be achievable by MINRES' on some grids.

From Figure 3.6 note that it is possible that the curve of $\eta^{(k)}$ may fall below the line of true a posteriori estimate η. However, as the iteration proceeds, ultimately the sequence $\{\eta^{(k)}\}$ converges with some accuracy to η.

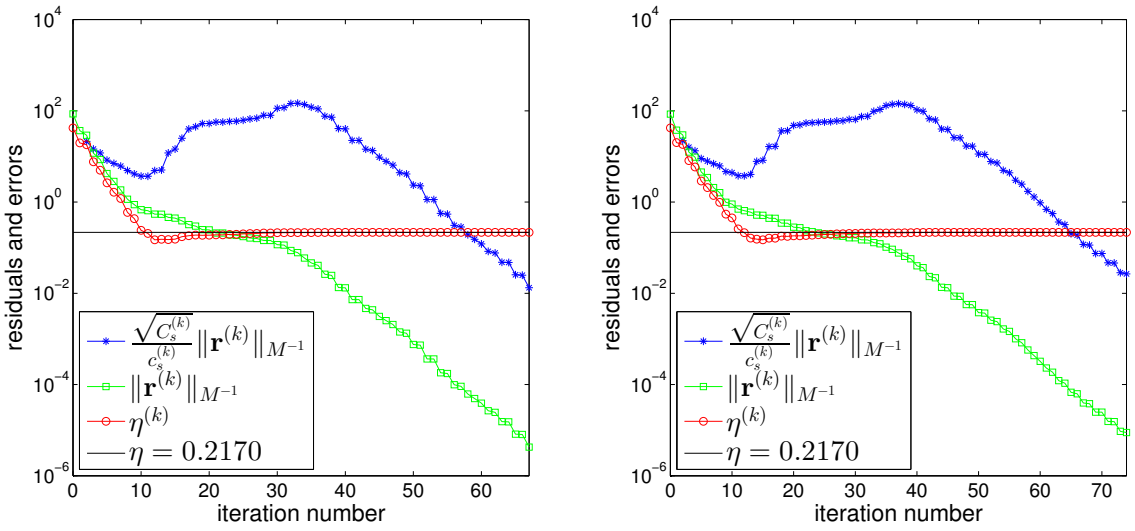

Figure 3.6: Errors vs iteration number for block ideal (left) and block AMG (right) preconditioned MINRES on a 128×384 grid for Stokes test problem 2.

The results from both the test problems illustrate that employing an optimal balanced black-box stopping strategy not only avoids unnecessary computations but also rules out premature stopping of the preconditioned MINRES solver.

3.6 Summary

An optimal balanced black-box stopping test is devised in this chapter in MINRES with preconditioning for solving (saddle point) symmetric indefinite linear systems arising from FEM discretization of an underlying PDE (Stokes equations in particular). The constants in the balanced stopping test are estimated cheaply on-the-fly. This is achieved by exploiting the relationship between Ritz, harmonic Ritz values (obtained from the Lanczos process in preconditioned MINRES) and the relevant eigenvalues of the preconditioned matrix involved in the balanced stopping test. Typically, employing such a balanced stopping strategy would avoid premature stopping and generally leads to huge computational savings. The stopping strategy presented here has extended the work done in this direction by [Silvester and Simoncini, 2011]. In particular, the methodology presented here for deriving the balanced stopping test is different from

that in [Silvester and Simoncini, 2011]. Also, the constant involved their stopping test has been 'improved' in the sense that one can now stop optimally a few iterations earlier than using their stopping test.

Chapter 4

Balanced iterative stopping for nonsymmetric systems I

Publication

- The material presented in this chapter will soon be submitted for publication.
- The devised balanced stopping test in GMRES, BICGSTAB(ℓ), and TFQMR solvers for solving nonsymmetric linear systems arising from FEM approximation of (parametric) convection-diffusion equations has resulted in the functions: `CD_GMRES`, `CD_BICGSTAB(`ℓ`)`, and `CD_TFQMR` in the software IFISS [Elman et al., 2014b].

In the previous chapters an optimal balanced black-box stopping test was devised in MINRES for solving symmetric positive-definite and symmetric indefinite linear systems. An optimal balanced black-box stopping test in GMRES [Saad and Schultz, 1986] for solving nonsymmetric linear systems arising from FEM approximation of a PDE is presented here. The same balanced stopping test is applied to optimally stop BICG [Fletcher, 1976], its variants BICGSTAB(ℓ) [Sleijpen and Fokkema, 1993], and TFQMR [Freund, 1993] for solving nonsymmetric systems. The devised balanced stopping test in BICGSTAB(ℓ) and TFQMR resolves the issue of stopping optimally for these solvers. These iterative methods do not satisfy any optimality conditions, that is, there is no algebraic quantity that is always guaranteed to decrease monotonically with respect to iteration count [Barret et al., 1987, section 2.3.8]. Hence, stopping optimally is always an issue associated with these solvers.

Balanced stopping criterion for nonsymmetric linear systems arising from FEM discretization of a PDE have been studied in detail in [Arioli et al., 2005] and [Wu, 2003, chapter 5]. In [Arioli et al., 2005] the stopping criterion is based on a priori approximation error bounds, while in [Wu, 2003, chapter 5] although the stopping test is based on a posteriori approximation error estimators, it involves user-defined parameters for stopping. The contribution of the present work to the existing literature is that it states the precise constants (and hence it is an optimal balanced black-box stopping test) involved in the balanced stopping test. Computing these constants for huge linear systems can be expensive. In order to compute these constants *on-the-fly*, a balanced stopping test based on MINRES for solving the corresponding discrete normal equations is also proposed in section 4.5.

In chapters 2 and 3, it was observed that a 'tight' a posteriori error estimator is required for devising a balanced stopping test. Unlike stochastic diffusion equations, tight a posteriori error estimators for parameterized convection-diffusion equations have not yet been devised. Thus, stochastic collocation together with 'appropriate'[1] FEM discretization is employed to discretize this PDE. This results in solving a finite number of linear systems. Devising a balanced stopping test for all these nonsymmetric linear systems is no more or less general than devising a balanced stopping test for

[1] Appropriate here implies that the discrete linear system arising from FEM discretization should be nonsymmetric, thereby mirroring the non self-adjoint nature of the continuous convection-diffusion operator.

a single nonsymmetric linear system arising from FEM discretization of deterministic convection-diffusion equations.

This chapter consists of 6 sections. The target linear algebra problem is set up in section 4.1 and the natural norm for measuring the errors is identified. An overview of preconditioned GMRES, BICGSTAB(ℓ), and TFQMR is presented in section 4.2. The balanced stopping test is developed in section 4.3. In section 4.4 a set of computational results that can be reproduced using the IFISS toolbox [Elman et al., 2014b] are presented. These results will confirm the effectiveness of the balanced stopping strategy. A theoretical analysis towards cheap computation of the constants involved in the devised stopping test is considered in section 4.5. A summary of the chapter is presented in section 4.6.

4.1 Deterministic convection-diffusion equations

Convection-diffusion equations are used for modelling various phenomena in physical, biological, and engineering sciences such as the transfer and diffusion of pollutants, drift-diffusion equations in semi-conductor physics, temperature of a fluid moving along a heated wall etc.; see [Eriksson et al., 1996, chapter 18].

To quote [Elman et al., 2014a, p. 282] "The discrete convection-diffusion equation has been used as an "archetypical" nonsymmetric system, in much the same way as the discrete Poisson equation is regarded as the "definitive" symmetric positive-definite system." Thus, the discrete linear systems arising from FEM discretization of convection-diffusion equations are chosen as representatives for devising a balanced stopping criterion in iterative solvers for nonsymmetric linear systems.

Following the notation in [Elman et al., 2014a, p. 234], the steady-state scalar convection-diffusion solution $u(\vec{x}) : D \to \mathbb{R}$ satisfies

$$-\nabla \cdot \epsilon(\vec{x})\nabla u(\vec{x}) + \vec{w}(\vec{x}) \cdot \nabla u(\vec{x}) = f(\vec{x}), \quad \forall\, \vec{x} \in D \subset \mathbb{R}^d\, (d = 2,\, 3), \tag{4.1a}$$

$$u(\vec{x}) = g_D(\vec{x}), \quad \forall\, \vec{x} \in \partial D_D, \tag{4.1b}$$

$$\nabla u(\vec{x}) \cdot \vec{n} = g_N(\vec{x}), \quad \forall\, \vec{x} \in \partial D_N = \partial D \setminus \partial D_D. \tag{4.1c}$$

Here D is the spatial domain, $\vec{w}$ denotes the wind, and $\epsilon := \kappa I$ is the isotropic permeability tensor, $\kappa : D \to \mathbb{R}$. The quantities f, g_D, g_N are given functions and $\vec{n}$

denotes the normal to boundary ∂D, which is the union of the Dirichlet (∂D_D) and the Neumann (∂D_N) spatial boundary.

For the simplicity of exposition, the diffusion coefficient $\epsilon > 0$ will be assumed to be independent of the spatial coordinates. Also, it will be assumed that $\nabla \cdot \vec{w} = 0$.

The presence of the convection term in (4.1) often results in the formation of layers (exponential boundary layers/ shear layers) in the solution. Therefore, the FEM approximation has to be constructed intelligently taking into account the relative contributions of convection and diffusion in (4.1). The Peclet number $\mathcal{P}$ encapsulates these contributions globally and is defined as

$$\mathcal{P} := \frac{|\vec{w}|\mathrm{L}}{\epsilon}, \tag{4.2}$$

where L denotes a characteristic length scale for D and $|\cdot|$ is some appropriate measure.

If $\mathcal{P} \leq 1$, (4.1) is diffusion dominated otherwise convection is more relevant. In fact the rth mesh element Peclet number

$$\mathcal{P}_h^r := \frac{|\vec{w_r}|h_r}{2\epsilon}, \qquad h_r - r\text{th element mesh parameter}, \;\; w_r - r\text{th element wind}, \tag{4.3}$$

might be a better indicator of the relative contributions of convection and diffusion locally in the FEM mesh.

4.1.1 Weak formulation

The variational formulation of (4.1) is to find $u \in H_E^1$ such that

$$a(u, v) = l(v), \qquad \forall\, v \in H_{E_0}^1, \tag{4.4}$$

where

$$\begin{aligned} a(u,v) &:= \epsilon \int_D (\nabla u \cdot \nabla v) + \int_D (\vec{w} \cdot \nabla u) v, \\ l(v) &:= \int_D f v + \epsilon \int_{\partial D_N} g_N v, \\ H_E^1 &:= \{v \in H^1(D) \,|\, v = g_D \text{ on } \partial D_D\}, \\ H_{E_0}^1 &:= \{v \in H^1(D) \,|\, v = 0 \text{ on } \partial D_D\}. \end{aligned}$$

Notice that the bilinear form $a(\cdot, \cdot)$ is the sum of a self-adjoint operator (diffusion operator) and an almost skew self-adjoint convection operator.[2] It can be easily verified

[2]The convection operator is completely skew self-adjoint provided Neumann boundary conditions are absent or they exist only at the characteristic boundary [Elman et al., 2014a, p. 241].

that $a(\cdot,\cdot)$ and $l(\cdot)$ satisfy the conditions [Elman et al., 2014a, p. 243] of the Lax–Milgram theorem. Thus, there exists a unique solution to the weak formulation (4.4). The Galerkin FEM formulation of (4.4) is presented next and also the *natural* norm for measuring errors will be identified.

4.1.2 Galerkin FEM formulation

Find $u_h \in S_E^h$ such that

$$a(u_h, v_h) \;=\; l(v_h), \qquad \forall\, v_h \in S_0^h, \tag{4.5}$$

where S_E^h and S_0^h are finite dimensional subspaces of H_E^1 and $H_{E_0}^1$ respectively.

[Wathen, 2007] advocates that a natural norm for a function u in the Sobolev space $H_{E_0}^1$ is the L^2 norm of its gradient, that is, $\|\nabla u\|_2$. However, this need not be the only meaningful norm for measuring errors associated with (4.5). An alternative norm known as the *streamline diffusion norm* is also discussed in [Elman et al., 2014a, p. 252]. This norm arises when the streamline diffusion method introduced by [Hughes and Brooks, 1979] is used for overcoming the drawbacks associated with the Galerkin discretization.[3] This leads to a slightly different FEM formulation to (4.5).

4.1.3 Streamline diffusion FEM formulation

Find $u_h \in S_E^h$ such that

$$a_{\mathrm{sd}}(u_h, v_h) \;=\; l_{\mathrm{sd}}(v_h), \qquad \forall\, v_h \in S_0^h, \tag{4.6}$$

where if δ denotes the stabilization parameter [Elman et al., 2014a, p. 253] and if $g_N = 0$, then

$$\begin{aligned} a_{\mathrm{sd}}(u_h, v_h) \;&:=\; \epsilon \int_D (\nabla u_h \cdot \nabla v_h) \;+\; \int_D (\vec{w} \cdot \nabla u_h) v_h \;+\; \delta \int_D (\vec{w} \cdot \nabla u_h)\,(\vec{w} \cdot \nabla v_h) \\ &\quad -\; \delta\epsilon \sum_k \int_{\triangle_k} (\nabla^2 u_h)\,(\vec{w} \cdot \nabla v_h),^{4} \\ l_{\mathrm{sd}}(v_h) \;&:=\; \int_D f v_h \;+\; \delta \int_D f(\vec{w} \cdot \nabla v_h). \end{aligned}$$

[3] Galerkin approximation for (4.1) is inaccurate if the mesh is not fine enough to resolve the layers in the solution and these inaccuracies may also propagate and pollute the approximated solution in regions where the exact solution is well behaved. An alternative way to handle boundary layers is by using Shishkin grids; see [Shishkin, 1992].

The corresponding streamline diffusion norm is

$$\|u_h\|_{\text{sd}} := (\epsilon \|\nabla u_h\|_2^2 + \delta \|\vec{w} \cdot \nabla u_h\|_2^2)^{1/2}. \tag{4.7}$$

For convection dominated problems, that is, for large Peclet numbers (small ϵ), the solution u_h is dominated by its behaviour along the streamlines, and hence $\|u_h\|_{\text{sd}}$ which involves the streamline derivative $\|\vec{w} \cdot \nabla u_h\|_2$ is a more meaningful measure than $\|\nabla u_h\|_2$.

The streamline diffusion method is closely related to the methodology of [Brezzi et al., 1998], which employs interior finite element basis functions to gather information interior to elements. This enhances the quality of the corresponding discrete solution. For a more detailed discussion; see [Elman et al., 2014a, p. 247 ff.].

The IFISS toolbox employs streamline diffusion stabilization for solving (4.1), but measures errors in the L^2 norm of the gradient. The balanced stopping test will be based on this norm. However, the stopping methodology can easily be modified to cater to the streamline diffusion norm.

Having formulated the streamline diffusion FEM formulation, the target linear system is set up in the next subsection.

4.1.4 Matrix formulation

Let $\{\phi_i\}_{i=1}^n$ be a basis for S_0^h. The basis functions could be the piecewise linear ($\boldsymbol{P}_1$) finite elements, or the piecewise bilinear ($\boldsymbol{Q}_1$) finite elements etc. By augmenting this basis with $\{\phi_i\}_{j=n+1}^{n+n_\partial}$, an arbitrary $u_h \in S_E^h$ can be expressed (loosely defined) as a basis[5] expansion

$$u_h = \sum_{j=1}^{n+n_\partial} x_j \phi_j.$$

The function $\sum_{j=n+1}^{n+n_\partial} x_j \phi_j$ interpolates the boundary data g_D on ∂D_D. Plugging this expansion in (4.6), and enforcing the condition $v_h = \phi_i$, the discrete problem is to find $\{x_j\}_{j=1}^n$ such that [Elman et al., 2014a, p. 272]

$$\sum_{j=1}^{n} a_{\text{sd}}(\phi_j, \phi_i) x_j = \hat{l}_{\text{sd}}(\phi_i), \qquad i = 1, 2, \ldots, n,$$

[4]The elementwise sum makes sense if it is assumed that the functions in S_0^h lie in $H^2(\triangle_k)$ when restricted to $\triangle_k$.

[5]The space H_E^1 is not a vector space unless $g_D = 0$.

where

$$\begin{aligned}\hat{l}_{\text{sd}}(\phi_i) := & \int_D f\phi_i + \int_{\partial D_N} g_N\phi_i - \sum_{j=n+1}^{n+n_\partial}\left(\epsilon\int_D(\nabla\phi_j\cdot\nabla\phi_i) + \int_D(\vec{w}\cdot\nabla\phi_j)\phi_i\right)x_j \\ & - \sum_{j=n+1}^{n+n_\partial}\left(\sum_{k'}\delta_{k'}\int_{\triangle_{k'}}(\vec{w}\cdot\nabla\phi_j)(\vec{w}\cdot\nabla\phi_i)\right)x_j.\end{aligned} \tag{4.8}$$

The second sum in (4.8) is over those elements $\triangle_{k'}$ that have boundaries that intersect with the Dirichlet boundary ∂D_D. In matrix form, one needs to find $\mathbf{x}\in\mathbb{R}^n$ such that

$$F\mathbf{x} = \mathbf{b} \qquad \Longleftrightarrow M^{-1}F\mathbf{x} = M^{-1}\mathbf{b},$$

where

$$\begin{aligned} F &= [f_{ij}]\in\mathbb{R}^{n\times n}, & f_{ij} &:= a_{\text{sd}}(\phi_j,\phi_i), \\ \mathbf{b} &= [b_i]\in\mathbb{R}^n, & b_i &:= \hat{l}_{\text{sd}}(\phi_i), \\ & M \text{ is a preconditioner.} \end{aligned} \tag{4.9}$$

When lower order (piecewise linear or piecewise bilinear) finite elements along with stabilization are employed, the coefficient matrix F has the form

$$F = \epsilon A + N + S,$$

where

$$\begin{aligned} A &= [a_{ij}]\in\mathbb{R}^{n\times n}, & a_{ij} &:= \int_D \nabla\phi_j\cdot\nabla\phi_i, \\ N &= [n_{ij}]\in\mathbb{R}^{n\times n}, & n_{ij} &:= \int_D(\vec{w}\cdot\nabla\phi_j)\phi_i, \\ S &= [s_{ij}]\in\mathbb{R}^{n\times n}, & s_{ij} &:= \sum_k \delta_k\int_{\triangle_k}(\vec{w}\cdot\nabla\phi_j)(\vec{w}\cdot\nabla\phi_i). \end{aligned} \tag{4.10}$$

For FEM discretization without streamline diffusion stabilization, $F=\epsilon A+N$ for finite elements of any order.

The matrices under consideration are quite structured. The matrix A is symmetric and positive-definite provided Dirichlet boundary conditions exist over an interval ($\int_{\partial D_D}\neq 0$), however small. The stabilization matrix S is symmetric and positive-semidefinite. The matrix N is a skew-symmetric matrix [Elman et al., 2014a, p. 241, pp. 271–272]. Thus, F is a nonsymmetric matrix that will be assumed to be invertible throughout this chapter. Iterative solvers like GMRES, BICGSTAB(ℓ), and TFQMR are popular for solving nonsymmetric linear systems. A discussion on these methods is presented in the following section.

4.2 Fast Krylov solvers for nonsymmetric systems

Krylov methods for solving nonsymmetric linear systems can essentially be classified into two categories. These solvers either satisfy some optimality condition or they are inexpensive in the sense that the number of arithmetic operations does not depend on the iteration count, that is, there is only a fixed amount of work per iteration. The latter category Krylov methods are characterized for having short-term recurrences of fixed length. The GMRES method falls in the optimality satisfying category while BICG, BICGSTAB(ℓ), and TFQMR methods come under the suboptimal class. [Faber and Manteuffel, 1984, 1987] have shown that for general nonsymmetric linear systems, there is no Krylov subspace method which is both inexpensive and optimal in some sense. However, a common feature of all these iterative methods is their connection to the Lanczos [Lanczos, 1950] or the Arnoldi [Arnoldi, 1951] methods for estimating the eigenvalues of matrices. This relation is perhaps because the asymptotic convergence behaviour of iterative methods is dependent on the spectrum of the coefficient matrix.

4.2.1 An overview of GMRES

Suppose that $\mathbf{r}^{(0)} = \mathbf{b} - F\mathbf{x}^{(0)}$ is the initial residual with starting vector $\mathbf{x}^{(0)}$. The kth GMRES iterate $\mathbf{x}^{(k)}$ is in the translated Krylov space

$$\mathbf{x}^{(0)} + \operatorname{span}\{\mathbf{r}^{(0)}, F\mathbf{r}^{(0)}, F^2\mathbf{r}^{(0)}, \ldots, F^{k-1}\mathbf{r}^{(0)}\}.$$

This implies that the kth residual $\mathbf{r}^{(k)}$ lies in the translated Krylov space

$$\mathbf{r}^{(0)} + \operatorname{span}\{F\mathbf{r}^{(0)}, F^2\mathbf{r}^{(0)}, \ldots, F^{k}\mathbf{r}^{(0)}\}. \tag{4.11}$$

Choosing the residual with the minimal Euclidean norm is the defining characteristic of the GMRES method. Also, since Krylov spaces form an ascending chain, $\|\mathbf{r}^{(k)}\|$ is monotonically decreasing.[6] From (4.11), it follows that if Π_k denotes the set of real polynomials of degree less than or equal to k, then $\mathbf{r}^{(k)} = p_k(F)\mathbf{r}^{(0)}$, where $p_k(0) = 1$, $p_k \in \Pi_k$. The polynomial p_k is chosen such that

$$\|\mathbf{r}^{(k)}\| = \min_{p_k \in \Pi_k,\ p_k(0)=1} \|p_k(F)\mathbf{r}^{(0)}\|. \tag{4.12}$$

In order to derive convergence bounds, some additional assumptions are required. Suppose that F is diagonalizable, then $F = VDV^{-1}$ where V is the matrix whose

[6] There might be stagnation of $\|\mathbf{r}^{(k)}\|$ for some iteration counts; see [Meurant, 2014].

columns are the eigenvectors of F and D is the diagonal matrix of eigenvalues λ_j's of F. Also, $p_k(F) = Vp_k(D)V^{-1}$. It follows that $\|\mathbf{r}^{(k)}\| = \min_{p_k \in \Pi_k,\, p_k(0)=1} \|p_k(F)\mathbf{r}^{(0)}\|$ implies

$$\begin{aligned} \min_{p_k \in \Pi_k,\, p_k(0)=1} \|p_k(F)\mathbf{r}^{(0)}\| &= \|Vp_k(D)V^{-1}\mathbf{r}^{(0)}\| \\ &\leq \|V\|\,\|V^{-1}\|\,\|p_k(D)\|\,\|\mathbf{r}^{(0)}\| \\ &= \|V\|\,\|V^{-1}\| \max_{\lambda_j} |p_k(\lambda_j)|\,\|\mathbf{r}^{(0)}\|, \end{aligned} \tag{4.13}$$

where in the penultimate step the subadditive property and the consistency of matrix-vector norms has been utilized. From (4.12) and (4.13), the following convergence bound is obtained [Elman et al., 2014a, p. 289].

$$\frac{\|\mathbf{r}^{(k)}\|}{\|\mathbf{r}^{(0)}\|} \leq \kappa(V) \min_{p_k \in \Pi_k,\, p_k(0)=1} \max_{\lambda_j} |p_k(\lambda_j)|, \tag{4.14}$$

where $\kappa(V)$ denotes the condition number of the matrix V. Equation (4.14) implies that the GMRES convergence is quite fast when eigenvalues of F are clustered away from the origin and F is nearly normal. In practice, F arising from FEM approximation is ill-conditioned. Theoretically, bound (4.14) indicates that preconditioning strategies for (4.9) should focus on clustering the eigenvalues of the preconditioned coefficient matrix away from the origin. Also, the bounds for $\kappa(V)$ and $\min_{p_k \in \Pi_k,\, p_k(0)=1} \max_{\lambda_j} |p_k(\lambda_j)|$ are anything but straightforward to compute [Elman et al., 2014a, p. 290]. Further discussions on GMRES convergence analysis and computable convergence bounds can be found in [Elman et al., 2014a, proposition 7.3, p. 291], [Liesen, 2000], [Liesen and Strakoš, 2005]. Closely associated with GMRES convergence is the notion of the asymptotic convergence factor

$$\rho := \lim_{k \to \infty} \left(\min_{p_k \in \Pi_k,\, p_k(0)=1} \max_{\lambda_j} |p_k(\lambda_j)| \right)^{1/k}. \tag{4.15}$$

This is the factor roughly by which $\|\mathbf{r}^{(k)}\|$ can be expected to decrease for large enough iteration count k. In practice, k is typically not too large. An upper bound on ρ can be found under certain conditions [Elman et al., 2014a, theorem 7.2, p. 291].

It was mentioned earlier that iterative methods and popular methods for estimating eigenvalues are closely related. This relationship between the GMRES and the Arnoldi methods is explored next.

Relation between Arnoldi and GMRES methods

The Arnoldi method computes a unitary similar transform of F, that is,

$$\begin{aligned} FV_k &= V_k H_k + h_{k+1,k}\,[0,\ldots,0,\mathbf{v}^{(k+1)}] = V_{k+1}\hat{H}_k, \\ H_k &= V_k^T F V_k, \qquad \text{if } h_{k+1,k} = 0, \end{aligned}$$

where H_k is an upper-Hessenberg matrix and $\hat{H}_k \in \mathbb{R}^{(k+1)\times k}$ is the matrix H_k with an additional final row $[0,\ldots,0,h_{k+1,k}]$. The columns of the matrix $V_k := [\mathbf{v}^{(1)},\ldots,\mathbf{v}^{(k)}]$ form an orthonormal basis for $K_k(F,\mathbf{v}^{(1)})$. Notice that this method requires all the vectors $\{\mathbf{v}^{(i)}\}_{i=1}^k$ for constructing $\mathbf{v}^{(k+1)}$. This storage requirement may become prohibitive for large k.

The GMRES method seeks the kth iterate $\mathbf{x}^{(k)} \in \mathbf{x}^{(0)} + K_k(F,\mathbf{r}^{(0)})$. Choosing $\mathbf{v}^{(1)} = \frac{\mathbf{r}^{(0)}}{\|\mathbf{r}^{(0)}\|}$ in the Arnoldi method generates an orthonormal basis $\{\mathbf{v}^{(1)},\ldots,\mathbf{v}^{(k)}\}$ for $K_k(F,\mathbf{r}^{(0)})$. Thus,

$$\mathbf{x}^{(k)} = \mathbf{x}^{(0)} + V_k\mathbf{y}^{(k)}, \qquad \text{for some } \mathbf{y}^{(k)} \in \mathbb{R}^k. \tag{4.16}$$

Then the kth residual is

$$\begin{aligned} \mathbf{r}^{(k)} &= \mathbf{b} - F\mathbf{x}^{(k)} \\ &= \mathbf{r}^{(0)} - FV_k\mathbf{y}^{(k)} \\ &= V_{k+1}\left(\|\mathbf{r}^{(0)}\|\mathbf{e}_1 - \hat{H}_k\mathbf{y}^{(k)}\right), \end{aligned} \tag{4.17}$$

where $\mathbf{e}_1 = (1,0,\ldots,0)^T \in \mathbb{R}^{k+1}$. Since V_{k+1} is orthonormal,

$$\|\mathbf{r}^{(k)}\| = \left\|\|\mathbf{r}^{(0)}\|\mathbf{e}_1 - \hat{H}_k\mathbf{y}^{(k)}\right\|. \tag{4.18}$$

The choice of $\mathbf{y}^{(k)}$ that minimizes (in the least-squares sense) the right-hand-side expression in (4.18) gives the residual with the minimal Euclidean norm.

The classic trade-off between optimality and storage of all k previous orthonormal basis vectors (for constructing an orthonormal basis of dimension $k+1$) implies that the work and storage costs of GMRES scale like $O(kn)$. When k is 'large', these storage costs become prohibitive. In order to minimize the storage costs, the GMRES process is restarted and this method is known as restarted GMRES; see [Saad, 2003, chapter 6, section 6.5.5], [Embree, 2003]. Sometimes restarts are done to ensure orthogonality of the basis vectors, which due to accumulation of rounding errors might become

nonorthogonal. A detailed discussion on GMRES can be found in [Saad, 2003, section 6.1–6.5].

For the symmetric positive-definite linear systems, the Arnoldi method reduces to the Lanczos method. This will be the focus of discussion in the next section.

4.2.2 An overview of suboptimal Krylov solvers

When the memory storage requirements are prohibitive, an alternative is to employ suboptimal Krylov methods like BICG, BICGSTAB(ℓ), TFQMR, conjugate gradient squared (CGS) [Sonneveld, 1989], and quasi-minimum residual (QMR) [Freund and Nachtigal, 1991], [Freund and Nachtigal, 1994] that utilize short-term recurrences of fixed length. This implies that the storage requirements and overhead (the work per iteration k) for such methods is independent of the dimension k of the Krylov space. As a result these methods are quite popular; see [Amritkar et al., 2015], [Ahuja et al., 2015]. The downside in employing these methods is the fact that they do not satisfy any optimality condition. Their convergence behaviour is quite irregular and precious little convergence theory is available to explain the erratic behaviour. Indeed, there is no guarantee that they would converge and a breakdown may occur before convergence. These breakdowns are in turn due to the breakdown in the underlying Lanczos biorthogonalization process upon which these methods are constructed; see [Saad, 1982], [Day, 1997].

Lanczos biorthogonalization method

The Lanczos biorthogonalization method for nonsymmetric matrices constructs a pair of orthogonal bases $\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(k)}\}$, $\{\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(k)}\}$ for the Krylov spaces $K_k(F, \mathbf{v}^{(1)})$ and $K_k(F^T, \mathbf{w}^{(1)})$ respectively. The biorthogonal Lanczos algorithm for nonsymmetric matrices [Saad, 2003, section 7.1–7.2] is given as follows. Let $\mathbf{w}^{(0)} = \mathbf{v}^{(0)} = 0$. Choose $\mathbf{v}^{(1)}, \mathbf{w}^{(1)}$ such that $(\mathbf{v}^{(1)})^T \mathbf{w}^{(1)} = 1$. Also, $\beta_1 = \delta_1 = 0$. For $j = 1, 2, \ldots$

$$\begin{aligned} \hat{\mathbf{v}}^{(j+1)} &= F\mathbf{v}^{(j)} - \alpha_j \mathbf{v}^{(j)} - \beta_j \mathbf{v}^{(j-1)}, \\ \hat{\mathbf{w}}^{(j+1)} &= F^T\mathbf{w}^{(j)} - \alpha_j \mathbf{w}^{(j)} - \delta_j \mathbf{w}^{(j-1)}, \end{aligned} \tag{4.19}$$

where $\alpha_j = (\mathbf{w}^{(j)})^T F\mathbf{v}^{(j)}$. The scalars δ_{j+1}, β_{j+1} are chosen such that

$$\frac{\hat{\mathbf{w}}^{(j+1)T}\hat{\mathbf{v}}^{(j+1)}}{\delta_{j+1}\beta_{j+1}} = 1.$$

There are several choices for $\beta_{j+1}, \delta_{j+1}$ that enforce this condition; for more details see [Saad, 2003, section 7.1]. The crucial point to note here is that this condition may be violated if either of the vectors $\hat{\mathbf{v}}^{(j)}, \hat{\mathbf{w}}^{(j)}$ are zero or if they are (nearly) orthogonal. This would lead to breakdown of the Lanczos biorthogonalization algorithm; see [Saad, 2003, section 7.1.2]. If the algorithm is applied for solving a linear system $F\mathbf{x} = \mathbf{b}$ then the breakdown due to zero vector(s) implies that a solution to either $F\mathbf{x} = \mathbf{b}$ or $F^T\mathbf{x} = \mathbf{b}$ has been found. The breakdown due to orthogonality is more problematic since then the algorithm stops without converging. To overcome this breakdown, few 'look-ahead' strategies like QMR etc., have been devised; see [Parlett et al., 1985], [Freund et al., 1993].

Utilization of the Lanczos biorthogonalization algorithm for solving linear systems leads to BICG algorithms and its variants like BICGSTAB(ℓ) etc.

BICG, BICGSTAB(ℓ), and TFQMR

To solve $F\mathbf{x} = \mathbf{b}$, let $\mathbf{r}^{(0)} = \mathbf{b} - F\mathbf{x}^{(0)}$ be the initial residual for any starting vector $\mathbf{x}^{(0)}$. Let $\hat{\mathbf{r}}^{(0)}$ be a nonzero vector which is not orthogonal to $\mathbf{r}^{(0)}$ and choose $\mathbf{v}^{(1)} = \dfrac{\mathbf{r}^{(0)}}{\|\mathbf{r}^{(0)}\|}$, $\mathbf{w}^{(1)} = \dfrac{\hat{\mathbf{r}}^{(0)}}{\|\hat{\mathbf{r}}^{(0)}\|}$. Using the Lanczos biorthogonalization method, BICG constructs a pair of orthogonal bases $\{\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(k)}\}$, $\{\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(k)}\}$ for the Krylov spaces $K_k(F, \mathbf{r}^{(0)})$, $K_k(F^T, \hat{\mathbf{r}}^{(0)})$ respectively. In matrix form [Saad, 2003, section 7.1]

$$\begin{aligned} FV_k &= V_k T_k + \delta_{k+1}[0, \ldots, 0, \mathbf{v}^{(k+1)}] = V_{k+1}\hat{T}_k, \\ W_k^T F V_k &= T_k, \\ W_k^T V_k &= \mathcal{I}, \qquad\qquad \mathcal{I} - \text{Identity matrix}, \end{aligned} \tag{4.20}$$

where $T_k := \underset{1\le j\le k}{\texttt{tridiag}}\,[\delta_j, \alpha_j, \beta_{j+1}]$ is a tridiagonal matrix, $V_k := [\mathbf{v}^{(1)}, \ldots, \mathbf{v}^{(k)}]$ and $W_k := [\mathbf{w}^{(1)}, \ldots, \mathbf{w}^{(k)}]$. The matrix $\hat{T}_k \in \mathbb{R}^{(k+1)\times k}$ is the tridiagonal matrix T_k with an additional final row $[0, \ldots, 0, \delta_{k+1}]$. The iterate $\mathbf{x}^{(k)} \in K_k(F, \mathbf{r}^{(0)})$ can be expressed as

$$\begin{aligned} \mathbf{x}^{(k)} &= \mathbf{x}^{(0)} + V_k\mathbf{y}^{(k)}, \qquad \text{for some } \mathbf{y}^{(k)} \in \mathbb{R}^k, \\ \Longleftrightarrow \mathbf{r}^{(k)} &= \mathbf{r}^{(0)} - FV_k\mathbf{y}^{(k)}. \end{aligned} \tag{4.21}$$

The BICG method seeks the residual $\mathbf{r}^{(k)}$ such that $W_k^T\mathbf{r}^{(k)} = 0$, that is, $\mathbf{r}^{(k)}$ is orthogonal to the auxiliary set of vectors $\{\mathbf{w}^{(j)}\}_{j=1}^k$.

Enforcing the orthogonality condition in (4.21) and applying (4.20) gives

$$T_k \mathbf{y}^{(k)} = W^T \mathbf{v}^{(1)} \|\mathbf{r}^{(0)}\| \quad \Longleftrightarrow \quad T_k \mathbf{y}^{(k)} = \|\mathbf{r}^{(0)}\| \mathbf{e}_1, \tag{4.22}$$

where $\mathbf{e}_1 = (1, \ldots, 0)^T \in \mathbb{R}^k$. The iterate $\mathbf{x}^{(k)}$ is obtained by solving (4.22) for $\mathbf{y}^{(k)}$ and using (4.21).

The suboptimal Krylov solvers [Saad, 2003, section 7.4.2] are quite sensitive to rounding errors. Also, the Euclidean norm of the BICG residuals do not have a global monotonic decrease with iteration count. This is a characteristic of all the suboptimal Krylov methods used for solving nonsymmetric linear systems. However, a 'more' smoother convergence of the residual Euclidean norm is achieved through variants of BICG like BICGSTAB(ℓ), TFQMR etc. These methods differ from BICG in as much that they do not employ the transpose of the coefficient matrix in their computations and incorporate 'look-ahead' strategies to overcome the breakdown of the underlying Lanczos biorthogonalization method. BICGSTAB(ℓ) in particular employs ℓ GMRES steps with each BICG iteration to smooth out to some extent the irregular behaviour of the iteration residual. For a detailed discussion see [Saad, 2003, section 7.3–7.5].

After surveying the state-of-art iterative solvers that can be employed for solving nonsymmetric linear systems, the balanced stopping methodology is presented in the next section.

4.3 A balanced stopping test

The methodology for the devised balanced stopping test in this section follows the flavour of chapters 2 and 3.

4.3.1 Error equation

Let $\mathbf{x} = (x_1, \ldots, x_n)^T$ denote the exact solution of $F\mathbf{x} = \mathbf{b}$ and $\mathbf{e}^{(k)} := \mathbf{x} - \mathbf{x}^{(k)}$ be the iteration error corresponding to the kth iterate $\mathbf{x}^{(k)} = (x_1^k, \ldots, x_n^k)^T$. From the triangle inequality at iteration k

$$\underbrace{\|u - u_h^{(k)}\|_E}_{\text{total error}} \leq \underbrace{\|u - u_h\|_E}_{\text{approximation error}} + \underbrace{\|u_h - u_h^{(k)}\|_E}_{\text{algebraic error}}, \tag{4.23}$$

where u_h is the exact FEM approximation and $u_h^{(k)}$ denotes FEM approximation corresponding to iterate $\mathbf{x}^{(k)}$. Also, u is the exact solution and $\|\cdot\|_E$ denotes the L^2

norm of the gradient. Expressing u_h and $u_h^{(k)}$ in terms of the (interior) basis functions, the algebraic error evaluates to

$$\begin{aligned} \|u_h - u_h^{(k)}\|_E^2 &= \|\nabla(u_h - u_h^{(k)})\|_2^2 \\ &= \int_D \nabla(u_h - u_h^{(k)}) \cdot \nabla(u_h - u_h^{(k)}) \\ &= \int_D \left(\nabla \sum_{j=1}^{n} (x_j - x_j^k)\phi_j \right) \cdot \left(\nabla \sum_{i=1}^{n} (x_i - x_i^k)\phi_i \right) \\ &= (\mathbf{e}^{(k)})^T A \mathbf{e}^{(k)} =: \|\mathbf{e}^{(k)}\|_A^2, \end{aligned} \tag{4.24}$$

where $\mathbf{e}^{(k)} := (x_1 - x_1^k, \ldots, x_n - x_n^k)^T$. The matrix A by construction (see (4.10)) is symmetric positive-definite. Hence, $\|\cdot\|_A$ defines a norm. Note that the natural norm for the continuous problem here translates into the *natural* norm—involving the matrix formed from symmetric positive-definite part of the coefficient matrix—for the corresponding discrete problem. This norm is a popular choice for measuring errors associated with the solution of nonsymmetric linear systems.

Verfürth pioneered the a posteriori error estimation techniques for the convection-diffusion equations [Verfürth, 1998]. For the current discussion, the local problem error estimation strategy estimator developed in [Elman et al., 2014a, pp. 264–265] will be used. That is, one can compute an error estimate $\eta^{(k)}$ that is equivalent to the total error (approximation error at the kth iteration) in the sense that

$$c_1\, \eta^{(k)} \leq \|u - u_h^{(k)}\|_E \leq C_1\, \eta^{(k)}, \quad \text{with } \frac{C_1}{c_1} \sim O(1). \tag{4.25}$$

If a posteriori error estimator η corresponds to the exact FEM approximation u_h then analogous to chapters 2 and 3, assuming the a posteriori error estimates η and $\eta^{(k)}$ to be close estimates of the (exact) approximation error and the total error at the kth iteration step respectively, (4.23) can be rewritten as

$$\eta^{(k)} \simeq \eta + \|\mathbf{e}^{(k)}\|_A, \qquad k = 0, 1, 2, \ldots \tag{4.26}$$

where $\|u_h - u_h^{(k)}\|_E = \|\mathbf{e}^{(k)}\|_A$ from (4.24). The relation $\simeq$ is a direct consequence of (4.25). Balancing the total error and algebraic error, the iteration will be stopped at iteration k^*, which is the smallest value of k such that

$$\|\mathbf{e}^{(k^*)}\|_A \leq \eta^{(k^*)}. \tag{4.27}$$

Since the iteration error $\mathbf{e}^{(k)}$ is difficult to compute, in order to utilize (4.27) it is necessary to obtain bounds on $\|\mathbf{e}^{(k)}\|_A$ which is the focus of the next section.

4.3.2 Tractable bounds on algebraic error

The goal in this section is to obtain upper (C) and lower bounds (c) on $\|\mathbf{e}^{(k)}\|_A$ which will be called the norm-equivalence bounds henceforth. Ideally, $\|\mathbf{e}^{(k)}\|_A$ should be bounded by a readily computable and monotonically decreasing quantity (if any) of the employed iterative solver. The reason being that as the iteration progresses, the accuracy of the discrete solution (and hence the FEM solution) keeps on improving.[7] This will be reflected in the balanced stopping test based on a monotonically decreasing quantity (if any) of the iterative solver. The Euclidean norm of the residual, which is readily computable for all Krylov iterative methods is monotonically decreasing in GMRES. So, bounds on $\|\mathbf{e}^{(k)}\|_A$ in terms of the surrogate norm $\|\mathbf{r}^{(k)}\|$ are obtained here, that is

$$c \leq \frac{\|\mathbf{e}^{(k)}\|_A}{\|\mathbf{r}^{(k)}\|} \leq C. \tag{4.28}$$

Since $\mathbf{e}^{(k)} = F^{-1}\mathbf{r}^{(k)}$, it follows that

$$\|\mathbf{e}^{(k)}\|_A^2 = (\mathbf{r}^{(k)})^T F^{-T} A F^{-1} \mathbf{r}^{(k)}. \tag{4.29}$$

From (4.28) and (4.29), it follows that the norm-equivalence bounds are obtained by calculating the extremal Rayleigh quotient bounds of $F^{-T}AF^{-1}$.

Rayleigh quotient

Computing the extremal Rayleigh quotient bounds is equivalent to finding the extremal eigenvalues of $F^{-T}AF^{-1}$, that is

$$\theta \leq \frac{\mathbf{v}^T F^{-T} A F^{-1} \mathbf{v}}{\mathbf{v}^T \mathbf{v}} \leq \Theta, \qquad \forall\, \mathbf{v} \in \mathbb{R}^n, \tag{4.30}$$

where θ, Θ are the smallest and the largest eigenvalues $F^{-T}AF^{-1}$ respectively. The extremal eigenvalue calculation of $F^{-T}AF^{-1}$ is equivalent to solving a generalized extremal eigenvalue problem for A and F^TF.

Find the extremal eigenvalues λ such that

$$A\mathbf{y} = \lambda F^T F \mathbf{y}, \qquad \mathbf{y} \in \mathbb{R}^n \text{ is an eigenvector.} \tag{4.31}$$

This is a symmetric positive-definite eigenvalue problem. The matrices F^TF and A are both symmetric positive-definite. Thus, (4.31) can be converted to a symmetric

[7] Assuming a non-divergent iterative method is used for solving the linear system.

positive-definite eigenvalue problem through a Cholesky factorization of F^TF. Hence, all the eigenvalues in (4.31) are real and greater than zero. If this was not the case, then obtaining a balanced stopping test based on norm-equivalence bounds might not be straightforward.

Using (4.28), (4.29), and (4.30) it follows that for $k = 0, 1, \ldots$

$$\sqrt{\theta} \leq \frac{\|\mathbf{e}^{(0)}\|_A}{\|\mathbf{r}^{(0)}\|}, \qquad \frac{\|\mathbf{e}^{(k)}\|_A}{\|\mathbf{r}^{(k)}\|} \leq \sqrt{\Theta}. \tag{4.32}$$

Equation (4.32) leads to the following upper bounds on $\|\mathbf{e}^{(k)}\|_A$, that is

$$\frac{\|\mathbf{e}^{(k)}\|_A}{\|\mathbf{e}^{(0)}\|_A} \leq \sqrt{\frac{\Theta}{\theta}}\,\frac{\|\mathbf{r}^{(k)}\|}{\|\mathbf{r}^{(0)}\|} \Longleftrightarrow \|\mathbf{e}^{(k)}\|_A \leq \sqrt{\frac{\Theta}{\theta}}\,\frac{\|\mathbf{r}^{(k)}\|}{\|\mathbf{r}^{(0)}\|}\|\mathbf{e}^{(0)}\|_A \Longleftrightarrow \|\mathbf{e}^{(k)}\|_A \leq \frac{\Theta}{\sqrt{\theta}}\|\mathbf{r}^{(k)}\|, \tag{4.33a}$$

$$\|\mathbf{e}^{(k)}\|_A \leq \sqrt{\Theta}\,\|\mathbf{r}^{(k)}\|. \tag{4.33b}$$

The quantities $\sqrt{\Theta}\|\mathbf{r}^{(k)}\|$, $\frac{\Theta}{\sqrt{\theta}}\|\mathbf{r}^{(k)}\|$ will be called weaker and stronger algebraic error bounds respectively for the rest of this chapter. The quantities $\sqrt{\Theta}$, $\frac{\Theta}{\sqrt{\theta}}$ will henceforth will be called weaker and stronger norm-equivalence constants respectively.[8]

The choice of norm-equivalence constants to be used in the employed balanced stopping test will be discussed next.

4.3.3 Stopping criterion

Based on the choice of the norm-equivalence constant, the following weaker and stronger balanced stopping criteria are obtained from (4.27) and (4.33) respectively. That is, we stop at iteration k^*, which is the smallest value of k such that

$$\sqrt{\Theta}\|\mathbf{r}^{(k^*)}\| \leq \eta^{(k^*)} \quad \Longleftrightarrow \quad \|\mathbf{r}^{(k^*)}\| \leq \frac{1}{\sqrt{\Theta}}\eta^{(k^*)}. \tag{4.34}$$

Henceforth, this will be called the weaker stopping test and the following will be known as the stronger stopping test

$$\frac{\Theta}{\sqrt{\theta}}\|\mathbf{r}^{(k^*)}\| \leq \eta^{(k^*)} \quad \Longleftrightarrow \quad \|\mathbf{r}^{(k^*)}\| \leq \frac{\sqrt{\theta}}{\Theta}\eta^{(k^*)}. \tag{4.35}$$

Note that the stopping criteria derived here can be used in iterative solvers for solving preconditioned nonsymmetric linear systems as well. The only difference from the

[8] This seemingly 'reciprocal' terminology is chosen because for balanced stopping the reciprocal of these constants are of primary interest. The above terminology has been adapted because $\frac{\sqrt{\theta}}{\Theta} \leq \frac{1}{\sqrt{\Theta}}$.

balanced stopping methodology of chapters 2 and 3 is the use of the Euclidean norm as the surrogate norm instead of using the preconditioner norm of the residual.

In terms of the number of iterations (and hence computational work and time) for convergence, the stronger stopping test cannot perform better than the weaker stopping test since $\frac{\sqrt{\theta}}{\Theta} \leq \frac{1}{\sqrt{\Theta}}$. Moreover, the stronger stopping test involves an additional overhead of computing the smallest eigenvalue. Thus, it would be prudent to employ the weaker stopping test whenever possible. A *crucial point* to observe is that if a posteriori error estimator overestimates the approximation error, it will be better to employ the stronger stopping test for otherwise the use of weaker stopping test might lead to premature stopping.

4.3.4 A posteriori error estimation

The a posteriori approximation error estimator employed here for the deterministic convection-diffusion equations (with piecewise bilinear rectangular finite elements) is reliable but need not always be efficient. It is reliable in the sense that the global upper bound on the true error does not depend on the mesh parameter h and the diffusion parameter ϵ. However, it might not always be possible that the a posteriori error estimate is a lower bound on the local (elemental) approximation error [Elman et al., 2014a, theorem 6.9, proposition 6.11, pp. 264–265]. According to [Elman et al., 2014a, p. 265], efficiency issue is generic for any local error estimator whenever boundary layers are not resolved by the FEM approximation. Hence, streamline diffusion stabilization is necessary for dealing adequately (but not completely!) with such situations.

To demonstrate that the employed a posteriori error estimator is a 'close' estimate of the approximation error, some computational results are presented for the test problem described in section 4.4. The a posteriori error η and the actual approximation error $\|u_{ref} - u_h\|_E := \|\nabla(u_{ref} - u_h)\|_2$ using a reference solution are tabulated in Table 4.1 for $2^l \times 2^l$ uniform and stretched grids respectively.

Since the exact solution to the model problem is not available, a reference solution u_{ref} is computed on a fine ($l = 9$) spatial 512×512 uniform and stretched grid respectively. This reference solution is then compared with the computed FEM solution u_h (which is linearly interpolated using MATLAB `interp2` function for compatible

Table 4.1: Approximation errors, a posteriori errors, and effectivity indices for convection-diffusion test problem on uniform (left) and stretched (right) grids.

l	η	$\|u_{ref} - u_h\|_E$	β_{eff}	$\mathcal{P}_h^{r\max}$
5	1.0562	4.1162	0.25	3.87
6	0.8556	2.6216	0.33	1.97
7	0.8018	1.5380	0.52	0.99
8	0.7855	0.7571	1.04	0.50

l	η	$\|u_{ref} - u_h\|_E$	β_{eff}	$\mathcal{P}_h^{r\max}$
5	0.8527	9.1846	0.09	9.86
6	0.8022	5.2547	0.15	5.97
7	0.7902	2.7423	0.29	3.49
8	0.7866	1.1563	0.68	2.00

comparison with the reference solution) for grids with $l = 5, 6, 7, 8$. The corresponding effectivity index $\beta_{\text{eff}} = \frac{\eta}{\|u_{ref} - u_h\|_E}$ is also presented. The columns for β_{eff} in Table 4.1 indicate that the a posteriori error estimator is an 'acceptably close' estimate of the approximation error. In fact as the mesh is refined and the layers in the solution are resolved, that is, maximum mesh Peclet number ($\mathcal{P}_h^{r\max}$) approaches ≤ 1, $\beta_{\text{eff}} \to 1$. Note that the computation of a posteriori error estimator employed here is quite cheap since it requires solving for a local 5×5 linear system on each element.

4.3.5 Computational logistics

The computational work involved in the balanced stopping test involves computation of eigenvalue(s) Θ and θ (for the stronger stopping test only), the Euclidean norm of the residual $\|\mathbf{r}^{(k)}\|$, and the a posteriori error estimator $\eta^{(k)}$. Among these, $\|\mathbf{r}^{(k)}\|$ is cheaply available in all Krylov solvers. Cheap, efficient, and reliable a posteriori estimators are also available for the present problem. If the a posteriori error computation is expensive, $\eta^{(k)}$ can be computed periodically (for example, every 4-5 iterations) to have a minor impact on the overall algorithmic cost. The computation of eigenvalues for systems of 'moderate size' (the cpu memory available) could be done easily through the MATLAB function `eigs` utilizing the sparsity of FEM matrices for this purpose. This eigenvalue computation for 'huge' matrices might be costly and alternative 'cheaper' methods for estimating these eigenvalues are therefore proposed in section 4.5.

4.4 Computational results

The computational results presented here are based only on the weaker stopping test since the a posteriori error estimator is a close (underestimate) of the approximation error; see Table 4.1. These results are presented for GMRES, BICGSTAB(2), and

TFQMR. The choice $\ell = 2$ is quite popular and widespread among practitioners; see [Elman et al., 2014a, p. 296]. Roundoff errors might pollute the residual norm computed from short-term recurrences for suboptimal Krylov solvers. In order to avoid these inaccuracies, $\|\mathbf{r}^{(k)}\|$ is computed here after forming the residual explicitly, that is, $\mathbf{r}^{(k)} = \mathbf{b} - F\mathbf{x}^{(k)}$. It is claimed here that in presence of tight a posteriori approximation error estimators, the balanced stopping test can be employed optimally for suboptimal iterative methods too provided breakdowns are handled adequately and these algorithms 'converge' at least to the accuracy of the true approximation error.

The computational experiments are carried out in the IFISS software in MATLAB. Four test problems based on (4.1) are present in IFISS. Computational results are presented here for the fourth test problem, which is characterized by a recirculating wind and has discontinuous Dirichlet boundary conditions leading to the formation of boundary layers near the corners of the domain [Elman et al., 2014a, p. 240]; see Figure 4.1.

- Recirculating wind
  $\vec{w} = (2x_2(1-x_1^2), -2x_1(1-x_2^2)),\ \forall\,(x_1,x_2) \in D.$
- Dirichlet discontinuous boundary conditions
  $u(x_1,-1) = 0\,,\ u(x_1,1) = 0,\ u(-1,x_2) = 0,\ u(1,x_2) = 1,\ \forall\,(x_1,x_2) \in D.$

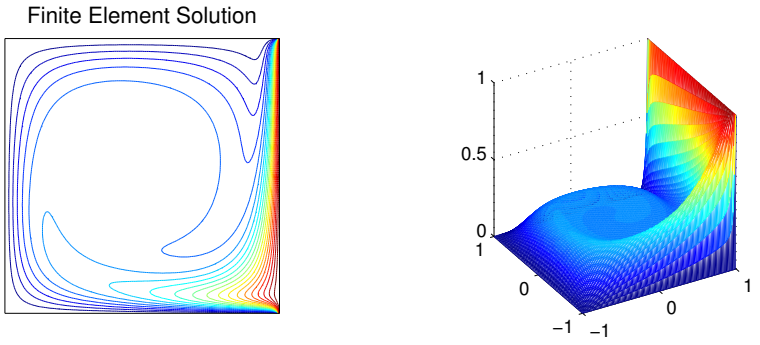

Figure 4.1: FEM solution surface and contour plots from the MATLAB backslash solution on a uniform grid for $l = 7$.

The convection-diffusion (CD) problem (4.1) is defined on $D = (-1,1) \times (-1,1)$ with the source function $f(x_1,x_2) = 0$, $\forall\,(x_1,x_2) \in D$. Rectangular piecewise bilinear ($\boldsymbol{Q}_1$) finite elements are used on $2^l \times 2^l$ uniform and stretched grids. The viscosity parameter $\epsilon = 1/64$ is fixed and the optimal inbuilt value of the *stabilization* parameter is used;

see [Elman et al., 2014a, p. 253]. This problem can be set up by choosing test problem `4` after running the driver `cd_testproblem` in IFISS.

There are four preconditioners built in IFISS for the discrete convection-diffusion problem. They are: diagonal (DIAG) preconditioner, that is, the diagonal matrix formed from the diagonal elements of F, incomplete LU (ILU), geometric multigrid (GMG), and algebraic multigrid (AMG) preconditioners; see [Elman et al., 2014a, chapter 7]. Note that the ILU, GMG, and AMG preconditioners are employed with their specified default settings in IFISS.

Let $\mathbf{x}$ denote the MATLAB backslash (Gaussian elimination) solution on each grid. Henceforth, this will be regarded as the true algebraic solution. This will be used for comparison with the result $\mathbf{x}^{(k^*)}$ computed using the balanced stopping test. From $\mathbf{x}$, the 'exact' a posteriori error estimate η is computed. The starting vector $\mathbf{x}^{(0)}$ is generated using the MATLAB function `rand`. The balanced stopping test that is used in preconditioned GMRES and BICGSTAB(ℓ) is implemented in `gmres_r` and `bicgstab_ell` in IFISS respectively, while the balanced stopping test in preconditioned TFQMR is incorporated in the existing MATLAB function for this solver. Also, let $\eta^{(k^*)}$ denote the a posteriori error estimate at the optimal stopping iteration k^* and $e_\eta^* := |\eta - \eta^{(k^*)}|$. These values are tabulated in Tables 4.2–4.7 for each preconditioner on every grid level for both uniform and stretched grids. The insights from these numbers are quite generic, which are summarised in the following paragraphs.

The e_η^* columns show that $\{\eta^{(k)}\}$ has converged with a good accuracy to the true a posteriori error estimate η at the balanced stopping iteration. To show the effectiveness of the balanced stopping test, the iteration counts k^* needed to satisfy the balanced stopping test have been compared with iteration counts $k_{\text{tol1}}, k_{\text{tol2}}$ needed to satisfy a fixed relative residual $\dfrac{\|\mathbf{r}^{(k)}\|}{\|\mathbf{r}^{(0)}\|}$ reduction tolerance of `1e-6` (which is the default tolerance in MATLAB solvers) and `1e-9` respectively. These tolerance values are a realistic user-input tolerance choices in the absence of a balanced stopping test. The user will not know in general the stopping point k^* a priori and is more likely to provide a tighter tolerance than actually required. This would lead to unnecessary computations. In any case, using a balanced stopping strategy would rule out premature stopping of the chosen iterative solver.

A comparison of the corresponding columns for iteration counts shows that for

the same approximation error, a significant number of iterations is saved by using the balanced stopping test. This would result in significant savings in computational work of the solver (as compared to using fixed relative residual $\frac{\|\mathbf{r}^{(k)}\|}{\|\mathbf{r}^{(0)}\|}$ reduction tolerance `1e-6` or tighter) if one were to solve the (preconditioned) linear systems arising from adaptive finite element for the chosen problem parameters. The linear systems that are solved are of size: 1089×1089, 4225×4225, 16641×16641, and 66049×66049. These computational savings are even more striking in light of the huge size of some of these systems.

Among the employed iterative methods, BICGSTAB(2) performs the best with each preconditioner. Between GMRES and TFQMR, GMRES converges slightly faster. However, using GMRES over TFQMR could be memory extensive in terms of storage. In any case, the balanced stopping test provides an optimal stopping point for suboptimal Krylov solvers like TFQMR etc. Indeed this is crucially dependent on the fact that these suboptimal solvers do not break down prematurely.

The iteration counts for the diagonal and the ILU preconditioner almost double with each grid refinement. On stretched grids, the convergence of the ILU and the diagonal preconditioner (in particular) improves drastically from their convergence on uniform grids. The iteration counts for these preconditioners on the stretched grids is in general almost halved or better from their iteration count on the corresponding uniform grid. This happens because grid stretch leads to a better handle on the boundary layers present in this problem. Thus, in some sense the diagonal matrix and the LU factors carry more information about the problem and hence converge faster than on the corresponding uniform grid.

Among the preconditioners, the rate of convergence in terms of iterations is similar for all the methods in the sense that GMG and AMG are the best preconditioners, while diagonal preconditioning is the worst with ILU somewhere in between. In fact the iteration counts k^* for GMG and AMG preconditioners indicate that their rate of convergence on uniform grids is mesh-independent [Elman et al., 2014a, p. 326] but rate of convergence of AMG on stretched grids is not mesh-independent; see iteration counts for AMG on stretched grids.

In order to gain further insight from the numerical experiments, the evolution of the following quantities—$\eta^{(k)}$ (red curve), $\|\mathbf{e}^{(k)}\|_A$ (cyan curve), $\|\mathbf{r}^{(k)}\|$ (green curve), and the weak algebraic error bound $\sqrt{\Theta}\|\mathbf{r}^{(k)}\|$ (blue curve) is also plotted.

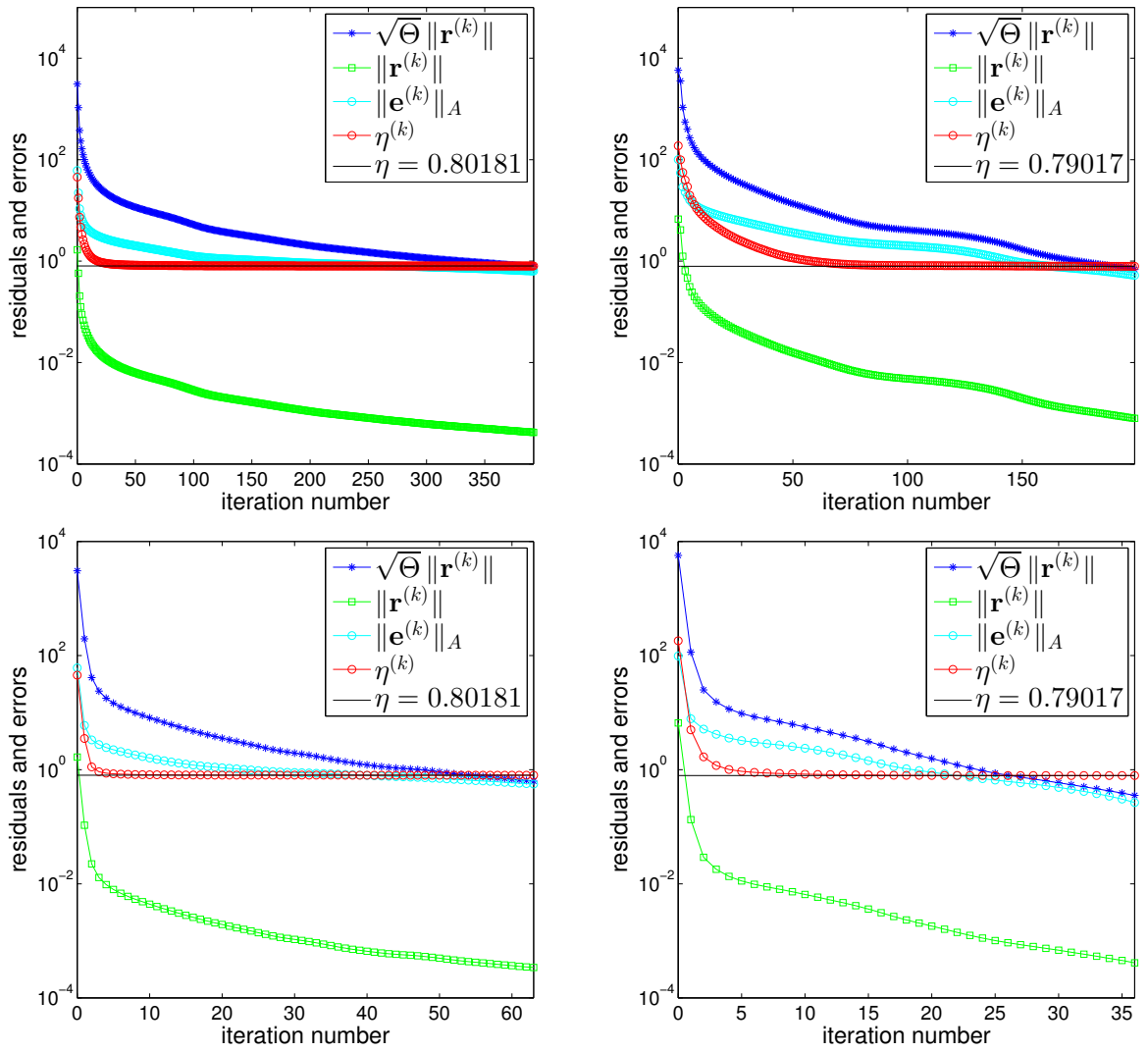

Figure 4.2: Errors vs iteration number for GMRES with DIAG (top) and ILU (bottom) preconditioning on a uniform (left) and stretched (right) grid for $l = 7$.

Table 4.2: GMRES iteration counts & errors for DIAG (top) & ILU (bottom) preconditioning on uniform (left) & stretched (right) grids for discrete CD system.

l	k_{tol1}	k_{tol2}	k^*	e^*_η
5	177	218	77	2.1e-3
6	381	476	172	5.3e-4
7	797	1001	383	1.2e-4
8	1501	1942	776	2.8e-5

l	k_{tol1}	k_{tol2}	k^*	e^*_η
5	97	127	42	8.8e-3
6	177	245	88	3.0e-3
7	349	518	190	1.4e-3
8	627	996	404	5.9e-4

l	k_{tol1}	k_{tol2}	k^*	e^*_η
5	19	24	7	1.9e-3
6	43	54	19	4.4e-4
7	113	144	54	1.4e-4
8	288	374	148	2.9e-5

l	k_{tol1}	k_{tol2}	k^*	e^*_η
5	16	21	7	4.4e-3
6	28	38	14	1.8e-3
7	52	73	27	7.9e-4
8	102	148	061	2.5e-4

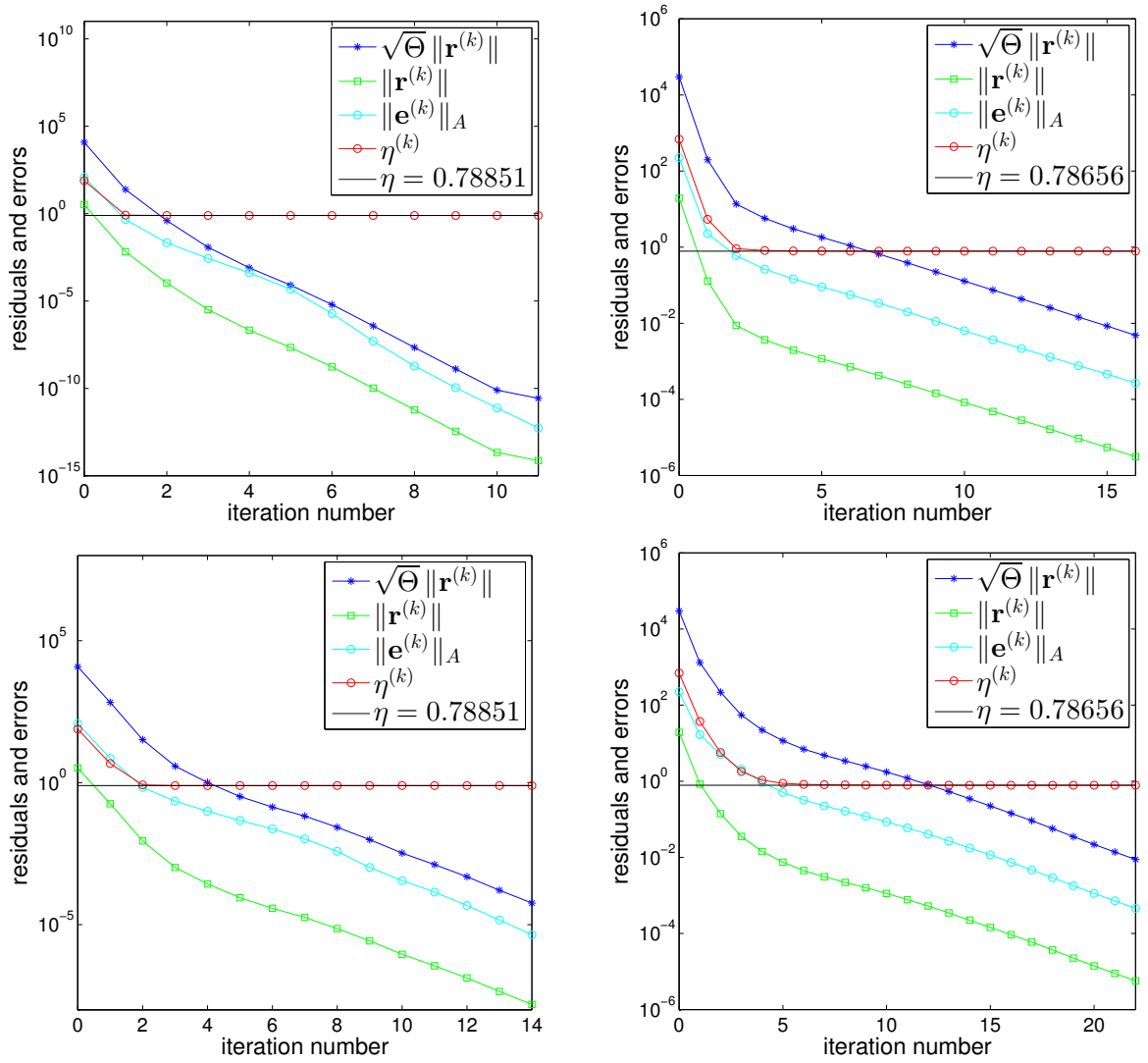

Figure 4.3: Errors vs iteration number for GMRES with GMG (top) and AMG (bottom) preconditioning on a uniform (left) and stretched (right) grid for $l = 8$.

Table 4.3: GMRES iteration counts & errors for GMG (top) & AMG (bottom) preconditioning on uniform (left) & stretched (right) grids for discrete CD system.

l	k_{tol1}	k_{tol2}	k^*	e^*_η
5	4	6	2	6.4e-5
6	3	6	2	3.4e-4
7	3	6	2	9.3e-5
8	3	5	2	1.0e-5

l	k_{tol1}	k_{tol2}	k^*	e^*_η
5	5	7	2	9.1e-4
6	4	8	3	2.6e-4
7	5	12	4	9.1e-4
8	4	15	7	3.9e-4

l	k_{tol1}	k_{tol2}	k^*	e^*_η
5	6	10	3	1.4e-3
6	7	11	4	7.7e-5
7	8	14	4	4.4e-5
8	7	14	5	1.8e-6

l	k_{tol1}	k_{tol2}	k^*	e^*_η
5	6	10	3	1.2e-3
6	6	12	4	1.0e-3
7	9	17	6	6.9e-4
8	12	28	13	2.6e-4

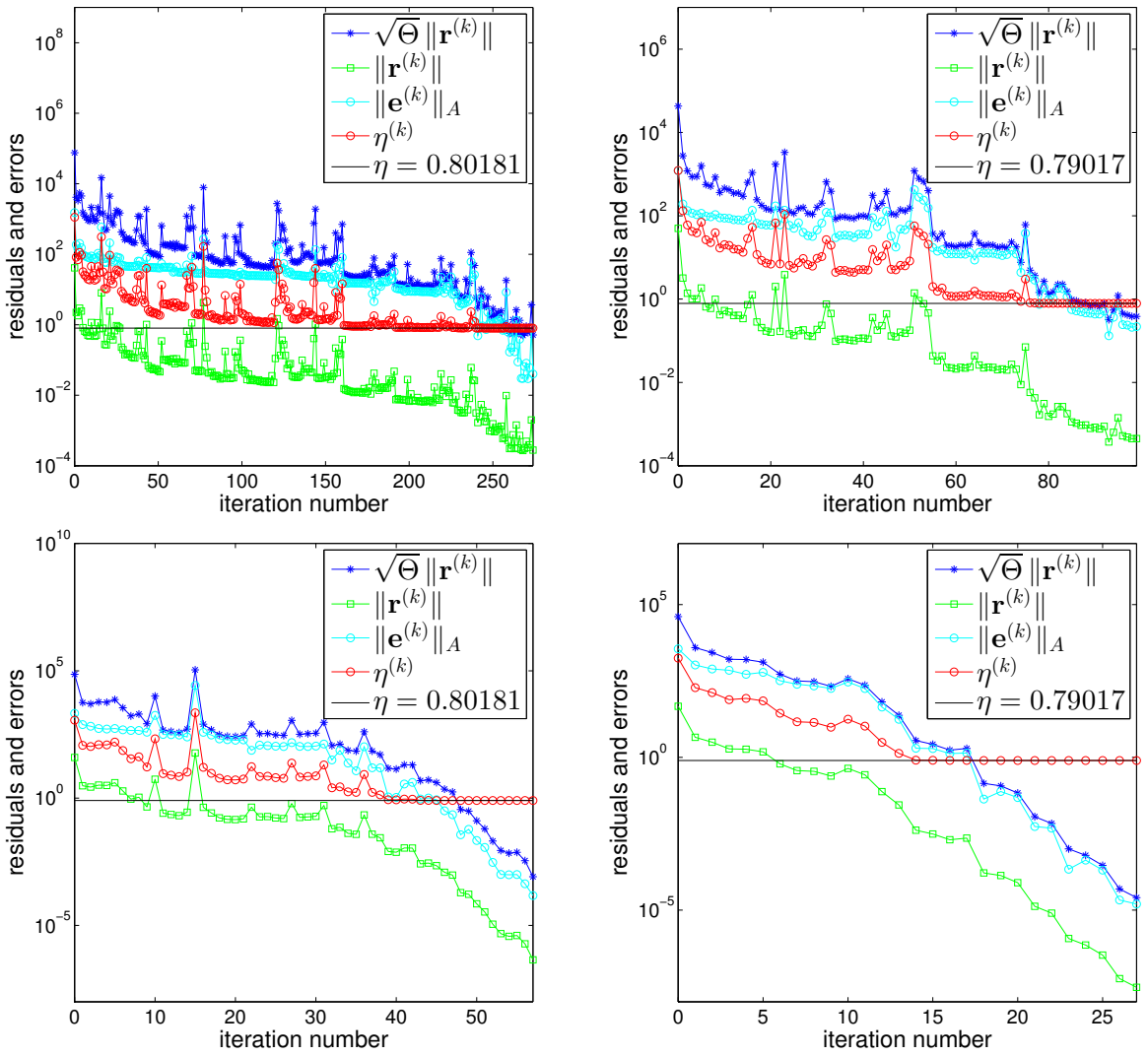

Figure 4.4: Errors vs iteration number for BICGSTAB(2) with DIAG (top) and ILU (bottom) preconditioning on a uniform (left) and stretched (right) grid for $l = 7$.

Table 4.4: BICGSTAB(2) iteration counts & errors for DIAG (top) & ILU (bottom) preconditioning on uniform (left) & stretched (right) grids for discrete CD system.

l	k_{tol1}	k_{tol2}	k^*	e^*_η
5	126	164	62	7.6e-4
6	286	376	136	5.3e-4
7	598	790	260	1.3e-4
8	1224	1488	544	5.7e-5

l	k_{tol1}	k_{tol2}	k^*	e^*_η
5	68	86	25	8.3e-3
6	118	170	41	8.7e-4
7	228	362	89	5.1e-4
8	502	918	192	2.9e-4

l	k_{tol1}	k_{tol2}	k^*	e^*_η
5	12	16	6	4.1e-5
6	30	38	15	1.7e-4
7	86	114	48	2.8e-5
8	236	290	124	3.5e-5

l	k_{tol1}	k_{tol2}	k^*	e^*_η
5	10	14	5	4.9e-5
6	18	24	9	5.1e-4
7	32	52	18	1.6e-5
8	58	96	036	1.2e-4

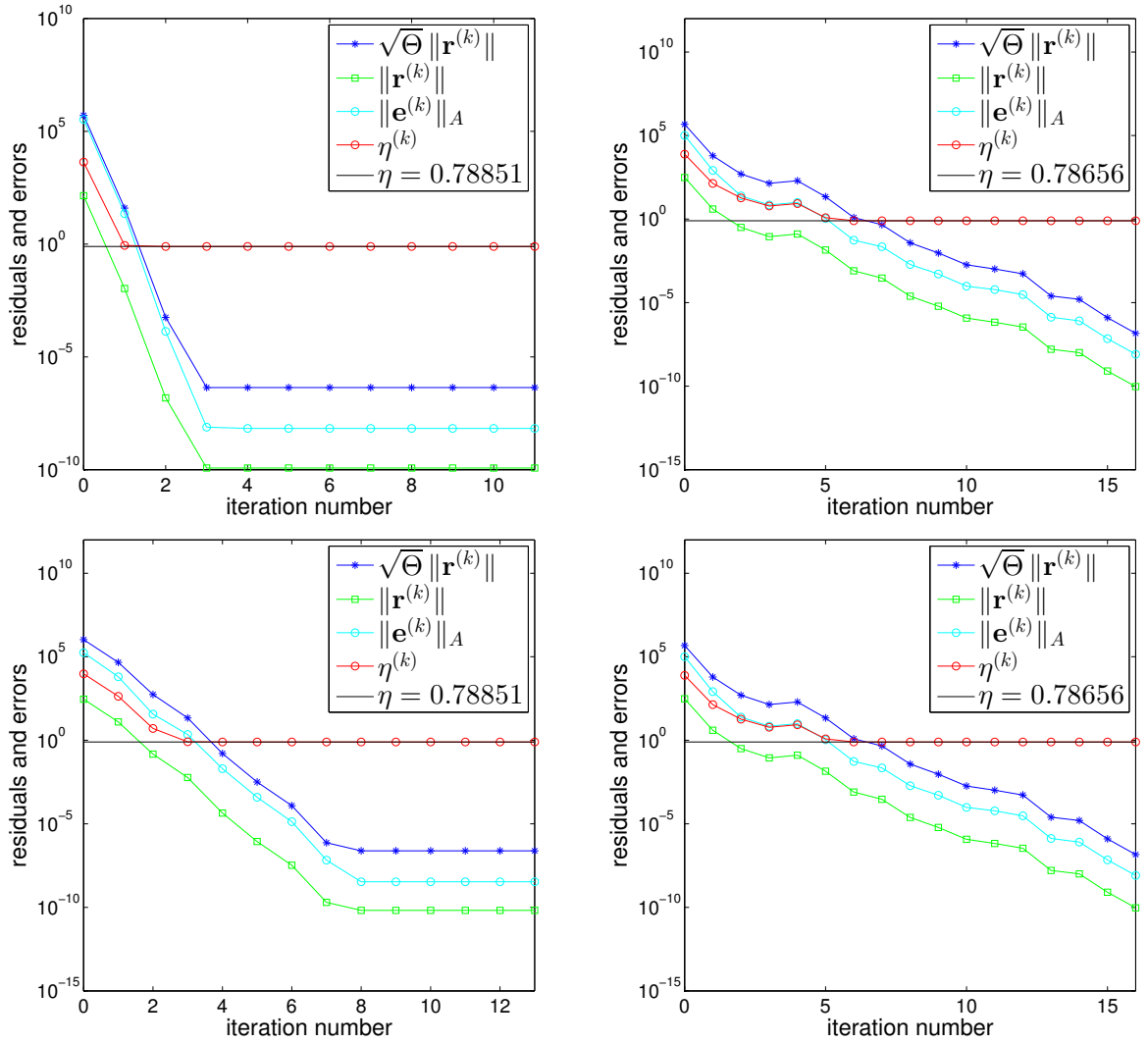

Figure 4.5: Errors vs iteration number for BICGSTAB(2) with GMG (top) and AMG (bottom) preconditioning on a uniform (left) and stretched (right) grid for $l = 8$.

Table 4.5: BICGSTAB(2) iteration counts & errors for GMG (top) & AMG (bottom) preconditioning on uniform (left) & stretched (right) grids for discrete CD system.

l	k_{tol1}	k_{tol2}	k^*	e^*_η
5	2	4	1	1.4e-4
6	2	4	2	2.2e-8
7	2	4	2	1.7e-8
8	2	4	2	1.1e-8

l	k_{tol1}	k_{tol2}	k^*	e^*_η
5	4	4	2	7.0e-7
6	4	4	2	5.7e-8
7	4	8	3	2.6e-6
8	2	8	5	3.6e-5

l	k_{tol1}	k_{tol2}	k^*	e^*_η
5	4	6	2	2.0e-4
6	4	6	3	3.7e-6
7	4	8	4	8.0e-6
8	4	8	4	3.4e-6

l	k_{tol1}	k_{tol2}	k^*	e^*_η
5	4	6	2	1.1e-5
6	4	6	2	6.6e-4
7	6	10	4	6.0e-6
8	8	18	7	1.7e-4

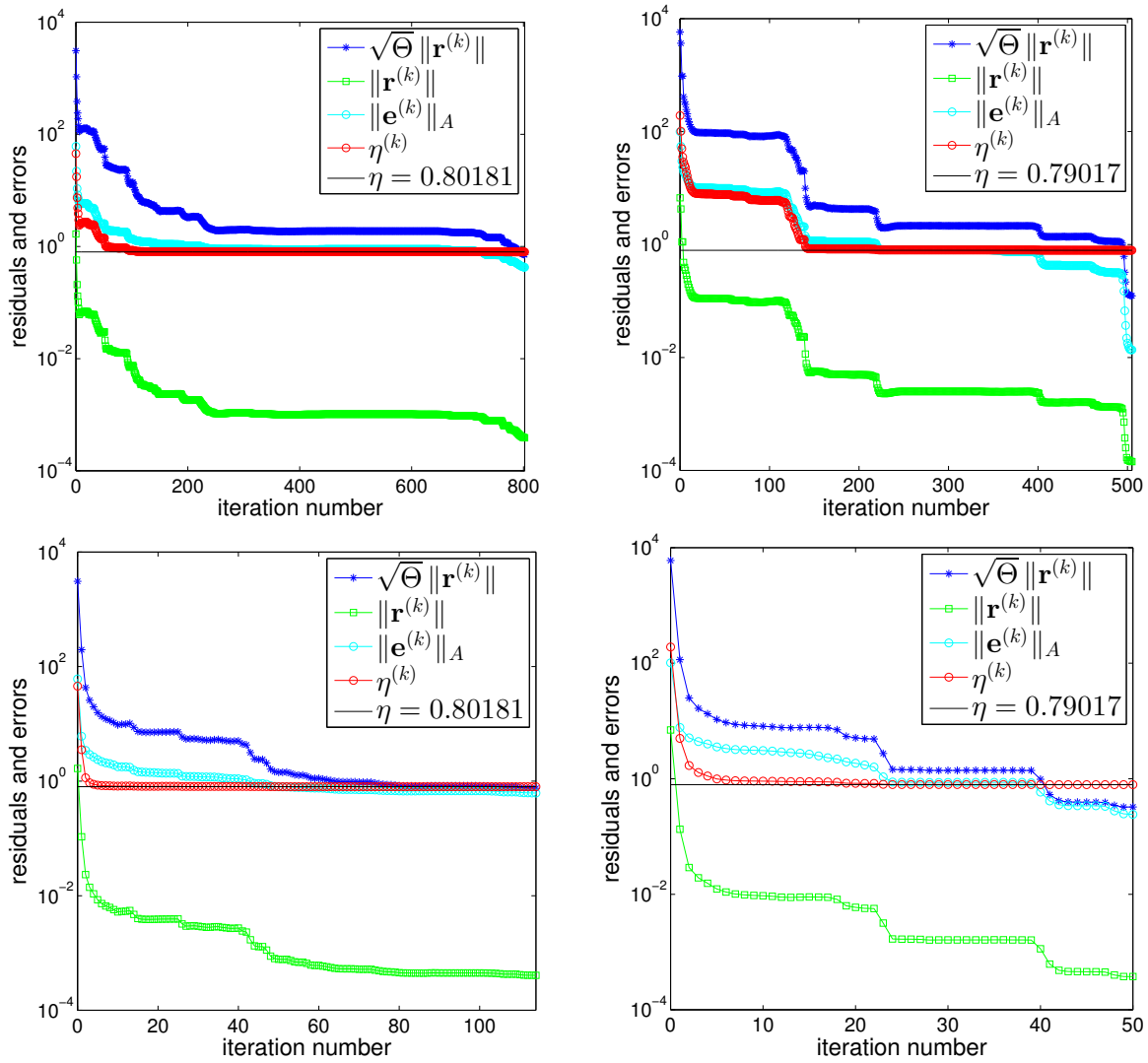

Figure 4.6: Errors vs iteration number for TFQMR with DIAG (top) and ILU (bottom) preconditioning on a uniform (left) and stretched (right) grid for $l = 7$.

Table 4.6: TFQMR iteration counts & errors for DIAG (top) & ILU (bottom) preconditioning on uniform (left) & stretched (right) grids for discrete CD system.

l	k_{tol1}	k_{tol2}	k^*	e^*_η
5	298	365	231	6.5e-3
6	707	806	334	1.3e-3
7	1521	1767	793	3.6e-4
8	3066	3516	2304	9.9e-5

l	k_{tol1}	k_{tol2}	k^*	e^*_η
5	163	209	107	7.0e-3
6	309	376	209	2.8e-3
7	574	871	496	6.1e-4
8	1209	1699	952	1.1e-4

l	k_{tol1}	k_{tol2}	k^*	e^*_η
5	32	37	15	7.5e-4
6	73	84	36	8.3e-5
7	193	234	105	4.4e-5
8	534	684	0345	4.8e-5

l	k_{tol1}	k_{tol2}	k^*	e^*_η
5	26	28	11	4.3e-3
6	43	56	21	2.5e-3
7	87	110	41	4.1e-4
8	173	218	112	2.5e-4

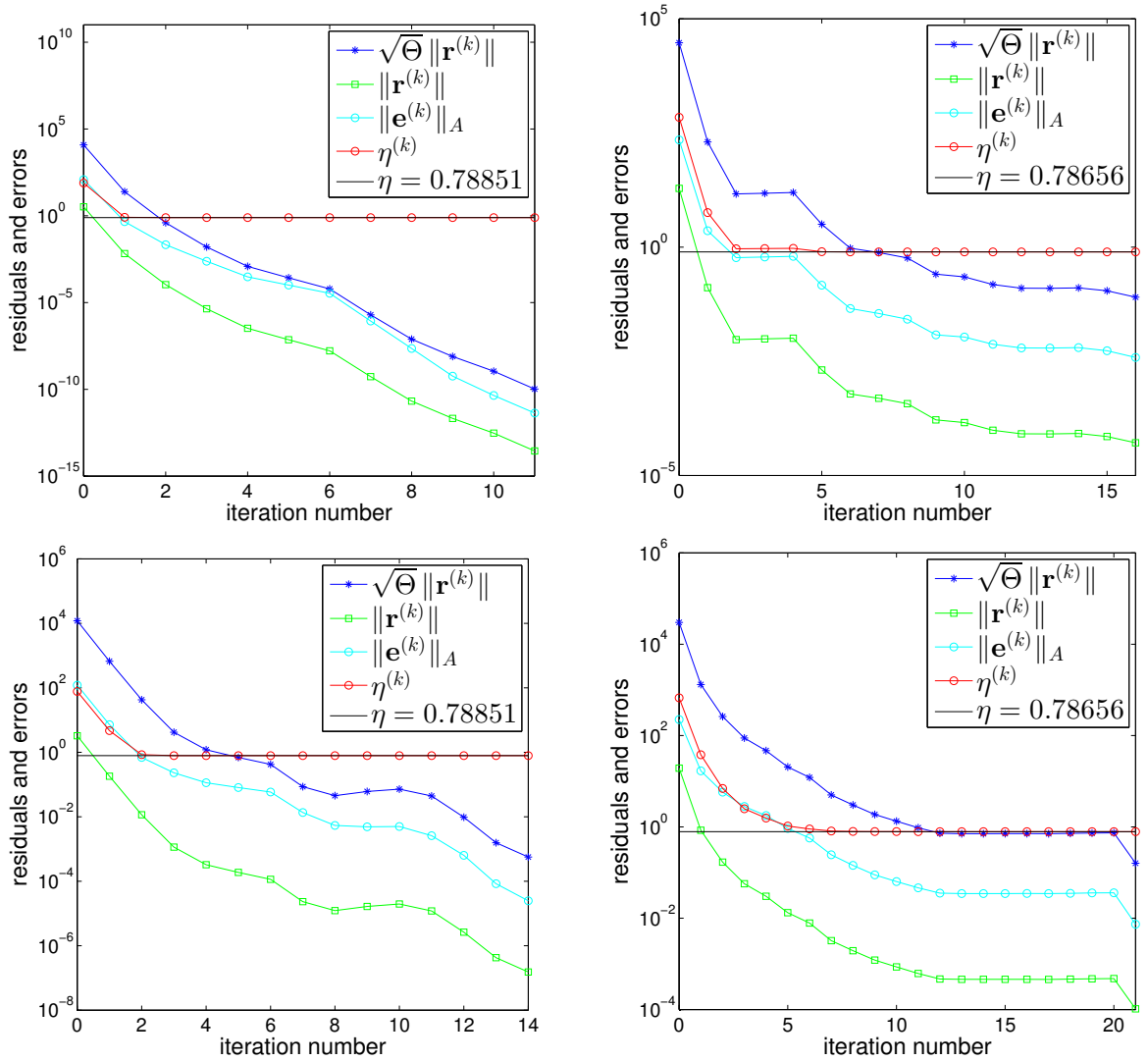

Figure 4.7: Errors vs iteration number for TFQMR with GMG (top) and AMG (bottom) preconditioning on a uniform (left) and stretched (right) grid for $l = 8$.

Table 4.7: TFQMR iteration counts & errors for GMG (top) & AMG (bottom) preconditioning on uniform (left) & stretched (right) grids for discrete CD system.

l	k_{tol1}	k_{tol2}	k^*	e^*_η
5	4	7	2	1.8e-4
6	5	8	2	1.8e-4
7	4	7	2	8.5e-5
8	3	6	2	1.2e-5

l	k_{tol1}	k_{tol2}	k^*	e^*_η
5	6	9	2	6.8e-4
6	6	10	3	4.9e-4
7	11	18	5	3.7e-4
8	15	29	7	5.6e-4

l	k_{tol1}	k_{tol2}	k^*	e^*_η
5	9	12	4	4.1e-5
6	7	14	4	4.4e-4
7	9	19	4	1.9e-5
8	8	17	5	4.3e-5

l	k_{tol1}	k_{tol2}	k^*	e^*_η
5	8	13	3	8.7e-4
6	9	14	5	1.5e-4
7	16	21	7	3.5e-4
8	26	43	12	5.8e-4

The balanced stopping test stops optimally when the blue curve is below the red curve. From (4.27) it follows that when the contribution of $\|\mathbf{e}^{(k)}\|_A$ to the sum $\eta + \|\mathbf{e}^{(k)}\|_A$ is insignificant,[9] $\{\eta^{(k)}\}$ (red curve) converges to η (black line). Indeed this is the case in all plots of Figures 4.2–4.7. In order to illustrate convergence, iterations have been continued for nine more steps after optimal stopping in each plot. This also illustrates optimal stopping at the correct iteration, that is $\{\eta^{(k)}\}$ converges to η on each plot. In Figures 4.2 and 4.3, it is noticed that after a initial burn in period, the rate of convergence of $\|\mathbf{r}^{(k)}\|$ is constant and is in fact the asymptotic convergence factor defined in (4.15). This burn in period for the 'bad' diagonal preconditioner is very large while for the 'best' preconditioners GMG and AMG the burn in period is negligible. Also, in Figures 4.2 and 4.3, Euclidean norm of the residual $\|\mathbf{r}^{(k)}\|$ is monotonically decreasing in GMRES while it exhibits irregular behaviour for BICGSTAB(2) and TFQMR; see Figures 4.4 and 4.6. However, a 'good' preconditioner smooths out the irregular behaviour to a large extent; see Figures 4.5, 4.7 for multigrid preconditioned BICGSTAB(2) and TFQMR respectively. These observations are consistent with the discussions in section 4.2.2.

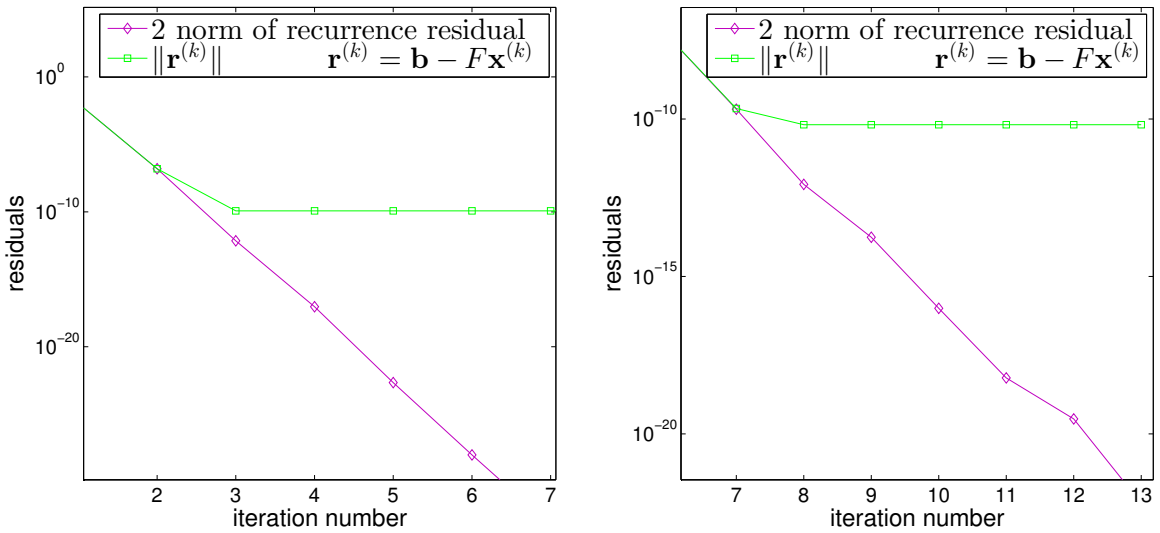

Figure 4.8: Computed and recurrence residuals 2-norm vs iteration number for GMG (left) and AMG (right) preconditioned BICGSTAB(2) on a uniform grid for $l = 8$.

Note that the plots for GMG and AMG preconditioned BICGSTAB(2) on uniform grid $l = 8$ in Figure 4.5 indicate that the computed residual ($\mathbf{r}^{(k)} = \mathbf{b} - F\mathbf{x}^{(k)}$)

[9]From visual inspection this seems to occur soon after $\|\mathbf{e}^{(k)}\|_A \leq \eta$. But in most cases both these quantities are unknown. So a priori knowledge of optimal stopping iterative step is generally difficult.

norm $\|\mathbf{r}^{(k)}\|$ stagnates after a few iterations. An investigation of the 2 norm of the residual obtained from BICGSTAB(2) recurrences suggests a possible cause for the stagnation behaviour of the computed residual. With respect to iteration count k, the evolution of $\|\mathbf{r}^{(k)}\|$ and 2-norm of the BICGSTAB(2) recurrence residual for both GMG and AMG preconditioning on the uniform grid $l = 8$ is plotted in Figure 4.8. From the plots, observe that the stagnation of $\|\mathbf{r}^{(k)}\|$ begins when the 2-norm of the corresponding residual obtained from BICGSTAB(2) recurrence is 'close' to machine epsilon (`2.2204e-16` here) and $\|\mathbf{r}^{(k)}\|$ completely stagnates when the 2-norm of the corresponding BICGSTAB(2) recurrence residual crosses machine epsilon. When the 2 norm of the corresponding BICGSTAB(2) recurrence residual is 'near' and crosses machine epsilon it implies that the numbers computed in the BICGSTAB(2) iterations have become quite 'small' in size and thus $\|\mathbf{r}^{(k)}\|$ which is computed using $\mathbf{x}^{(k)}$ of BICGSTAB(2) does not reduce further but stagnates.

The most important observation from the plots for BICGSTAB(2) and TFQMR is the following. After balanced stopping, the blue curve for algebraic error bound can jump above the red curve for $\eta^{(k)}$ (see plot of diagonal preconditioned BICGSTAB(2) on stretched grid in Figure 4.4); but this is not an issue since the solver has already stopped optimally. The main aim of these computational results is not to compare the convergence rates of various suboptimal Krylov solvers but to illustrate that an optimal balanced black-box stopping test can be devised for such solvers.

4.5 Cheap eigenvalue estimation in stopping test

As explained earlier, the computation of the extremal eigenvalues θ, Θ through the MATLAB command `eigs` could be expensive for huge linear systems. Some alternative approaches are suggested in this section to address this issue.

4.5.1 Solve the corresponding normal equations

One approach could be to solve $F^TF\mathbf{x} = F^T\mathbf{b}$—a symmetric positive-definite system– instead of solving the nonsymmetric system (4.9) and measure all the errors in the natural vector norm F^TF. Using preconditioned MINRES and balanced stopping methodology developed in chapter 2, the eigenvalues of interest can then be estimated on-the-fly by the corresponding Ritz values. However, the continuous analogue of the

vector norm F^TF might not be a physically meaningful norm for (4.4) and also the product of a vector with F^T might not be cheaply available. Interestingly, notice that choosing a preconditioner M such that $M^{-1}F$ is a symmetric positive-definite matrix is another alternative.

4.5.2 Information from spectrum of F

Suppose that F is a normal matrix. Then it admits an orthonormal decomposition $F = V^*DV \iff F^* = V^*D^*V$, where V is an orthonormal matrix and $D = (d_{jj})_{j=1}^n$ is a diagonal matrix with (complex) eigenvalues of F on its diagonal. Also, by (4.10) $A = \dfrac{F + F^*}{2\epsilon}$;[10] note that the matrices under discussion are real valued and hence symmetric and self-adjoint matrices are the same here.

Now consider the characteristic polynomial of A and F^*F, and let λ be its arbitrary eigenvalue

$$det\left(\frac{F + F^*}{2\epsilon} - \lambda F^*F\right) = det\left(\frac{D + D^*}{2\epsilon} - \lambda D^*D\right). \tag{4.36}$$

Here $det(.)$ denotes determinant and the fact that the determinant of a matrix is independent under similarity transforms has been used in (4.36). For $j = 1, 2, \ldots, n$, if $d_{jj} = p_j + iq_j \iff d_{jj}^* = p_j - iq_j$, where $i = \sqrt{-1}$, p_j and $q_j \in \mathbb{R}$. Also, note that the matrix $\dfrac{D + D^*}{2\epsilon} - \lambda D^*D$ is diagonal and its determinant is just the product of its diagonal entries. From the underlying discussion, and the characteristic equation associated with (4.36)

$$\lambda = \frac{p_j}{\epsilon(p_j^2 + q_j^2)}. \tag{4.37}$$

Thus, a knowledge of spectrum of F will help to compute the eigenvalues involved in the stopping test cheaply. However, this analysis is based on the assumption that F is normal, which is not always true. Also, the spectrum of F might not be easily available.

4.5.3 Information from parameters of the problem

A more practical approach that is still under research is obtain an estimate on the eigenvalues using the parameters of the problem, that is, ϵ, $\|\vec{w}\|$, and h (mesh step size

[10] When no stabilization is employed.

for uniform grids). Note that only an upper bound (estimate) for $\sqrt{\Theta}$ is required to use the weaker stopping test (4.34).

From Table 4.8 observe that Θ is essentially four times its value on previous grid. If such behaviour is known beforehand then one needs to compute Θ using MATLAB `eigs` on a coarse grid only. However, such information is rarely available in advance.

The 2 fold increase for $\sqrt{\Theta}$ on successive grids (keeping ϵ, $\|\vec{w}\|$ fixed) suggests that $\sqrt{\Theta}$ varies as $1/h$. Also, computations suggest that $\sqrt{\Theta}$ is also dependent on ϵ (keeping h, $\|\vec{w}\|$ fixed); see Table 4.8. Based on heuristics, it seems here that for a given h and ϵ, $1/2\epsilon h$ will be an upper bound on $\sqrt{\Theta}$. This is indeed the case as confirmed by our computations; see last two columns of each table in Table 4.8.

Table 4.8: Computed Θ from MATLAB `eigs` and its upper bound estimate for CD test problem on uniform grids for $\epsilon = 1/64$ (left) and $\epsilon = 1/200$ (right).

h	Θ	$\sqrt{\Theta}$	$1/2\epsilon h$
1/16	2.1279e+5	461.2917	512
1/32	8.5019e+5	922.0575	1024
1/64	3.3993e+6	1843.7190	2048
1/128	1.3596e+7	3686.4617	4096

h	Θ	$\sqrt{\Theta}$	$1/2\epsilon h$
1/16	2.0717e+6	1439.3401	1600
1/32	8.2995e+6	2880.8853	3200
1/64	3.3195e+7	5761.5102	6400
1/128	1.3277e+8	11522.5865	12800

Note that the dependence of Θ on $\|\vec{w}\|$ (keeping ϵ fixed) in still under investigation. Thus, it seems prudent to obtain an upper bound for $\sqrt{\Theta}$ in terms of the maximum mesh element Peclet number (since it incorporates the contribution of convection too). Since $\dfrac{1}{2\epsilon h} \leq \dfrac{1}{h^2}\dfrac{h\|\vec{w}\|}{2\epsilon}$, it follows from the definition of mesh element Peclet numbers in (4.3) that $\dfrac{1}{2\epsilon h} \leq \dfrac{1}{h^2}\mathcal{P}_h^{r_{\max}}$, where $\mathcal{P}_h^{r_{\max}}$ denotes the maximum mesh element Peclet number. The discussion presented here is based solely on heuristics/computations. A rigorous mathematical discussion in this direction is still under research.

4.6 Summary

An optimal balanced black-box stopping tests for solving nonsymmetric linear systems arising from (stochastic) FEM approximation of convection-diffusion equations has been devised in GMRES, BICGSTAB(ℓ), and TFQMR. In fact these algorithms can be modified to cater for solving nonsymmetric linear systems arising from other PDEs as well. The PDE origins of these systems have to be taken into account when devising an optimal balanced black-box stopping test. A balanced stopping criterion can be

constructed in the presence of a good preconditioner and a tight a posteriori error estimation routine.

Some methods have also been suggested to cheaply estimate the constants involved in the balanced stopping test. The balanced stopping test can result in significant computational savings and this aspect becomes especially significant when solving a (stochastic) PDE through collocation, adaptively, or in higher spatial dimensions using FEM. A balanced stopping test for memory inexpensive Krylov solvers such as BICGSTAB(ℓ), TFQMR etc., has also been devised. Currently, little convergence theory exists for such solvers and so the devised balanced stopping test is crucial to rule out premature stopping of such solvers.

Chapter 5

Balanced iterative stopping for nonsymmetric systems II

Publication

- The material presented in this chapter along with the work in chapter 4 will soon be submitted for publication.
- The devised balanced stopping test in GMRES for solving nonsymmetric linear systems arising (at every step of the nonlinear Picard or Newton iteration) from FEM approximation of (parametric) Navier–Stokes equations has resulted in the function `NAVIER_NEWTON_GMRES` in the software IFISS [Elman et al., 2014b].

Balanced stopping tests in iterative solvers for solving a single linear system arising from FEM approximation of a (stochastic) PDE has been devised in previous chapters. It is observed therein that a balanced stopping test not only avoids premature stopping of the employed iterative solver but also usually leads to huge computational savings. This motivates the use of a balanced stopping strategy when iteratively solving huge linear systems arising at every step from the associated linearized part of the employed nonlinear iterative solver (such as the Newton or Picard method). In any case, it would avoid premature stopping and hence provide a 'good' approximation to the linearized part of the nonlinear iterative solver at each (nonlinear) iterative step. Otherwise, the convergence (if at all) of the nonlinear iterative solver will be slow due to 'not so accurate' linear solves of the associated linearized part.

The deterministic Navier–Stokes equations is the underlying PDE considered here for illustrating the balanced stopping strategy since a posteriori approximation error estimation is not yet fully developed for stochastic Navier–Stokes equations. The FEM solution of Navier–Stokes equations involves solving a nonsymmetric linear system at each iterative step of the employed nonlinear iterative solver. Thus, an optimal balanced black-box stopping strategy in GMRES will be used here to solve these linear systems. This balanced strategy will only be a slight variant of the balanced stopping strategy developed in GMRES in chapter 4. Note that BICGSTAB(ℓ) and TFQMR could also be used for solving the nonsymmetric linear systems that arise here. However, the results are presented only for the GMRES solver.

Stopping criterion for the linearized part of the nonlinear iteration for the solution of Navier–Stokes equations has been devised by [Arioli and Loghin, 2008]. However, the stopping test therein is based on a priori error estimators while the stopping test presented in this chapter is based on a posteriori error estimation strategy.

This chapter consists of 5 sections. The Navier–Stokes PDE, its weak form, and FEM form is set up in section 5.1. The target nonlinear system is formulated in section 5.2 and the methods for solving it are discussed therein. The balanced stopping strategy in GMRES is presented in section 5.3 and its effectiveness is illustrated through some computational results in section 5.4. Section 5.5 contains a summary of this chapter.

5.1 Deterministic Navier–Stokes equations

Navier–Stokes equations form the fundamental model of an incompressible Newtonian fluid such as air etc. Similar to the Stokes equations, the steady-state Navier–Stokes solution $(\vec{u}, p)$ is defined on a spatial domain $D \subset \mathbb{R}^d$, $(d = 2, 3)$, where the vector valued velocity function $\vec{u}(\vec{x}) : D \to \mathbb{R}^d$ and the scalar valued pressure function $p(\vec{x}) : D \to \mathbb{R}$ satisfy [Elman et al., 2014a, p. 333 ff.]

$$-\nu\nabla\cdot\nabla\vec{u}(\vec{x}) + \vec{u}(\vec{x})\cdot\nabla\vec{u}(\vec{x}) + \nabla p(\vec{x}) = \vec{f}(\vec{x}), \qquad \forall\,\vec{x}\in D, \tag{5.1a}$$

$$\nabla\cdot\vec{u}(\vec{x}) = 0, \qquad \forall\,\vec{x}\in D, \tag{5.1b}$$

$$\vec{u}(\vec{x}) = \vec{w}(\vec{x}), \qquad \forall\,\vec{x}\in \partial D_D, \tag{5.1c}$$

$$\nu\nabla\vec{u}(\vec{x})\cdot\vec{n} - \vec{n}p(\vec{x}) = \vec{0}, \qquad \forall\,\vec{x}\in \partial D_N. \tag{5.1d}$$

The functions $\vec{f}$, $\vec{w}$ are given and ∂D_D, ∂D_N are the Dirichlet and Neumann parts respectively of the spatial boundary ∂D. Kinematic velocity $\nu > 0$ is given and $\vec{n}$ denotes the outward normal to ∂D. Note that the presence of the convective term gives the Navier–Stokes equations a nonlinear behaviour. Also, similar to convection-diffusion equations, the Reynolds number $\mathcal{R}$ [Elman et al., 2014a, p. 334] encapsulates here the relative contribution of convection and diffusion and is defined as

$$\mathcal{R} := \frac{|\vec{u}|\mathrm{L}}{\nu}, \tag{5.2}$$

where L is a characteristic length scale associated with D and $|\cdot|$ denotes some measure.

5.1.1 Weak formulation

The weak form of (5.1) is to find $\vec{u} \in \mathbf{H}^1_E(D)$ and $p \in L^2(D)$ such that

$$\nu\int_D \nabla\vec{u} : \nabla\vec{v} + \int_D (\vec{u}\cdot\nabla\vec{u})\cdot\vec{v} - \int_D p\,(\nabla\cdot\vec{v}) = \int_D \vec{f}\cdot\vec{v}, \qquad \forall\,\vec{v}\in\mathbf{H}^1_{E_0}(D), \tag{5.3a}$$

$$\int_D q\,(\nabla\cdot\vec{u}) = 0, \qquad \forall\, q\in L^2(D), \tag{5.3b}$$

where the spaces $\mathbf{H}^1_E(D), \mathbf{H}^1_{E_0}(D)$ are defined as in chapter 3 and $\nabla\vec{u} : \nabla\vec{v}$ denotes componentwise dot product. The solution of the nonlinear equations (5.3) is obtained through an iterative process where given $(\vec{u}^{(0)}, p^{(0)}) \in \mathbf{H}^1_E(D)\times L^2(D)$, a sequence $\{(\vec{u}^{(l+1)}, p^{(l+1)})\}_{l=0}^{\infty}$ of iterates satisfying (5.3) is constructed such that

$$\vec{u}^{(l+1)} = \vec{u}^{(l)} + \delta\vec{u}^{(l)}, \qquad p^{(l+1)} = p^{(l)} + \delta p^{(l)}. \tag{5.4}$$

Choosing finite dimensional subspaces $\mathbf{X}_E^h \subset \mathbf{H}_E^1(D)$, $\mathbf{X}_{E_0}^h \subset \mathbf{H}_{E_0}^1(D)$, and $M^h \subset L^2(D)$ leads to the mixed FEM formulation of (5.3).

5.1.2 Mixed FEM formulation

The mixed FEM formulation is to find $\vec{u}_h \in \mathbf{X}_E^1$ and $p_h \in M^h$ such that

$$\nu \int_D \nabla\vec{u}_h : \nabla\vec{v}_h \; + \int_D (\vec{u}_h \cdot \nabla\vec{u}_h)\cdot\vec{v}_h \; - \int_D p_h\,(\nabla\cdot\vec{v}_h) = \int_D \vec{f}\cdot\vec{v}_h, \qquad \forall\,\vec{v}_h \in \mathbf{X}_{E_0}^h, \tag{5.5a}$$

$$\int_D q_h\,(\nabla\cdot\vec{u}_h) \;= 0, \qquad \forall\, q_h \in M^h. \tag{5.5b}$$

The solution of (5.5) and hence (5.3) involves nonlinear iterations that requires solving a linearized problem at each iterative step. This aspect is discussed in the next section.

5.2 Nonlinear FEM iteration

Starting with an initial guess $(\vec{u}_h^{(0)}, p_h^{(0)}) \in \mathbf{X}_E^1 \times M^h$, the finite element analogue of (5.4) is to construct a sequence $\{(\vec{u}_h^{(l+1)}, p_h^{(l+1)})\}_{l=0}^{\infty}$ of iterates satisfying (5.5) such that [Elman et al., 2014a, pp. 344, 341]

$$\vec{u}_h^{(l+1)} = \vec{u}_h^{(l)} + \delta\vec{u}_h^{(l)}, \qquad p_h^{(l+1)} = p_h^{(l)} + \delta p_h^{(l)}. \tag{5.6}$$

Plugging (5.6) in (5.5) gives

$$D(\vec{u}_h^{(l)}, \delta\vec{u}_h^{(l)}, \vec{v}_h) \; + \nu \int_D \nabla\delta\vec{u}_h^{(l)} : \nabla\vec{v}_h \; - \int_D \delta p_h^{(l)}\,(\nabla\cdot\vec{v}_h) = R^{(l)}(\vec{v}_h), \qquad \forall\,\vec{v}_h \in \mathbf{X}_{E_0}^h, \tag{5.7a}$$

$$\int_D q_h\,(\nabla\cdot\vec{u}_h^{(l)}) \;= r^{(l)}(\vec{q}_h), \qquad \forall\, q_h \in M^h, \tag{5.7b}$$

where

$$R^{(l)}(\vec{v}_h) = \int_D \vec{f}\cdot\vec{v}_h \; - \int_D (\vec{u}_h^{(l)}\cdot\nabla\vec{u}_h^{(l)})\cdot\vec{v}_h \; - \nu\int_D \nabla\vec{u}_h^{(l)} : \nabla\vec{v}_h \; + \int_D p_h^{(l)}\,(\nabla\cdot\vec{v}_h),$$

$$r^{(l)}(\vec{q}_h) = \; - \int_D q_h\,(\nabla\cdot\vec{u}_h^{(l)}),$$

$$D(\vec{u}_h^{(l)}, \delta\vec{u}_h^{(l)}, \vec{v}_h) = \int_D (\delta\vec{u}_h^{(l)}\cdot\nabla\delta\vec{u}_h^{(l)})\cdot\vec{v}_h \; + \int_D (\delta\vec{u}_h^{(l)}\cdot\nabla\vec{u}_h^{(l)})\cdot\vec{v}_h \; + \int_D (\vec{u}_h^{(l)}\cdot\nabla\delta\vec{u}_h^{(l)})\cdot\vec{v}_h.$$

5.2.1 Newton iteration

Dropping the quadratic term $\int_D (\delta\vec{u}_h^{(l)} \cdot \nabla\delta\vec{u}_h^{(l)}) \cdot \vec{v}_h$ of D and substituting in (5.7) leads to solving a linear problem for the *Newton* correction $(\delta\vec{u}^{(l)}, \delta p^{(l)})$ at the lth iterative step. That is, $\forall\,(\vec{v}_h, q_h) \in \mathbf{X}_{E_0}^h \times M^h$, find $(\delta\vec{u}_h^{(l)}, \delta p_h^{(l)}) \in \mathbf{X}_{E_0}^h \times M^h$ such that

$$\begin{aligned} \int_D (\delta\vec{u}_h^{(l)} \cdot \nabla\vec{u}_h^{(l)}) \cdot \vec{v}_h \;+\; \int_D (\vec{u}_h^{(l)} \cdot \nabla\delta\vec{u}_h^{(l)}) \cdot \vec{v}_h \;+\; \nu \int_D \nabla\delta\vec{u}_h^{(l)} : \nabla\vec{v}_h \\ - \int_D \delta p_h^{(l)}\,(\nabla \cdot \vec{v}_h) \;=\; R^{(l)}(\vec{v}_h), \\ \int_D q_h\,(\nabla \cdot \vec{u}_h^{(l)}) \;=\; r^{(l)}(\vec{q}_h). \end{aligned} \tag{5.8}$$

5.2.2 Picard iteration

Further linearization is achieved by dropping the linear term $\int_D (\delta\vec{u}_h^{(l)} \cdot \nabla\vec{u}_h^{(l)}) \cdot \vec{v}_h$ in (5.7). This leads to solving a linear problem for the *Picard* correction $(\delta\vec{u}^{(l)}, \delta p^{(l)})$ at the lth iterative step. That is, $\forall\,(\vec{v}_h, q_h) \in \mathbf{X}_{E_0}^h \times M^h$, find $(\delta\vec{u}_h^{(l)}, \delta p_h^{(l)}) \in \mathbf{X}_{E_0}^h \times M^h$ such that

$$\begin{aligned} \int_D (\vec{u}_h^{(l)} \cdot \nabla\delta\vec{u}_h^{(l)}) \cdot \vec{v}_h \;+\; \nu \int_D \nabla\delta\vec{u}_h^{(l)} : \nabla\vec{v}_h \;-\; \int_D \delta p_h^{(l)}\,(\nabla \cdot \vec{v}_h) \;=\; R^{(l)}(\vec{v}_h), \\ \int_D q_h\,(\nabla \cdot \vec{u}_h^{(l)}) \;=\; r^{(l)}(\vec{q}_h). \end{aligned} \tag{5.9}$$

In both cases, the solution $(\delta\vec{u}_h^{(l)}, \delta p_h^{(l)})$ is used to update the next (nonlinear) solution iterate $(\vec{u}_h^{(l+1)}, p_h^{(l+1)})$ through (5.6).

5.2.3 Matrix formulation

Let $\{\vec{\phi}_j\}_{j=1}^{n_u}$ be a basis for $\mathbf{X}_{E_0}^h$. Then any $\delta\vec{u}_h^{(l)} \in \mathbf{X}_{E_0}^h$ can be expressed as

$$\delta\vec{u}_h^{(l)} = \sum_{j=1}^{n_u} \Delta u_j^{(l)} \vec{\phi}_j, \qquad \Delta u_j^{(l)} \in \mathbb{R}. \tag{5.10}$$

Also, $\{\vec{\phi}_j\}_{j=1}^{n_u}$ can be extended (loosely speaking)[1] to form a basis for $\mathbf{X}_E^h$, so that any $\vec{u}_h^{(l)} \in \mathbf{X}_E^h$ can be expanded as

$$\vec{u}_h^{(l)} \;=\; \sum_{j=1}^{n_u+n_\partial} u_j^{(l)} \vec{\phi}_j, \qquad u_j^{(l)} \in \mathbb{R}, \tag{5.11}$$

where the term $\sum_{j=n_u+1}^{n_u+n_\partial} u_j^{(l)} \vec{\phi}_j$ interpolates the boundary data on ∂D_D.

[1] $\mathbf{X}_E^h$ is not a vector space unless $\vec{w} = 0$.

Similarly, if $\{\psi_k\}_{k=1}^{n_p}$ be a basis for M^h, then any $p_h^{(l)}, \delta p_h^{(l)} \in M^h$ has an expression

$$p_h^{(l)} = \sum_{k=1}^{n_p} p_k^{(l)} \psi_k, \qquad \delta p_h^{(l)} = \sum_{k=1}^{n_p} \Delta p_k^{(l)} \psi_k, \qquad p_k^{(l)}, \Delta p_k^{(l)} \in \mathbb{R}. \tag{5.12}$$

Since $\vec{u}_h^{(l)}, p_h^{(l)}$ are known from the previous iterative step, their basis coefficients in (5.11), and (5.12) are known too.

Using (5.10), (5.11), and (5.12) in (5.8) leads to the following discrete (Newton) system of linear equations at the lth nonlinear iterative step.

$$\begin{bmatrix} \nu\mathbf{A} + \mathbf{N}^{(l)} + \mathbf{W}^{(l)} & B^T \\ B & O \end{bmatrix} \begin{bmatrix} \Delta\mathbf{u}^{(l)} \\ \Delta\mathbf{p}^{(l)} \end{bmatrix} = \begin{bmatrix} \mathbf{f}^{(l)} \\ \mathbf{g}^{(l)} \end{bmatrix}. \tag{5.13}$$

The matrices $\mathbf{A}$ and B were encountered in chapter 3. The symmetric positive-definite matrix $\mathbf{A}$ (vector-Laplacian matrix) is the block diagonal matrix with the usual FEM stiffness matrix on its diagonals and the matrix B is the divergence matrix. The matrix $\mathbf{N}^{(l)}$ is the *vector convection matrix*, the scalar version of which was introduced in chapter 4. Also, W is known as the *Newton derivative matrix*. Solution vectors $\Delta\mathbf{u}^{(l)} = [\Delta u_1^{(l)}, \ldots, \Delta u_{n_u}^{(l)}]^T \in \mathbb{R}^{n_u}$, $\Delta\mathbf{p}^{(l)} = [\Delta p_1^{(l)}, \ldots, \Delta p_{n_p}^{(l)}]^T \in \mathbb{R}^{n_p}$ and the entries of $\mathbf{A}$, B, $\mathbf{N}^{(l)}$, $\mathbf{W}^{(l)}$, $\mathbf{f}^{(l)}$, and $\mathbf{g}^{(l)}$ are given by [Elman et al., 2014a, p. 348]

$$\begin{aligned}
\mathbf{A} &= [a_{ij}] \in \mathbb{R}^{n_u \times n_u}, & a_{ij} &:= \int_D \nabla\vec{\phi}_i : \nabla\vec{\phi}_j, \\
B &= [b_{kj}] \in \mathbb{R}^{n_p \times n_u}, & b_{kj} &:= -\int_D \psi_k \, (\nabla \cdot \vec{\phi}_j), \\
\mathbf{N}^{(l)} &= [n_{ij}] \in \mathbb{R}^{n_u \times n_u}, & n_{ij} &:= \int_D (\vec{u}_h^{(l)} \cdot \nabla\vec{\phi}_j) \cdot \vec{\phi}_i, \\
\mathbf{W}^{(l)} &= [w_{ij}] \in \mathbb{R}^{n_u \times n_u}, & w_{ij} &:= \int_D (\vec{\phi}_j \cdot \nabla\vec{u}_h^{(l)}) \cdot \vec{\phi}_i, \\
\mathbf{f}^{(l)} &= [f_i] \in \mathbb{R}^{n_u}, & f_i &:= \int_D \vec{f} \cdot \vec{\phi}_i - \int_D (\vec{u}_h^{(l)} \cdot \nabla\vec{u}_h^{(l)}) \cdot \vec{\phi}_i \\
& & & \quad - \nu \int_D \nabla\vec{u}_h^{(l)} : \nabla\vec{\phi}_i + \int_D p_h^{(l)} \, (\nabla \cdot \vec{\phi}_i), \\
\mathbf{g}^{(l)} &= [g_k] \in \mathbb{R}^{n_p}, & g_k &:= \int_D \psi_k \, (\nabla \cdot \vec{u}_h^{(l)}).
\end{aligned} \tag{5.14}$$

Note the dependence of vector convection matrix, Newton derivative matrix, and right-hand-side vectors on the nonlinear iterative step.

Dropping the Newton derivative matrix in (5.13) results in the linear system arising from Picard iteration

$$\begin{bmatrix} \nu\mathbf{A} + \mathbf{N}^{(l)} & B^T \\ B & O \end{bmatrix} \begin{bmatrix} \Delta\mathbf{u}^{(l)} \\ \Delta\mathbf{p}^{(l)} \end{bmatrix} = \begin{bmatrix} \mathbf{f}^{(l)} \\ \mathbf{g}^{(l)} \end{bmatrix}. \tag{5.15}$$

In any case, the coefficient matrix in (5.13) or (5.15) is nonsymmetric.[2] Thus, Krylov solvers like GMRES, BIGSTAB(ℓ) etc., will be used for solving the associated linear systems (5.13) or (5.15). In the next section, a balanced stopping criterion is presented in GMRES for solving (5.13) or (5.15).

5.3 A balanced stopping test

Similar to chapter 3, a natural norm for measuring the errors arising from weak (and mixed FEM approximation (5.5)) approximation (5.3) is

$$\|(\vec{u},p)\|_{\mathcal{E}} := \|\nabla\vec{u}\|_2 + \|p\|_2, \qquad \forall\,(\vec{u},p)\in \mathbf{H}^1_{E_0}(D)\times L^2(D). \tag{5.16}$$

The associated vector norm $\|\cdot\|_E$ is defined as

$$\|\mathbf{e}\|_E := \sqrt{\mathbf{e}^T E\mathbf{e}} = \sqrt{\mathbf{e}_1^T\mathbf{A}\mathbf{e}_1 + \mathbf{e}_2^T Q\mathbf{e}_2}, \qquad \forall\,\mathbf{e}=[\mathbf{e}_1^T,\mathbf{e}_2^T]^T\in\mathbb{R}^{n_u+n_p}, \tag{5.17}$$

where $E := \begin{bmatrix} \mathbf{A} & O \\ O & Q \end{bmatrix}$. Here E is a symmetric positive-definite matrix and therefore $\|\cdot\|_E$ is a norm on $\mathbb{R}^{n_u+n_p}$. Also, $Q=[q_{kj}]$, $q_{kj} := \displaystyle\int_D \psi_k\psi_j$, $\forall\, k,\, j=1,\ldots,n_p$ is the pressure mass matrix defined in chapter 3.

For any $(\vec{v}_h,q_h)\in\mathbf{X}^h_{E_0}\times M^h$, $\|\cdot\|_{\mathcal{E}}$ is equivalent to $\|\cdot\|_E$ in the sense that

$$\sqrt{\mathbf{v}^T\mathbf{A}\mathbf{v}+\mathbf{q}^TQ\mathbf{q}} \;\le\; \|(\vec{v}_h,q_h)\|_{\mathcal{E}} \;\le\; \sqrt{2}\sqrt{\mathbf{v}^T\mathbf{A}\mathbf{v}+\mathbf{q}^TQ\mathbf{q}}, \tag{5.18}$$

where $\mathbf{v}$, $\mathbf{q}$ are the coordinates of $\vec{v}_h$, q_h with respect to velocity and pressure basis respectively. The case of $\|\cdot\|_{0,2}$ norm can be handled in exactly the same manner as in chapter 3.

5.3.1 Error equation

At the lth nonlinear iteration, in presence of 'tight' a posteriori error estimators $\eta^{(l_k)}$, it follows from the triangle inequality at linear iteration k that

$$\eta^{(l_k)} \simeq \eta^{(l)} + \|\mathbf{e}^{(l_k)}\|_E, \qquad k=0,1,2,\ldots. \tag{5.19}$$

If the norm $\|\mathbf{e}^{(l_k)}\|_E$ of the iteration error $\mathbf{e}^{(l_k)}$ is 'small', then it follows from (5.19) that $\{\eta^{(l_k)}\}$ converges to $\eta^{(l)}$. Here $\mathbf{e}^{(l_k)} = [(\Delta\mathbf{u}^{(l)})^T,(\Delta\mathbf{p}^{(l)})^T]^T - [(\Delta\mathbf{u}^{(l_k)})^T,(\Delta\mathbf{p}^{(l_k)})^T]^T$.

[2] A stabilization matrix similar to the Stokes equations is employed (for lower order finite elements) in place of the zero block of the coefficient matrix [Elman et al., 2014a, p. 349].

So, one would stop optimally when the algebraic error balances the total error, that is, stop at the first iteration l_{k^*} such that

$$\|\mathbf{e}^{(l_{k^*})}\|_E \leq \eta^{(l_{k^*})}. \tag{5.20}$$

The a posteriori error estimator $\eta^{(l_k)}$ is equivalent to the total error (approximation error at the kth iteration) in the sense that

$$c_1\,\eta^{(l_k)} \leq \|\nabla(\delta\vec{u}^{(l)} - \delta\vec{u}_h^{(l_k)})\|_2 + \|\delta p^{(l)} - \delta p_h^{(l_k)}\|_2 \leq C_1\,\eta^{(l_k)}, \quad \text{with } \frac{C_1}{c_1} \sim O(1). \tag{5.21}$$

At the lth iterative step $\vec{u}_h^{(l)}, p_h^{(l)}$ is known. It follows from (5.6) that $\delta\vec{u}_h^{(l)} = \vec{u}_h^{(l+1)} - \vec{u}_h^{(l)}$ and $\delta p_h^{(l)} = p_h^{(l+1)} - p_h^{(l)}$. This implies that (5.13) or (5.15) essentially solves for the basis coefficients of $(\vec{u}_h^{(l+1)}, p_h^{(l+1)})$. *Thus, essentially one can use the same a posteriori approximation error estimators to estimate approximation errors a posteriori for* $(\delta\vec{u}_h^{(l)}, \delta p_h^{(l)})$ as those for $(\vec{u}_h^{(l+1)}, p_h^{(l+1)})$.[3]

Obtaining tractable bounds on the quantity $\|\mathbf{e}^{(l_k)}\|_E$ is the goal of the next section. This will lead to a balanced stopping test for linear iteration which is discussed next.

5.3.2 Tractable bounds on algebraic error

Let $N_S^{(l)}$ denote the coefficient matrix of the (linearized) discrete Navier–Stokes system in (5.13) or (5.15) at the lth iterative step. The Euclidean norm $\|\mathbf{r}^{(l_k)}\|$ of the iteration residual $\mathbf{r}^{(l_k)} = [(\mathbf{f}^{(l)})^T, (\mathbf{g}^{(l)})^T]^T - {N_S}^{(l)}[(\Delta\mathbf{u}^{(l_k)})^T, (\Delta\mathbf{p}^{(l_k)})^T]^T$ is readily available and monotonically decreasing with iteration count k in (preconditioned) GMRES. So, upper and lower bounds on the usually unobservable quantity, that is, the error $\|\mathbf{e}^{(l_k)}\|_E^2 = (\mathbf{r}^{(l_k)})^T({N_S}^{(l)})^{-T}E({N_S}^{(l)})^{-1}\mathbf{r}^{(l_k)}$ are obtained in terms of the surrogate norm $\|\mathbf{r}^{(l_k)}\|$.

Similar to chapter 4, this involves solving for extremal Rayleigh quotient bounds for $({N_S}^{(l)})^{-T}E{N_S}^{(l)^{-1}}$. This in turn leads to solving for extremal eigenvalues θ, Θ, which are the minimum and the maximum eigenvalues respectively of the generalized eigenvalue problem for E and $({N_S}^{(l)})^T{N_S}^{(l)}$. Note that E is a symmetric positive-definite matrix. So, a Cholesky decomposition (theoretically) of the symmetric positive-definite matrix $({N_S}^{(l)})^T{N_S}^{(l)}$ converts the generalized eigenvalue problem for E and

[3]This is not a rigorous mathematical statement. A proof for this statement is an ongoing research.

$(N_S{}^{(l)})^T N_S{}^{(l)}$ into a symmetric positive-definite algebraic eigenvalue problem. Thus, θ and Θ are both real and greater than zero.

Since $\theta \leq \dfrac{(\mathbf{r}^{(l_k)})^T (N_S{}^{(l)})^{-T} E (N_S{}^{(l)})^{-1} \mathbf{r}^{(l_k)}}{(\mathbf{r}^{(l_k)})^T \mathbf{r}^{(l_k)}} \leq \Theta$, $k = 0, 1, 2, \ldots$, it follows that

$$\sqrt{\theta} \leq \frac{\|\mathbf{e}^{(0)}\|_E}{\|\mathbf{r}^{(0)}\|}, \qquad \frac{\|\mathbf{e}^{(l_k)}\|_E}{\|\mathbf{r}^{(l_k)}\|} \leq \sqrt{\Theta}. \tag{5.22}$$

Equation (5.22) leads to the following upper bounds on $\|\mathbf{e}^{(k)}\|_E$, that is

$$\frac{\|\mathbf{e}^{(l_k)}\|_E}{\|\mathbf{e}^{(0)}\|_E} \leq \sqrt{\frac{\Theta}{\theta}} \frac{\|\mathbf{r}^{(l_k)}\|}{\|\mathbf{r}^{(0)}\|} \Longleftrightarrow \|\mathbf{e}^{(l_k)}\|_E \leq \sqrt{\frac{\Theta}{\theta}} \frac{\|\mathbf{r}^{(l_k)}\|}{\|\mathbf{r}^{(0)}\|} \|\mathbf{e}^{(0)}\|_E \Longleftrightarrow \|\mathbf{e}^{(l_k)}\|_E \leq \frac{\Theta}{\sqrt{\theta}} \|\mathbf{r}^{(l_k)}\|, \tag{5.23a}$$

$$\|\mathbf{e}^{(l_k)}\|_E \leq \sqrt{\Theta}\, \|\mathbf{r}^{(l_k)}\|. \tag{5.23b}$$

5.3.3 Stopping criterion for linearized iteration

In light of (5.23a) and (5.23b), the stopping test (5.20) becomes: stop at the first iteration l_{k^*} such that

$$\sqrt{\Theta} \|\mathbf{r}^{(l_{k^*})}\| \leq \eta^{(l_{k^*})} \quad \Longleftrightarrow \quad \|\mathbf{r}^{(l_{k^*})}\| \leq \frac{1}{\sqrt{\Theta}} \eta^{(l_{k^*})}. \tag{5.24}$$

$$\frac{\Theta}{\sqrt{\theta}} \|\mathbf{r}^{(l_{k^*})}\| \leq \eta^{(l_{k^*})} \quad \Longleftrightarrow \quad \|\mathbf{r}^{(l_{k^*})}\| \leq \frac{\sqrt{\theta}}{\Theta} \eta^{(l_{k^*})}. \tag{5.25}$$

Stopping test (5.24) will be called the stronger stopping test and (5.25) will be known as the weaker stopping test. Similar to chapter 4, note that the stopping criteria derived here can be used in iterative solvers for solving preconditioned nonsymmetric linear systems as well. Also, as in chapter 4, a *crucial point* to note here is that the weaker stopping test can be used as long as the a posteriori error estimator provides a 'tight' underestimation of the true error. In case of 'tight' overestimation, the stronger stopping test should be employed.

An optimal balanced black-box stopping test can also be devised for nonlinear iteration (5.6). This is discussed in the next subsection.

5.3.4 Stopping criterion for nonlinear iteration

It follows from (5.6) that

$$\begin{aligned} & (\vec{u}_h^{(l+1)} - \vec{u}_h^{(l)}, p_h^{(l+1)} - p_h^{(l)}) = (\delta\vec{u}_h^{(l)}, \delta p_h^{(l)}) \\ \Longleftrightarrow\ & \|(\vec{u}_h^{(l+1)} - \vec{u}_h^{(l)}, p_h^{(l+1)} - p_h^{(l)})\|_{\mathcal{E}} = \|(\delta\vec{u}_h^{(l)}, \delta p_h^{(l)})\|_{\mathcal{E}} \\ \Longleftrightarrow\ & \|\nabla(\vec{u}_h^{(l+1)} - \vec{u}_h^{(l)})\|_2 + \|p_h^{(l+1)} - p_h^{(l)}\|_2 = \|\nabla\delta\vec{u}_h^{(l)}\|_2 + \|\delta p_h^{(l)}\|_2. \end{aligned} \tag{5.26}$$

Note that

$$\begin{aligned}\|\nabla(\vec{u}_h^{(l+1)} - \vec{u}_h^{(l)})\|_2 &= \|\nabla(\vec{u}_h^{(l+1)} - \vec{u}) - \nabla(\vec{u}_h^{(l)} - \vec{u})\|_2,\\ \|p_h^{(l+1)} - p_h^{(l)}\|_2 &= \|(p_h^{(l+1)} - p) - (p_h^{(l)} - p)\|_2.\end{aligned} \tag{5.27}$$

Here $(\vec{u}, p)$ is the true solution. Since norm of difference is greater than or equal to the difference of norms, it follows from (5.27) that

$$\begin{aligned}\|\nabla(\vec{u}_h^{(l+1)} - \vec{u}_h^{(l)})\|_2 \geq\ & \|\nabla(\vec{u}_h^{(l+1)} - \vec{u})\|_2 - \|\nabla(\vec{u} - \vec{u}_h^{(l)})\|_2,\\ \|p_h^{(l+1)} - p_h^{(l)}\|_2 \geq\ & \|(p_h^{(l+1)} - p)\|_2 - \|(p - p_h^{(l)})\|_2.\end{aligned} \tag{5.28}$$

Plugging (5.28) in (5.26) leads to

$$\|\nabla(\vec{u}_h^{(l+1)} - \vec{u})\|_2 + \|(p_h^{(l+1)} - p)\|_2 \leq\ (\|\nabla(\vec{u}_h^{(l)} - \vec{u})\|_2 + \|p_h^{(l)} - p\|_2) + \|\nabla\delta\vec{u}_h^{(l)}\|_2 \ +\ \|\delta p_h^{(l)}\|_2. \tag{5.29}$$

From (5.17) and (5.18) it follows that

$$\|\nabla\delta\vec{u}_h^{(l)}\|_2 \ +\ \|\delta p_h^{(l)}\|_2 \ \simeq\ \sqrt{(\Delta\mathbf{u}^{(l)})^T\mathbf{A}\Delta\mathbf{u}^{(l)} + (\Delta\mathbf{p}^{(l)})^T Q\Delta\mathbf{p}^{(l)}}. \tag{5.30}$$

In presence of 'tight' a posteriori error estimator $\eta_{\text{sol}}^{(l)}$, which is equivalent to the approximation error at the lth nonlinear iteration in the sense that

$$c_1\,\eta_{\text{sol}}^{(l)} \ \leq\ \|\nabla(\vec{u}_h^{(l)} - \vec{u})\|_2 \ +\ \|p_h^{(l)} - p\|_2 \ \leq\ C_1\,\eta_{\text{sol}}^{(l)}, \quad \text{with } \frac{C_1}{c_1} \sim O(1), \tag{5.31}$$

using (5.30) and (5.31) in (5.29) leads to

$$\eta_{\text{sol}}^{(l+1)} \ \simeq\ \eta_{\text{sol}}^{(l)} \ +\ \sqrt{(\Delta\mathbf{u}^{(l)})^T\mathbf{A}\Delta\mathbf{u}^{(l)} + (\Delta\mathbf{p}^{(l)})^T Q\Delta\mathbf{p}^{(l)}}. \tag{5.32}$$

Using the balanced stopping criterion (5.24) or (5.25) for linear iteration, $(\Delta\mathbf{u}^{(l)}, \Delta\mathbf{p}^{(l)})$ is replaced by $(\Delta\mathbf{u}^{(l_{k^*})}, \Delta\mathbf{p}^{(l_{k^*})})$[4] in (5.32) which becomes

$$\eta_{\text{sol}}^{(l+1)} \ \simeq\ \eta_{\text{sol}}^{(l)} \ +\ \sqrt{(\Delta\mathbf{u}^{(l_{k^*})})^T\mathbf{A}\Delta\mathbf{u}^{(l_{k^*})} + (\Delta\mathbf{p}^{(l_{k^*})})^T Q\Delta\mathbf{p}^{(l_{k^*})}}. \tag{5.33}$$

Note that $\{\eta_{\text{sol}}^{(l)}\}$ ultimately converges to true a posteriori approximation error estimate η_{sol}. So $\forall\, l \geq \hat{l}$ (say), $\eta_{\text{sol}}^{(l)}$ are $\eta_{\text{sol}}^{(l+1)}$ are essentially the same. Using this idea, one can optimally stop the nonlinear iteration when the contribution from linearized part (that is $\sqrt{(\Delta\mathbf{u}^{(l_{k^*})})^T\mathbf{A}\Delta\mathbf{u}^{(l_{k^*})} + (\Delta\mathbf{p}^{(l_{k^*})})^T Q\Delta\mathbf{p}^{(l_{k^*})}}$) in (5.33) is insignificant. Thus, stop the nonlinear iteration at l^* which is the smallest value of $(l+1)$ such that

$$\sqrt{(\Delta\mathbf{u}^{(l^*_{k^*})})^T\mathbf{A}\Delta\mathbf{u}^{(l^*_{k^*})} + (\Delta\mathbf{p}^{(l^*_{k^*})})^T Q\Delta\mathbf{p}^{(l^*_{k^*})}} \ \leq\ \eta_{\text{sol}}^{(l^*+1)}. \tag{5.34}$$

[4] This k^* will in general be different for different l.

The resulting algorithm NAVIER_NEWTON_GMRES is presented in Figure 5.1. Note that the coefficient matrix $N_S^{(l)}$ of (5.13) is never assembled for GMRES_Navier_balanced. Instead intelligent matrix-vector products are carried out using the structure of $N_S^{(l)}$ (see the coefficient matrix structure in (5.13)). The same is true for any choice of a preconditioner $M_S^{(l)}$. Also, a random initial guess can be used for each call of GMRES_Navier_balanced. Note that in practice, the a posteriori error estimate $\eta_{\text{sol}}^{(l+1)}$ should be computed (and hence the stopping test (5.34) be tested) periodically. The algorithm in Figure 5.1 can easily be modified to cater to this situation. The same holds true for the (linearized) balanced stopping inside GMRES_Navier_balanced.

Algorithm: NAVIER_NEWTON_GMRES
given functions GMRES_Navier_balanced, matvec$\mathbf{A}$, matvecQ, Navier_error_est

. . . . . .

solve the corresponding Stokes problem to obtain starting guess: $(\vec{u}_h^{(0)}, p_h^{(0)})$

. . . . . .

for $l = 0, 1, 2, \ldots$ until convergence do

Inner iteration (GMRES solver)

% GMRES_Navier_balanced: solves (5.13) using preconditioned GMRES with balanced stopping (5.24) or (5.25)

% Coefficient matrix ${N_S}^{(l)}$, right-hand-side $[{\mathbf{f}^{(l)}}^T, {\mathbf{g}^{(l)}}^T]^T$, preconditioner ${M_S}^{(l)}$

compute the vector of basis coefficients for $\delta\vec{u}_h^{(l)}$ and $p_h^{(l)}$:
$[{\Delta\mathbf{u}^{(l)}}^T, {\Delta\mathbf{p}^{(l)}}^T]^T$ = GMRES_Navier_balanced$({N_S}^{(l)}, [{\mathbf{f}^{(l)}}^T, {\mathbf{g}^{(l)}}^T]^T, {M_S}^{(l)})$

Outer iteration (Nonlinear solver)

update solution: $\vec{u}_h^{(l+1)} = \vec{u}_h^{(l)} + \delta\vec{u}_h^{(l)}, \qquad p_h^{(l+1)} = p_h^{(l)} + \delta p_h^{(l)}$

% Navier_error_est computes the a posteriori error estimate
compute a posteriori error estimate: $\eta_{\text{sol}}^{(l+1)}$ = Navier_error_est $(\vec{u}_h^{(l+1)}, p_h^{(l+1)})$

% matvec$\mathbf{A}(\cdot)$, matvecQ$(\cdot)$ compute the action of $\mathbf{A}$ and Q on a vector respectively.
stopping test:
if $\sqrt{(\Delta\mathbf{u}^{(l)})^T \texttt{matvec}\mathbf{A}(\Delta\mathbf{u}^{(l)}) + (\Delta\mathbf{p}^{(l)})^T \texttt{matvecQ}(\Delta\mathbf{p}^{(l)})} \leq \eta_{\text{sol}}^{(l+1)}$, convergence

enddo

Figure 5.1: The NAVIER_NEWTON_GMRES algorithm expressed in pseudo-code.

5.3.5 A posteriori error estimation

Similar to the Stokes equations in chapter 3, computation of a posteriori error estimates for the Navier–Stokes mixed FEM formulation entails solving local Poisson problems for each component of velocity [Elman et al., 2014a, p. 352 ff.]. In fact it has been stated in [Elman et al., 2014a, proposition 8.9, p. 354] that a posteriori error estimators for stabilized $\boldsymbol{Q}_1$-$\boldsymbol{P}_0$ rectangular finite elements are reliable in the sense that the global upper bound on the approximation error does not depend on the parameters of the continuous problem. Thus, results presented in the section 5.4 are based on stabilized $\boldsymbol{Q}_1$-$\boldsymbol{P}_0$ rectangular finite elements.

5.3.6 Computational logistics

At the lth nonlinear iteration, $\|\mathbf{r}^{(l_k)}\|$ is readily available as a by-product of GMRES iteration. The eigenvalues Θ and θ involved in the (linear) stopping test (5.25) are computed using MATLAB `eigs`. The a posteriori error estimators should be computed periodically for both the linear and nonlinear stopping test to minimize overall algorithmic cost. Also, a cheap but an additional cost arises in computing the matrix-vector products in the left-hand-side of the nonlinear balanced stopping test.

5.4 Computational results

Results of some computational experiments in IFISS are presented in this section as a proof-of-concept. The test problem for this purpose is the flow over a backward-facing step problem; see [Gresho et al., 1993], [Powell and Silvester, 2012]. In order to illustrate the robustness of the linear and the nonlinear balanced stopping test (5.25) and (5.34) respectively, results are presented here for various values of viscosity (hence varying Reynolds number) and grid levels (g) for $\boldsymbol{Q}_1$-$\boldsymbol{P}_0$ rectangular finite elements on $2^g \times (2^g \times 3)$ grids.

Since no stabilization for the convection term is inbuilt in IFISS for the Navier–Stokes equations, the a posteriori error estimator is expected to overestimate the true error. Thus, employing the weaker stopping test (5.24) for linear iterations might lead to premature stopping. Hence, the stronger stopping test (5.25) will be used

here. The modified pressure convection-diffusion preconditioner [Elman et al., 2014a, chapter 9] is employed as a preconditioner for all cases in the GMRES solver for solving the linear(ized) system arising at each nonlinear iterative step. Moreover, results are presented here only for the Newton iterations. However, the balanced stopping criterion for both linear and nonlinear iterations is applicable to Picard iterations as well. Also, note that the initial guess for the Newton iteration in each case is the (inbuilt) solution of the corresponding Stokes problem.

At each grid level and for various values of viscosity, a reference 'true' solution is computed. This is done by solving the test problem using Newton iteration to a tight nonlinear relative residual tolerance of `1e-12`. From this true solution, 'true' a posteriori error estimate η_{sol} is computed. Also, let the difference between the true a posteriori error estimate and the computed a posteriori error estimate at the nonlinear iteration l be denoted by $e_{\eta_{\text{sol}}}^{(l_{k^*})} := |\eta_{\text{sol}}^{(l_{k^*})} - \eta_{\text{sol}}|$.

Similarly, on each grid level, a 'true' MATLAB backslash solution is computed for linear system arising at each step of the nonlinear iteration. From this true solution, 'true' a posteriori error estimate $\eta^{(l)}$ is also computed. Also, let the difference between the true a posteriori error estimate and the computed a posteriori error estimate[5] at linear stopping iteration k be denoted by $e_{\eta}^{(l_{k^*})} := |\eta^{(l_{k^*})} - \eta^{(l)}|$. Each linear system was also solved using GMRES to a (iteration) relative residual tolerance of `1e-6` and `1e-9` for a comparison with balanced stopping GMRES solver. The same preconditioner and the same initial random vector is used in all these solvers for solving any particular linear system. Also, let ${l_k}_{\text{tol1}}$, ${l_k}_{\text{tol2}}$ denote the number of iterations needed to satisfy GMRES relative residual tolerance of `1e-6` and `1e-9` respectively.

The Navier–Stokes PDE (5.1) is defined on a L-shaped (flow over a backward-facing step) domain $D = (-1, 5) \times (-1, 1) \setminus (-1, 0] \times (-1, 0]$. Poiseuille flow profile is imposed on the inflow boundary ($x_1 = -1, 0 \leq x_2 \leq 1$), $\vec{x} = (x_1, x_2) \in D$ and zero velocity condition is imposed on the walls. Neumann boundary conditions are defined everywhere on the outflow boundary ($x_1 = 5, -1 < x_2 < 1$). The forcing term $\vec{f}$ is zero. This problem can be generated in IFISS by choosing example `2` when running the driver `navier_testproblem` [Elman et al., 2014a, p. 335]. The balanced stopping test in GMRES is implemented in IFISS function `gmres_r` while the nonlinear balanced

[5]This a posteriori approximation error estimate is for the linearized part $(\delta\vec{u}_k^{(l_k)}, \delta p_h^{(l_k)})$.

stopping test is incorporated in the function `solve_step_navier` in IFISS.

Table 5.1: Navier–Stokes test problem solved using Newton iteration on a 16×64 grid with $\nu = 1/50$.

l	$l_{k_{\text{tol1}}}$	$l_{k_{\text{tol2}}}$	l_{k^*}	$e_\eta^{(l_{k^*})}$	$e_{\eta_{\text{sol}}}^{(l_{k^*})}$	Θ	θ
1	35	45	26	4.0e-04	1.5e-01	1.6e+04	2.4e-02
2	47	59	38	3.0e-06	1.5e-04	2.7e+04	2.4e-02

Table 5.2: Navier–Stokes test problem solved using Newton iteration on a 32×96 grid with $\nu = 1/50$.

l	$l_{k_{\text{tol1}}}$	$l_{k_{\text{tol2}}}$	l_{k^*}	$e_\eta^{(l_{k^*})}$	$e_{\eta_{\text{sol}}}^{(l_{k^*})}$	Θ	θ
1	31	42	23	2.0e-04	7.2e-02	6.4e+04	8.5e-02
2	41	52	34	2.3e-07	4.5e-04	1.0e+05	8.5e-02
3	45	56	37	5.0e-09	9.1e-07	1.0e+05	8.5e-02

Table 5.3: Navier–Stokes test problem solved using Newton iteration on a 64×192 grid with $\nu = 1/50$.

l	$l_{k_{\text{tol1}}}$	$l_{k_{\text{tol2}}}$	l_{k^*}	$e_\eta^{(l_{k^*})}$	$e_{\eta_{\text{sol}}}^{(l_{k^*})}$	Θ	θ
1	29	39	23	3.6e-05	3.1e-02	2.6e+05	2.2e-01
2	38	48	33	7.1e-08	3.3e-04	4.4e+05	2.2e-01
3	42	53	36	1.3e-09	5.2e-07	4.0e+05	2.1e-01

From Tables 5.1, 5.2, 5.3, 5.4, 5.5, and 5.6, a comparison of $l_{k_{\text{tol1}}}$, $l_{k_{\text{tol2}}}$ numbers with the corresponding l_{k^*} values shows that employing the linear stopping test (5.25) leads to savings in iteration counts. In each table at the lth Newton iteration and at the linear balanced stopping iteration l_{k^*}, $e_\eta^{(l_{k^*})}$ columns show that the preconditioned GMRES solution of the linearized part has converged with some accuracy to the true linearized solution. In other words, $\{\eta^{(l_k)}\}$ has converged with some accuracy to true $\eta^{(l)}$. At the nonlinear balanced stopping iteration l^*, $e_{\eta_{\text{sol}}}^{(l_{k^*})}$ columns exhibit convergence with some accuracy of $\{\eta_{\text{sol}}^{(l_{k^*})}\}$ to the true a posteriori approximation error estimate η_{sol}.

The eigenvalues Θ, θ used in the linear stopping criterion are also tabulated in these tables. These numbers exhibit some structure thereby suggesting that there might be an expression for these quantities in terms of the parameters of the problem. However, this aspect has not been investigated in this thesis.

Table 5.4: Navier–Stokes test problem solved using Newton iteration on a 16×64 grid with $\nu = 1/100$.

l	$l_{k_{\text{tol1}}}$	$l_{k_{\text{tol2}}}$	l_{k^*}	$e_\eta^{(l_{k^*})}$	$e_{\eta_{\text{sol}}}^{(l_{k^*})}$	Θ	θ
1	48	62	40	4.5e-04	1.1e+00	5.5e+04	6.3e-03
2	79	96	73	8.4e-07	3.8e-02	2.6e+05	6.3e-03

Table 5.5: Navier–Stokes test problem solved using Newton iteration on a 32×96 grid with $\nu = 1/100$.

l	$l_{k_{\text{tol1}}}$	$l_{k_{\text{tol2}}}$	l_{k^*}	$e_\eta^{(l_{k^*})}$	$e_{\eta_{\text{sol}}}^{(l_{k^*})}$	Θ	θ
1	40	54	33	8.3e-05	5.8e-01	2.2e+05	2.4e-02
2	62	77	59	3.8e-08	1.4e-02	1.2e+06	2.4e-02
3	66	81	62	4.0e-08	5.6e-04	7.4e+05	2.4e-02

Table 5.6: Navier–Stokes test problem solved using Newton iteration on a 64×192 grid with $\nu = 1/100$.

l	$l_{k_{\text{tol1}}}$	$l_{k_{\text{tol2}}}$	l_{k^*}	$e_\eta^{(l_{k^*})}$	$e_{\eta_{\text{sol}}}^{(l_{k^*})}$	Θ	θ
1	35	46	46	2.0e-05	2.9e-01	9.0e+05	8.5e-02
2	53	66	52	2.0e-08	7.1e-03	4.9e+06	8.5e-02
3	55	68	52	1.8e-08	2.2e-04	3.0e+06	8.5e-02
4	59	73	55	8.2e-10	8.4e-07	2.6e+06	8.5e-02

Evolution of errors with iteration number are plotted in Figure 5.2 at 4th Newton iteration on 64×192 grid for $\nu = 1/100$. On the plot for linear iteration observe that at the balanced linear stopping iteration l_{k^*}, the red curve for $\eta^{(l_k)}$ converges with some accuracy to the black line for $\eta^{(l)}$. This convergence is further illustrated by continuing for 9 more iterations after balanced linear stopping. Note that $\{\eta^{(l_k)}\}$ converges to $\eta^{(l)}$ when $\|\mathbf{e}^{(l_k)}\|_E$ (cyan) curve goes below the (black) line for $\eta^{(l)}$. However, as mentioned in previous chapters, iteration error $\mathbf{e}^{(l_k)}$ is rarely known a priori. Also, on the plot for Newton iteration (right) notice that at the balanced stopping nonlinear iteration number four, the curve for $\eta_{\text{sol}}^{(l_{k^*})}$ converges with some accuracy to the black line for the true a posteriori approximation error estimate η_{sol}. This convergence is further illustrated by continuing for 2 more iterations after balanced nonlinear stopping.

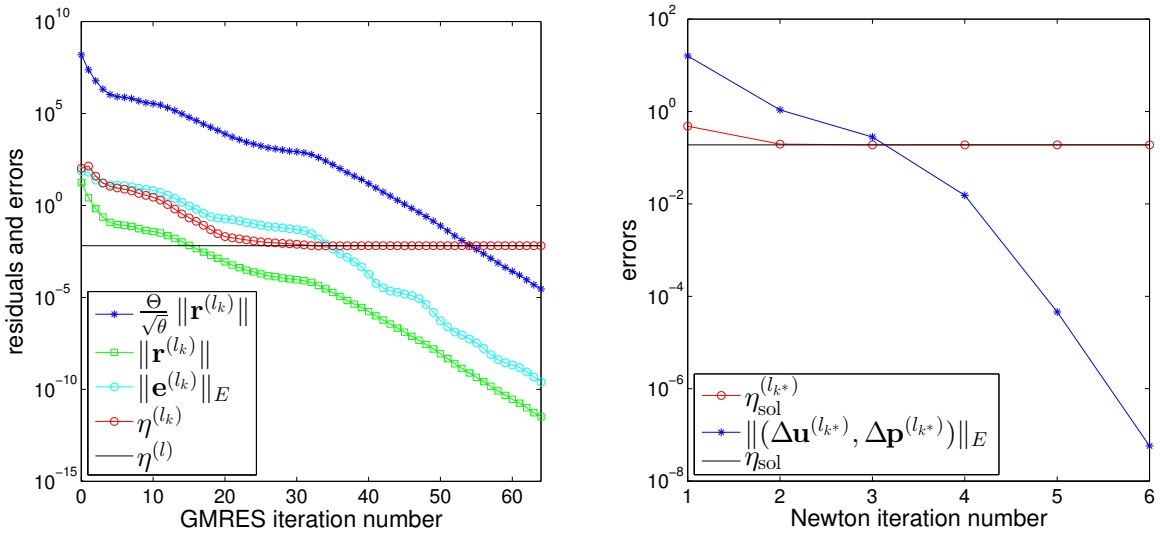

Figure 5.2: Errors vs iteration number for Navier–Stokes test problem on a 64×192 grid with $\nu = 1/100$ for Newton iteration (right) and linear (GMRES) iteration (left) at $l = 4$th Newton iteration.

5.5 Summary

An optimal balanced black-box stopping criterion for linear (GMRES, BICGSTAB(ℓ), TFQMR etc.) iterations associated with Newton or Picard nonlinear iterations have been devised for solving Navier–Stokes equations. Moreover, an optimal balanced black-box stopping criterion for nonlinear (Newton or Picard) iterations has also been formulated; see [Syamsudhuha and Silvester, 2003] for alternative nonlinear iteration stopping strategies. Using such balanced stopping tests may not only save unnecessary computational work of the employed linear and nonlinear iterative solver but also rules out premature stopping.

The balanced stopping strategies presented here are quite generic. They can be suitably modified to cater for varied linear and nonlinear iterative procedures arising from FEM approximation of a (stochastic) PDE provided cheap and tight a posteriori approximation error estimators are available along with cheap tractable bounds on the relevant unobservable errors.

Chapter 6

Open questions

Every solution strategy in science gives rise to new queries. Following are some of the questions arising from the research material presented in this thesis.

- How to obtain a cheap and tight a priori or posteriori FEM approximation error estimate for the various PDEs considered in this thesis?

Note that computing the a posteriori error estimates is the major additional cost in employing an iterative method with an optimal balanced black-box stopping strategy. Thus, a procedure for obtaining a cheap and tight a priori or a posteriori approximation error estimates will help to reduce this algorithmic cost.

- How to obtain a cheap and tight posteriori FEM approximation error estimate for stochastic convection-diffusion equations, stochastic Stokes equations, stochastic Navier–Stokes equations?

Using an iterative solver with an optimal balanced black-box stopping strategy is advantageous in practice when it is applied for solving discrete systems arising from stochastic PDEs.

- How to obtain cheaply the eigenvalues involved in the balanced stopping tests in iterative solvers for solving nonsymmetric linear systems?

By investigating the relationship between the eigenvalues of the discrete problem and the eigenvalues of the continuous problem it must be possible in some way to obtain an estimate of the eigenvalues of the discrete problem cheaply.

- Is it possible to extend the idea of error balancing and tractable bounds for the unobservable errors to obtain an optimal balanced black-box strategy for

multilevel Monte-Carlo methods, optimization problems involving solution of a linear system at each iterative step (PDE constrained optimization) etc.?

Bibliography

D. J. Acheson. Elementary Fluid Dynamics. Oxford University Press, UK, 1990. First Edition.

K. Ahuja, P. Benner, E. de Sturler, and L. Feng. Recycling BiCGSTAB with an application to parametric model order reduction. SIAM J. Sci. Comput., 37(5): S429–S446, 2015. https://doi.org/10.1137/140972433.

M. Ainsworth and J. Oden. A posteriori error estimators for Stokes and Oseen equations. SIAM J. Numer. Anal., 34(1):228–245, 1997. https://doi.org/10.1137/S0036142994264092.

A. Amritkar, E. de Sturler, K. Swirydowicz, D. Tafti, and K. Ahuja. Recycling Krylov subspaces for CFD applications and a new hybrid recycling solver. Journal of Comp. Phy., 303:222–237, 2015. https://doi.org/10.1016/j.jcp.2015.09.040.

M. Arioli. A stopping criterion for the conjugate gradient algorithm in a finite element method framework. Numer. Math., 97:1–24, 2004. https://doi.org/10.1007/s00211-003-0500-y.

M. Arioli and D. Loghin. Stopping criteria for mixed finite element problems. Elec. Trans. on Numer. Anal., 29:178–192, 2008. https://etna.ricam.oeaw.ac.at/vol.29.2007-2008/pp178-192.dir/pp178-192.pdf.

M. Arioli, D. Loghin, and A. J. Wathen. Stopping criteria for iterations in finite element methods. Numer. Math., 99(3):381–410, 2005. https://doi.org/10.1007/s00211-004-0568-z.

W. E. Arnoldi. The principle of minimized iterations in the solution of the matrix eigenvalue problem. Quarterly of Applied Mathematics, 9(1):17–29, 1951. https://doi.org/10.1090/qam/42792.

O. Axelsson. Iterative Solution Methods. Cambridge University Press, UK, 1994. First Edition.

I. Babuška, R. Tempone, and G. E. Zouraris. Galerkin finite element approximations of stochastic elliptic partial differential equations. SIAM J. Numer. Anal., 42(2): 800–825, 2004. https://doi.org/10.1137/S0036142902418680.

I. Babuška, F. Nobile, and R. Tempone. A stochastic collocation method for elliptic partial differential equations with random input data. SIAM J. Numer. Anal., 45 (3):1005–1034, 2007. https://doi.org/10.1137/050645142.

Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst. Templates for the solution of Algebraic Eigenvalue Problems: A Practical guide. SIAM, USA, 2000.

R. Barret, M. Berry, T. Chan, J. Demmel, J. Donato, J. Dongarra, V. Eijkhout, R. Pozo, C. Romine, and H. A. Van Der Vorst. Templates for the Solution of Linear Systems: Building Blocks for Iterative Methods. SIAM, USA, 1987.

M. Benzi, G. H. Golub, and J. Liesen. Numerical solution of saddle point problems. Acta Numerica, 14:1–137, 2005. https://doi.org/10.1017/S0962492904000212.

A. Bespalov and D. Silvester. Efficient adaptive stochastic Galerkin methods for parametric operator equations. SIAM J. Sci. Comput., 38(4):A2118–A2140, 2016. https://doi.org/10.1137/15M1027048.

A. Bespalov, C. Powell, and D. Silvester. Energy norm a posteriori error estimation for parametric operator equations. SIAM J. Sci. Comput., 36(2):A339–A363, 2014. https://doi.org/10.1137/130916849.

S. C. Brenner and L. R. Scott. The Mathematical Theory of Finite Element Methods. Springer, USA, 2008. Third Edition.

F. Brezzi, D. Marini, and A. Russo. Applications of the pseudo residual-free bubbles to the stabilization of convection-diffusion problems. Comput. Methods Appl. Mech. Engrg., 166(1-2):51–63, 1998.
https://doi.org/10.1016/S0045-7825(98)00082-6.

T. Butler, P. Constantine, and T. Wildey. A posteriori error analysis of parameterized linear systems using spectral methods. SIAM J. Matrix Anal. Appl., 33(1):195–209, 2012. https://doi.org/10.1137/110840522.

E. V. Chizhonkov and M. A. Olshanskii. On the domain geometry dependence of the LBB condition. ESAIM: M2AN, 34(5):935–951, 2000. https://doi.org/10.1051/m2an:2000110.

P. G. Constantine, D. F. Gleich, and G. Iaccarino. Spectral methods for parameterized matrix equations. SIAM J. Matrix Anal. Appl., 31(5):2681–2699, 2010. https://doi.org/10.1137/090755965.

T. A. Davies, S. Rajamanickam, and W. M. Sid-Lakhdar. A survey of direct methods for sparse linear systems. Acta Numerica, 25:383–566, 2016. https://doi.org/10.1017/S0962492916000076.

D. Day. An efficient implementation of the nonsymmetric Lanczos algorithm. SIAM J. Matrix Anal. Appl., 18(3):566–589, 1997. https://doi.org/10.1137/S0895479895292503.

M. K. Deb, I. M. Babuška, and J. T. Oden. Solution of stochastic partial differential equations using Galerkin finite element techniques. Comput. Methods Appl. Mech. Engrg., 190(48):6359–6372, 2001. https://doi.org/10.1016/S0045-7825(01)00237-7.

M. Eiermann, O. G. Ernst, and E. Ullmann. Computational aspects of the stochastic finite element method. Comput. Visual Sci, 10(1):3–15, 2007. https://doi.org/10.1007/s00791-006-0047-4.

M. Eigel, C. J. Gittelson, C. Schwab, and E. Zander. Adaptive stochastic Galerkin FEM. Comput. Methods Appl. Mech. Engrg., 270:247–269, 2014. https://doi.org/10.1016/j.cma.2013.11.015.

H. Elman, D. Silvester, and A. Wathen. Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics. Oxford University Press, UK, 2014a. Second Edition.

H. C. Elman and D. J. Silvester. Collocation methods for exploring perturbations in linear stability analysis. arxiv:1703.04796, 2017. https://arxiv.org/abs/1703.04796.

H. C. Elman, O. G. Ernst, D. P. O' Leary, and M. Stewart. Efficient iterative algorithms for the stochastic finite element method with application to acoustic scattering. Comput. Methods Appl. Mech. Engrg., 194(9–11):1037–1055, 2005. https://doi.org/10.1016/j.cma.2004.06.028.

H. C. Elman, A. Ramage, and D. J. Silvester. IFISS: A computational laboratory for investigating incompressible flow problems. SIAM Review, 56(2):261–273, 2014b. https://doi.org/10.1137/120891393.

M. Embree. The tortoise and the hare restart GMRES. SIAM Review, 45(2):259–266, 2003. https://doi.org/10.1137/S003614450139961.

K. Eriksson, D. Estep, P. Hansbo, and C. Johnson. Computational Differential Equations. Cambridge University Press, USA, 1996. First Edition.

V. Faber and T. Manteuffel. Necessary and sufficient conditions for the existence of a conjugate gradient method. SIAM J. Numer. Anal., 21(2):352–362, 1984. https://doi.org/10.1137/0721026.

V. Faber and T.A. Manteuffel. Orthogonal error methods. SIAM J. Numer. Anal., 24 (1):170–187, 1987. https://doi.org/10.1137/0724014.

B. Fischer. Polynomial Based Iteration Methods for Symmetric Linear Systems. SIAM, USA, 2011. Second Edition.

R. Fletcher. Conjugate gradient methods for indefinite systems. Springer-Verlag, 506: 73–89, 1976. https://doi.org/10.1007/BFb0080116.

R. W. Freund. A transpose-free quasi-minimal residual algorithm for non-hermitian linear systems. SIAM J. Sci. Comput., 14(2):470–482, 1993. https://doi.org/10.1137/0914029.

R. W. Freund and N. M. Nachtigal. QMR: a quasi-minimal residual method for non-hermitian linear systems. Numer. Math., 60(1):315–339, 1991. https://doi.org/10.1007/BF01385726.

R. W. Freund and N. M. Nachtigal. An implementation of the QMR method based on coupled two-term recurrences. SIAM J. Sci. Comput., 15(2):313–337, 1994. https://doi.org/10.1137/0915022.

R. W. Freund, M. H. Gutknecht, and N. M. Nachtigal. An implementation of the look-ahead Lanczos algorithm for non-hermitian matrices. SIAM J. Sci. Comput., 14(1):137–158, 1993. https://doi.org/10.1137/0914009.

W. Gautschi. Orthogonal Polynomials Computation and Approximation. Oxford University Press, UK, 2004. First Edition.

R. G. Ghanem and R. M. Kruger. Numerical solution of spectral stochastic finite element systems. Comput. Methods Appl. Mech. Engrg., 129(3):289–303, 1996. https://doi.org/10.1016/0045-7825(95)00909-4.

R. G. Ghanem and P. D. Spanos. Stochastic finite elements: a spectral approach. Springer-Verlag, USA, 1991. First Edition.

G. H. Golub and C. F. Van Loan. Matrix Computations. The John Hopkins University Press, USA, 2013. Fourth Edition.

A. D. Gordon. Solving PDEs with random data by stochastic collocation. The University of Manchester, UK, 2013. PhD Thesis. https://www.escholar.manchester.ac.uk/uk-ac-man-scw:184261.

A. Greenbaum. Iterative Methods for Solving Linear Systems. SIAM, USA, 1997. First Edition.

P. M. Gresho, D. K. Gartling, J. R. Torczynski, K. A. Cliffe, K. H. Winters, T. J. Garratt, A. Spence, and J. W. Goodrich. Is the steady viscous incompressible two-dimensional flow over a backward-facing step at Re = 800 stable? Int. J. Numer. Methods Fluids, 17(6):501–541, 1993. https://doi.org/10.1002/fld.1650170605.

D. F. Griffiths and D. J. Higham. Learning LaTeX. SIAM, USA, 2016. Second Edition.

M. D. Gunzburger, C. G. Webster, and G. Zhang. Stochastic finite element methods for partial differential equations with random input data. Acta Numerica, 23:521–650, 2014. https://doi.org/10.1017/S0962492914000075.

M. Hestenes and E. Stiefel. Methods of conjugate gradients for solving linear systems. Journal of Research of National Bureau of Standards, 49(6):409–436, 1952. http://nvlpubs.nist.gov/nistpubs/jres/049/jresv49n6p409_A1b.pdf.

D. J. Higham and N. J. Higham. MATLAB Guide. SIAM, USA, 2017. Third Edition.

N. J. Higham. Accuracy and Stability of Numerical Algorithms. SIAM, USA, 2002. Second Edition.

R. A. Horn and C. R. Johnson. Matrix Analysis. Cambridge University Press, USA, 2013. Second Edition.

T. J. R. Hughes and A. Brooks. A multi-dimensional upwind scheme with no crosswind diffusion. In: Finite Element Methods for Convection Dominated Flows, ASME Winter Annual Meeting, T. Hughes (Ed.), New York, USA, 34:19–35, 1979. https://www.researchgate.net/publication/297681092.

P. Jiránek, Z. Strakos, and M. Vohralík. A posteriori error estimates including algebraic error and stopping criteria for iterative solvers. SIAM J. Sci. Comput., 32(3):1567–1590, 2010. https://doi.org/10.1137/08073706X.

D. Kay and D. Silvester. A posteriori error estimation for stabilized mixed approximations of the Stokes equations. SIAM J. Sci. Comput., 24(1):1321–1336, 1999. https://doi.org/10.1137/S1064827598333715.

A. Klimke and B. Wohlmuth. Algorithm 847: spinterp: piecewise multilinear hierarchical sparse grid interpolation in MATLAB. ACM Trans. Math. Softw., 31 (4):561–579, 2005. https://doi.org/10.1145/1114268.1114275.

C. Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. J. Research Nat. Bur. Standards, 45(4):255–282, 1950. https://doi.org/10.6028/jres.045.026.

J. Liesen. Computable convergence bounds for GMRES. SIAM J. Matrix Anal. Appl., 21(3):882–903, 2000. https://doi.org/10.1137/S0895479898341669.

J. Liesen and Z. Strakos. GMRES convergence analysis for a convection-diffusion model problem. SIAM J. Sci. Comput., 26(6):1989–2009, 2005. https://doi.org/10.1137/S1064827503430746.

J. Liesen and Z. Strakoš. Krylov Subspace Methods, Principles and Analysis. Oxford University Press, UK, 2012. First Edition.

G. J. Lord, C. E. Powell, and T. Shardlow. An Introduction to Computational Stochastic PDEs. Cambridge University Press, UK, 2014. First Edition.

K. Mardal and R. Winther. Preconditioning discretizations of systems of partial differential equations. Numerical Linear Algebra with Applications, 18(1):1–40, 2011. https://doi.org/10.1002/nla.716.

G. Meurant. Necessary and sufficient conditions for GMRES complete and partial stagnation. Applied Numerical Mathematics, 75:100–107, 2014. https://doi.org/10.1016/j.apnum.2013.02.008.

J. T. Oden and L. F. Demkowicz. Applied Functional Analysis. CRC Press, USA, 1996. First Edition.

C. C. Paige and M. A. Saunders. Solution of sparse indefinite systems of linear equations. SIAM J. Numer. Anal., 12(4):617–629, 1975. https://doi.org/10.1137/0712047.

B. N. Parlett. The Symmetric Eigenvalue Problem. SIAM, USA, 1998.

B. N. Parlett, D. R. Taylor, and Z. A. Liu. A look-ahead Lanczos algorithm for unsymmetric matrices. Mathematics of Computation, 44(169):105–124, 1985. https://doi.org/10.1090/S0025-5718-1985-0771034-2.

D. A. Di Pietro, E. Flauraud, M. Vohralík, and S. Yousef. A posteriori error estimates, stopping criteria, and adaptivity for multiphase compositional refinement for thermal multiphase compositional flows in porous media. Journal of Comp. Phy., 276:163–187, 2014a. https://doi.org/10.1016/j.jcp.2014.06.061.

D. A. Di Pietro, M. Vohralík, and S. Yousef. An a posteriori-based, fully adaptive algorithm with adaptive stopping criteria and mesh refinement for thermal multiphase compositional flows in porous media. Comput. Math. Appl., 68(12 B): 2331–2347, 2014b. https://doi.org/10.1016/j.camwa.2014.08.008.

C. E. Powell and H. C. Elman. Block-diagonal preconditioning for spectral stochastic finite-element systems. IMA J. Numer. Anal., 29(2):350–375, 2009. https://doi.org/10.1093/imanum/drn014.

C. E. Powell and D. J. Silvester. Preconditioning steady-state Navier–Stokes equations with random data. SIAM J. Sci. Comput., 34(5):A2482–A2506, 2012. https://doi.org/10.1137/120870578.

C. E. Powell, D. Silvester, and V. Simoncini. An efficient reduced basis solver for stochastic Galerkin matrix equations. SIAM J. Sci. Comput., 39(1):A141–A163, 2017. https://doi.org/10.1137/15M1032399.

Y. Saad. The Lanczos biorthogonalization algorithm and other oblique projection methods for solving large unsymmetric systems. SIAM J. Numer. Anal., 19(3): 485–506, 1982. https://doi.org/10.1137/0719031.

Y. Saad. Iterative Methods for Sparse Linear Systems. SIAM, USA, 2003. Second Edition.

Y. Saad and M. Schultz. A generalized minimal residual algorithm for solving nonsymmetric linear systems. SIAM J. Sci. Comput., 7(3):856–869, 1986. https://doi.org/10.1137/0907058.

G. I. Shishkin. Methods of constructing grid approximations for singularly perturbed boundary-value problems. Condensing grid methods. Russian J. Numer. Anal. Math. Modelling, 7(6):537–562, 1992. https://doi.org/10.1515/rnam.1992.7.6.537.

D. Silvester and Pranjal. An optimal solver for linear systems arising from stochastic FEM approximation of diffusion equations with random coefficients. SIAM/ASA J. Uncertainty Quantification, 4(1):298–311, 2016. https://doi.org/10.1137/15M1017740.

D. J. Silvester and V. Simoncini. An optimal iterative solver for symmetric indefinite systems stemming from mixed approximation. ACM Trans. Math. Softw., 37(4), 2011. https://doi.org/10.1145/1916461.1916466.

D. J. Silvester, A. Bespalov, and C. E. Powell. S-IFISS version 1.01, March 2015. https://www.manchester.ac.uk/ifiss/s-ifiss1.0.tar.gz.

G. L. G. Sleijpen and D. R. Fokkema. BICGSTAB(L) for linear equations involving unsymmetric matrices with complex spectrum. Elec. Trans. Numer. Anal., 1:11–32, 1993. https://etna.mcs.kent.edu/vol.1.1993/pp11-32.dir/pp11-32.pdf.

D. Sloan, E. Süli, and S. Vandewalle. Partial Differential Equations, Volume 7 in Numerical Analysis 2000. Elsevier, USA, 2001. First Edition.

R. C. Smith. Uncertainty Quantification: Theory, Implementation, and Applications. SIAM, USA, 2014. First Edition.

P. Sonneveld. CGS, a fast Lanczos-type solver for nonsymmetric linear systems. SIAM J. Sci. Stat. Comput., 10(1):36–52, 1989. https://doi.org/10.1137/0910004.

Z. Strakoš and P. Tichý. Error estimation in preconditioned conjugate gradients. BIT Numerical Mathematics, 45(4):789–817, 2005. https://doi.org/10.1007/s10543-005-0032-1.

W. Strauss. Partial Differential Equations: An Introduction. John Wiley & Sons, Inc, USA, 2008. Second Edition.

Syamsudhuha and D. J. Silvester. Efficient solution of the steady-state Navier–Stokes equations using a multigrid preconditioned Newton–Krylov method. Int. J. Numer. Methods Fluids, 43(12):1407–1427, 2003. https://doi.org/10.1002/fld.627.

R. Verfürth. A posteriori error estimators for convection-diffusion equations. Numer. Math., 80(4):641–663, 1998. https://doi.org/10.1007/s002110050381.

A. Wathen. Preconditioning and convergence in the right norm. Int. J. Comput. Math., 84(8):1199–1209, 2007. https://doi.org/10.1080/00207160701355961.

A. J. Wathen. Preconditioning. Acta Numerica, 24:329–376, 2015. https://doi.org/10.1017/S0962492915000021.

N. Wiener. The homogeneous chaos. Amer. J. Math., 60(4):897–936, 1938. https://doi.org/10.2307/2371268.

C. T. Wu. On the implementation of an accurate and efficient solver for convection-diffusion equations. University of Maryland, USA, 2003. PhD thesis. https://drum.lib.umd.edu/handle/1903/32.

Appendix A

Some definitions and theorems

A.1 Linear algebra concepts

Definition A.1.1 (Vector space)**.** [Horn and Johnson, 2013, p. 1–2]
Let V be a nonempty set and F be a field. Then V is called a vector space over F—denoted by $V(F)$—if $\forall\, x, y, z \in V$ and $\forall\, a, b \in F$, the following conditions are satisfied.

(i) $x + y \in V$.

(ii) $x + y = y + x$.

(iii) $(x + y) + z = x + (y + z)$.

(iv) $x + 0 = x$. The element 0 is called the *additive identity*.

(v) $x + (-x) = 0$. That is for every element $x \in V, \exists$ an element $-x \in V$ known as the *inverse* of x.

(vi) $a \odot x \in F$.

(vii) $(a + b) \odot x = a \odot x + b \odot x$.

(viii) $a \odot (x + y) = a \odot x + a \odot y$.

(ix) $(a \odot b) \odot x = a \odot (b \odot x)$.

(x) $e \odot x = x$. The element $e \in F$ is known as the multiplicative identity.

The binary operation $+ : V \times V \to V$ is known as vector addition and the binary operation $\odot : F \times V \to V$ is known as scalar multiplication and is usually denoted

multiplicatively, that is, $a \odot b$ by ab. Common examples of vector spaces are $\mathbb{R}^d(\mathbb{R})$ and $\mathbb{C}^d(\mathbb{R}), d = 1, 2, 3, \ldots$.

Definition A.1.2 (Vector subspace)**.** [Horn and Johnson, 2013, p. 2]
A subset W of vector space $V(F)$ is called a (vector) subspace of V if W itself is a vector space over F with respect to the vector addition and scalar multiplication as defined on V.

Definition A.1.3 (Basis and dimension)**.** [Horn and Johnson, 2013, p. 2–3]
A linearly independent set $B \subset V$ is called a basis for vector space $V(F)$ if it also spans V, that is, $L(B) = V$. Here $L(B)$ is the set of all possible finite linear combinations of the elements of B.

The number of elements in any basis of $V(F)$ is known as the dimension of vector space $V(F)$. If dimension is finite then $V(F)$ is called finite dimensional else $V(F)$ is called infinite dimensional.

Definition A.1.4 (Inner-product)**.** [Horn and Johnson, 2013, p. 315]
Let $V(F)$ be a vector space. A function $\langle \cdot, \cdot \rangle : V \times V \to F$ is an inner-product if $\forall\, x, y, z \in V$ and $\forall\, c \in F$, the following properties are satisfied.

(i) $\langle x, x \rangle \geq 0$,

(ii) $\langle x, x \rangle = 0$ iff $x = 0$,

(iii) $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$,

(iv) $\langle cx, y \rangle = c\langle x, y \rangle$,

(v) $\langle x, y \rangle = \overline{\langle y, x \rangle}$.

Note that $\overline{\langle \cdot, \cdot \rangle}$ represents the complex conjugate of $\langle \cdot, \cdot \rangle$.

Inner-products essentially define a rule for multiplication of elements of V with each other which is not provided by the axioms of vector space.

An Inner-product induces a norm $\| \cdot \|$ on $V(F)$, that is, $\|x\| := \sqrt{\langle x, x \rangle}$, for any $x \in V$. Thus, inner-product spaces are always endowed with at least one candidate for measuring size of vectors. Also, the space $(V, \| \cdot \|)$ is the associated normed linear space.

Definition A.1.5 (Orthogonality)**.** [Horn and Johnson, 2013, p. 15]
Any two vectors $x, y \in V(F)$ are called orthogonal with respect to an inner product $\langle \cdot, \cdot \rangle$ defined on V iff $\langle x, y \rangle = 0$. In addition if $\langle x, x \rangle = 1$ and $\langle y, y \rangle = 1$, then the vectors x, y are called orthonormal.

Note that there exists a basis of orthonormal vectors for every finite dimensional vector space. This concept is very useful in practice where the entries of FEM matrices are usually the inner products of orthonormal basis functions of the underlying FEM trial and test spaces. This results in sparse FEM matrices whose structure can be utilized efficiently by iterative solvers.

Definition A.1.6 (Hilbert space)**.** [Brenner and Scott, 2008, p. 51]
Let $(V, \langle \cdot, \cdot \rangle)$ be an inner-product space. If the associated normed linear space $(V, \| \cdot \|)$ is complete, that is, every Cauchy sequence in V converges with respect to the norm $\| \cdot \|$ in V, then $(V, \langle \cdot, \cdot \rangle)$ is called a Hilbert space.

Definition A.1.7 (Boundedness and coerciveness)**.** [Brenner and Scott, 2008, p. 57]
A bilinear form $a(\cdot, \cdot)$ on a normed vector space $(V, \| \cdot \|_V)$ is said to be *bounded or continuous* iff $\exists\, C < \infty$ such that $|a(u, v)| \leq C \|u\|_V \|v\|_V, \forall\, u, v \in V$ and *coercive* on V iff $\exists \alpha > 0$ such that $a(v, v) \geq \alpha \|v\|_V^2, \forall\, v \in V$.

A.2 Some special types of matrices

Let $M_{m,n}$ (m, n are positive integers) denote the space of all $m \times n$ matrices with entries over the field of complex numbers $\mathbb{C}$. Then with respect to matrix addition as vector addition and scalar matrix multiplication as scalar multiplication, $M_{m,n}(\mathbb{C})$ is a vector space.

Definition A.2.1 (Symmetric matrix)**.** [Horn and Johnson, 2013, p. 7]
A matrix $H \in M_{n,n}(\mathbb{C})$ is symmetric iff $H = H^T$. Here H^T denotes the transpose of matrix H.

Definition A.2.2 (Nonsymmetric matrix)**.** A matrix $H \in M_{n,n}(\mathbb{C})$ is nonsymmetric iff H is not symmetric.

Definition A.2.3 (Skew-symmetric matrix)**.** [Horn and Johnson, 2013, p. 7]
A matrix $H \in M_{n,n}(\mathbb{C})$ is skew-symmetric iff $H = -H^T$.

Definition A.2.4 (Hermitian matrix)**.** [Horn and Johnson, 2013, p. 7]
A matrix $H \in M_{n,n}(\mathbb{C})$ is Hermitian iff $H = H^*$. Here H^* denotes the conjugate transpose of matrix H.

Note that for matrices with real entries, the Hermitian property is the same as the symmetric property.

Definition A.2.5 (Positive-(semi)definite matrix)**.** [Horn and Johnson, 2013, p. 429]
A Hermitian matrix $H \in M_{n,n}(\mathbb{C})$ is positive-semidefinite iff $\mathbf{x}^* H \mathbf{x} \geq 0, \ \forall\, x \neq 0 \in \mathbb{C}^n$.
A Hermitian matrix $H \in M_{n,n}(\mathbb{C})$ is positive-definite iff $\mathbf{x}^* H \mathbf{x} > 0, \ \forall\, x \neq 0 \in \mathbb{C}^n$.

Definition A.2.6 (Negative-(semi)definite matrix)**.** [Horn and Johnson, 2013, p. 429]
A Hermitian matrix $H \in M_{n,n}(\mathbb{C})$ is negative-semidefinite iff $\mathbf{x}^* H \mathbf{x} \leq 0, \ \forall\, x \neq 0 \in \mathbb{C}^n$.
A Hermitian matrix $H \in M_{n,n}(\mathbb{C})$ is negative-definite iff $\mathbf{x}^* H \mathbf{x} < 0, \ \forall\, x \neq 0 \in \mathbb{C}^n$.

Definition A.2.7 (Indefinite matrix)**.** [Horn and Johnson, 2013, p. 429]
A matrix $H \in M_{n,n}(\mathbb{C})$ is indefinite iff H is neither positive-semidefinite nor negative-semidefinite.

Definition A.2.8 (Eigenvalues and eigenvectors)**.** [Horn and Johnson, 2013, p. 44]
Let $H \in M_{n,n}(\mathbb{C})$. A scalar $\lambda \in \mathbb{C}$ and a nonzero vector $\mathbf{x} \in \mathbb{C}^{\mathbf{n}}$ is called a eigenvalue and eigenvector respectively of H if it satisfies $H\mathbf{x} = \lambda \mathbf{x}$.

Definition A.2.9 (Similarity)**.** [Horn and Johnson, 2013, p. 58]
Two matrices $H_1, H_2 \in M_{n,n}(\mathbb{C})$ are similar iff $\exists$ a nonsingular matrix $S \in M_{n,n}(\mathbb{C})$ such that $H_2 = S^{-1} H_1 S$.

Definition A.2.10 (Congruency)**.** [Horn and Johnson, 2013, p. 281]
Two matrices $H_1, H_2 \in M_{n,n}(\mathbb{C})$ are congruent iff $\exists$ a nonsingular matrix $S \in M_{n,n}(\mathbb{C})$ such that $H_2 = S^* H_1 S$.

A.3 Relevant theorems

Theorem A.3.1 (Cholesky factorization)**.** *Let $A \in M_{n,n}$ be Hermitian. Then A is positive-definite iff $\exists$ a (unique) lower triangular matrix with positive diagonal entries such that $A = LL^*$ [Horn and Johnson, 2013, corollary 7.2.9, p. 441].*

Theorem A.3.2 (Eigenvalues of similar matrices)**.** *Similar matrices have same eigenvalues [Horn and Johnson, 2013, corollary 1.3.4, p. 58].*

Theorem A.3.3 (Eigenvalues of Hermitian (symmetric) matrix)**.** *All eigenvalues of a Hermitian (symmetric) matrix are real [Horn and Johnson, 2013, theorem 2.5.6, p. 135].*

Theorem A.3.4 (Eigenvalues of Hermitian (symmetric) positive-definite matrix)**.** *All eigenvalues of a Hermitian (symmetric) positive-definite matrix are real and greater than zero [Horn and Johnson, 2013, theorem 7.2.1, p. 438].*

Theorem A.3.5 (Sylvester's law of inertia)**.** *Hermitian matrices* $H_1, H_2 \in M_{n,n}(\mathbb{C})$ *are congruent iff they have the same number of positive eigenvalues and the same number of negative eigenvalues [Horn and Johnson, 2013, theorem 4.5.8, p. 282].*

Theorem A.3.6 (Lax–Milgram)**.** *For a continuous, coercive bilinear form* $a(\cdot,\cdot)$ *and a continuous linear functional* $F \in V'$ *defined on a Hilbert space* $(V, \langle\cdot,\cdot\rangle)$*,* $\exists$ *a unique* $u \in V$ *such that* $a(u,v) = F(v)$*,* $\forall\, v \in V$ *[Brenner and Scott, 2008, theorem 2.7.7, p. 62]. Here* V' *denotes the dual space of* V*.*

Appendix B

Sample MATLAB runs of test problems in thesis

Some sample runs of the test problems presented in this thesis are given here. Note that all these sample runs have been formatted for better understanding. So, they differ in display from their actual runs in IFISS and S-IFISS toolbox in MATLAB.

B.1 Stochastic diffusion test problem 1

This sample run is produced using S-IFISS toolbox version 1.01 in MATLAB since the results in chapter 2 of this thesis correspond to computations carried out in this version; the current version of this toolbox used in MATLAB is 1.03. Also, `SPD_MINRES` is not yet incorporated in S-IFISS to be called internally from the script `stoch_square_diff` which sets up the FEM and linear algebra logistics. In fact `SPD_MINRES` is currently called externally, that is, from the MATLAB prompt. However, the sample run that is presented below has been formatted to reflect the final version where `SPD_MINRES` will be called internally. Also, in an actual implementation, a posteriori error estimates will be computed periodically, unlike here, where it is computed at each iteration. So, the time taken by the MINRES solver for computing the final solution is not included here. The data generated by the run here corresponds to the diffusion test problem 1 in chapter 2 with mean-based MINRES preconditioning on a uniform grid with $h = 1/32$, $p = 3$, $m = 5$, and $\sigma = 0.5$.

```
>> stoch_diff_testproblem
specification of reference stochastic diffusion problem.
choose specific example
1 L-shaped domain, synthetic random coefficient, constant source
2 Square domain, analytic KL expansion, non-constant source
3 channel domain, analytic KL expansion, trivial mean solution
```

```
4 channel domain, analytic KL expansion, nontrivial mean solution
5 Square domain, Eigel synthetic random coefficient, constant source
6 Square domain, Powell synthetic random coefficient, constant source
: 2

        1 file(s) copied.
        1 file(s) copied.
        1 file(s) copied.
        1 file(s) copied.

WARNING: you must select the default [-1,1] x [-1,1] square domain!

Grid generation for unit square  domain.
grid parameter: 3 for underlying 8x8 grid (default is 16x16) : 6
uniform/stretched grid (1/2) (default is uniform) : 1
[0,1] or [-1,1] square (enter 1/2) (default is [-1,1]) : 2
Number of random variables? (default is 5) : 5
Total polynomial degree? (default is 3) : 3
setting up KL expansion data
standard deviation? (default 0.3) : 0.5
correlation length in x-direction? (default 2.0) : 2
correlation length in y-direction? (default 2.0) : 2
Q1/Q2 approximation 1/2? (default Q1) : 1
save results for reference 1/0 (yes/no)? (default no) : 0
estimate error a posteriori 1/0 (yes/no)? (default yes) : 1
MINRES solution of the linear system ...
setting up stochastic Q1 diffusion matrices...  done

Call to SPD_MINRES with error control...
discrete parametric diffusion system...
mean-based preconditioner is used

Bingo! optimal convergence in  17 iterations
final estimated error is 3.3338e-02

k    Estimated-Error    Algebraic-Bound    Residual-Error
1     9.5210e+00         5.6148e+01         6.2904e+01
2     4.3453e+00         2.2306e+01         2.1291e+01
3     2.6261e+00         1.2357e+01         1.0044e+01
4     1.6513e+00         6.4619e+00         4.4699e+00
5     1.2752e+00         4.5724e+00         2.8116e+00
6     8.5125e-01         2.8425e+00         1.4687e+00
7     6.8418e-01         2.4265e+00         1.1109e+00
8     4.2364e-01         1.8590e+00         7.2594e-01
9     2.6503e-01         1.3266e+00         4.8434e-01
10    1.3325e-01         6.7028e-01         2.3518e-01
11    1.0274e-01         4.7473e-01         1.6424e-01
12    6.8965e-02         2.8906e-01         9.8153e-02
13    5.5823e-02         2.2256e-01         7.4401e-02
14    3.9996e-02         1.3254e-01         4.3340e-02
15    3.5562e-02         8.1927e-02         2.6521e-02
```

```
16    3.3766e-02          4.3677e-02          1.4037e-02
17    3.3338e-02          3.2408e-02          1.0370e-02      Stop here!

k       XQ-Error           YQ-Error           YP-Error
1     4.9115e+00         2.4532e+00         5.5107e+00
2     2.4767e+00         1.2384e+00         1.8833e+00
3     1.5622e+00         7.8086e-01         8.9226e-01
4     1.0134e+00         5.0633e-01         4.0025e-01
5     7.9041e-01         3.9496e-01         2.5401e-01
6     5.3163e-01         2.6587e-01         1.3414e-01
7     4.2783e-01         2.1405e-01         1.0199e-01
8     2.6462e-01         1.3235e-01         6.6278e-02
9     1.6536e-01         8.2476e-02         4.4115e-02
10    8.3592e-02         4.0671e-02         2.1725e-02
11    6.4905e-02         3.0726e-02         1.5571e-02
12    4.4246e-02         1.9262e-02         9.9368e-03
13    3.6313e-02         1.4399e-02         8.0253e-03
14    2.6940e-02         7.5688e-03         5.7895e-03
15    2.4430e-02         4.8565e-03         4.8876e-03
16    2.3434e-02         3.3194e-03         4.4548e-03
17    2.3193e-02         2.8646e-03         4.3790e-03

Eigenvalue convergence
k       Smallest    Largest
1       1.2551      1.2551
2       0.9111      1.5715
3       0.6607      1.7117
4       0.4785      1.8093
5       0.3781      1.8447
6       0.2670      1.8728
7       0.2096      1.8797
8       0.1525      1.8837
9       0.1333      1.8863
10      0.1231      1.8909
11      0.1197      1.8937
12      0.1153      1.8968
13      0.1117      1.8980
14      0.1069      1.8993
15      0.1048      1.9005
16      0.1033      1.9029
17      0.1024      1.9044

Linear solver statistics:
initial guess is random       ndof is 236600

Stochastic parameters and computed statistics:
5 active random variables | polynomial degree is  3
standard deviation = 5.0000e-01
correlation x-length = 2          correlation y-length = 2
maximum mean value = 7.979e-02    maximum variance value = 1.741e-03
```

B.2 Stokes equations test problem 1

This sample run is produced using IFISS toolbox version 3.3 in MATLAB since the results in chapter 3 of this thesis correspond to computations carried out in this version; the current version of this toolbox used in MATLAB is 3.5. Also, `SADDLE_MINRES` is not yet incorporated in IFISS to be called internally from the script `stokes_square_diff` which sets up the FEM and linear algebra logistics. In fact `SADDLE_MINRES` is called externally presently, that is, from the MATLAB prompt. However, the sample run presented below has been formatted to reflect the final version where `SADDLE_MINRES` will be called internally. Also, in an actual implementation, a posteriori error estimates will be computed periodically, unlike here, where it is computed at each iteration. So, the time taken by the MINRES solver for computing the final solution is not included here. The data generated by the run here corresponds to the Stokes test problem 1 in chapter 3 with block AMG preconditioning on a uniform grid with $h = 1/128$.

```
>> stokes_testproblem
specification of reference Stokes problem.
choose specific example (default is cavity)
     1  Channel domain
     2  Flow over a backward facing step
     3  Lid driven cavity
     4  Colliding flow
 : 4
        1 file(s) copied.
        1 file(s) copied.


Grid generation for cavity domain.
grid parameter: 3 for underlying 8x8 grid (default is 16x16) : 8
uniform/stretched grid (1/2) (default is uniform) : 1
Q1-Q1/Q1-P0/Q2-Q1/Q2-P1: 1/2/3/4? (default Q1-P0) : 2
setting up Q1-P0 matrices...  done
system matrices saved in square_stokes_nobc.mat ...
imposing (enclosed flow) boundary conditions ...
stabilization parameter (default is 1/4) : 1/4

Call to SADDLE_MINRES with error control ...
discrete Stokes system ...
iterative solution with preconditioned MINRES
maximum number of iterations? (default 100) : 100
preconditioner:
   0  none
   1  diagonal
   2  ideal block
```

```
   3  geometric multigrid block
   4  AMG block
default is AMG  : 4
AMG preconditioning...
AMG grid coarsening ...  6 grid levels constructed.
setup done.
PDJ/PGS smoother? 1/2 (point damped Jacobi) : 1
point damped Jacobi smoothing ..

Bingo! optimal convergence in  35 iterations
final estimated error is 3.0691e-01
```

k	Estimated-Error	Algebraic-Bound	Residual-Error
0	1.0227e+02		5.4237e+02
1	8.9877e+01		1.9460e+02
2	9.0299e+01	1.5434e+02	1.9124e+02
3	2.3297e+01	8.8748e+01	5.6919e+01
4	2.0280e+01	8.2246e+01	5.1567e+01
5	1.0618e+01	5.8310e+01	2.7436e+01
6	8.7140e+00	5.4948e+01	2.0689e+01
7	5.4809e+00	3.8570e+01	9.1642e+00
8	4.4629e+00	3.5400e+01	8.1567e+00
9	2.8637e+00	2.6878e+01	5.6753e+00
10	2.4072e+00	2.4293e+01	4.7866e+00
11	2.1757e+00	2.4028e+01	3.9990e+00
12	1.6644e+00	2.9380e+01	3.0711e+00
13	1.5139e+00	3.2123e+01	2.6221e+00
14	9.3009e-01	3.3599e+01	1.6881e+00
15	9.2572e-01	3.3628e+01	1.6805e+00
16	5.4101e-01	2.6909e+01	8.8737e-01
17	5.4323e-01	2.6919e+01	8.8650e-01
18	3.8487e-01	1.8525e+01	5.2757e-01
19	3.8556e-01	1.8494e+01	5.2609e-01
20	3.4542e-01	1.3662e+01	3.6515e-01
21	3.4093e-01	1.2339e+01	3.2301e-01
22	3.2409e-01	8.7885e+00	2.1483e-01
23	3.1907e-01	7.2919e+00	1.7234e-01
24	3.1350e-01	5.8741e+00	1.3431e-01
25	3.1145e-01	4.5582e+00	1.0083e-01
26	3.1059e-01	3.9845e+00	8.6765e-02
27	3.0841e-01	2.4474e+00	5.0832e-02
28	3.0829e-01	2.3764e+00	4.9233e-02
29	3.0755e-01	1.4830e+00	2.9788e-02
30	3.0755e-01	1.4810e+00	2.9740e-02
31	3.0712e-01	8.3231e-01	1.6276e-02
32	3.0717e-01	8.2186e-01	1.6062e-02

```
33      3.0697e-01          4.5011e-01          8.6641e-03
34      3.0696e-01          4.2203e-01          8.1141e-03
35      3.0691e-01          2.5398e-01          4.8521e-03     Stop here!


 Eigenvalue convergence
 k   Smallest-Minus      Largest-Minus     Smallest-Plus       Largest-Plus
 1
 2
 3    -9.6579e-01         -1.0061e+00       9.7323e-01          1.4768e+00
 4    -1.0341e+00         -9.9244e-01       9.6921e-01          1.4982e+00
 5    -1.1299e+00         -8.4547e-01       9.4926e-01          1.5192e+00
 6    -1.2194e+00         -7.6061e-01       9.4502e-01          1.5365e+00
 7    -1.2624e+00         -6.0745e-01       9.3499e-01          1.5531e+00
 8    -1.2697e+00         -5.9969e-01       9.3370e-01          1.5608e+00
 9    -1.2748e+00         -5.7508e-01       9.2813e-01          1.5663e+00
10    -1.2812e+00         -5.5677e-01       9.2507e-01          1.5733e+00
11    -1.2850e+00         -5.1233e-01       9.2073e-01          1.5771e+00
12    -1.2902e+00         -4.0669e-01       9.0875e-01          1.5823e+00
13    -1.2925e+00         -3.5971e-01       8.9765e-01          1.5852e+00
14    -1.2955e+00         -2.8289e-01       8.6227e-01          1.5928e+00
15    -1.2971e+00         -2.8235e-01       8.6209e-01          1.5953e+00
16    -1.3001e+00         -2.2968e-01       8.4436e-01          1.5997e+00
17    -1.3010e+00         -2.2963e-01       8.4434e-01          1.6012e+00
18    -1.3023e+00         -2.1368e-01       8.3475e-01          1.6032e+00
19    -1.3032e+00         -2.1362e-01       8.3470e-01          1.6042e+00
20    -1.3040e+00         -2.0714e-01       8.2711e-01          1.6054e+00
21    -1.3048e+00         -2.0505e-01       8.2537e-01          1.6062e+00
22    -1.3054e+00         -1.9820e-01       8.2071e-01          1.6070e+00
23    -1.3059e+00         -1.9494e-01       8.1896e-01          1.6079e+00
24    -1.3064e+00         -1.9178e-01       8.1714e-01          1.6085e+00
25    -1.3069e+00         -1.8868e-01       8.1516e-01          1.6093e+00
26    -1.3072e+00         -1.8722e-01       8.1428e-01          1.6097e+00
27    -1.3077e+00         -1.8291e-01       8.1150e-01          1.6108e+00
28    -1.3080e+00         -1.8270e-01       8.1140e-01          1.6111e+00
29    -1.3083e+00         -1.7993e-01       8.0988e-01          1.6117e+00
30    -1.3085e+00         -1.7992e-01       8.0988e-01          1.6120e+00
31    -1.3088e+00         -1.7757e-01       8.0882e-01          1.6125e+00
32    -1.3090e+00         -1.7753e-01       8.0880e-01          1.6126e+00
33    -1.3092e+00         -1.7620e-01       8.0781e-01          1.6129e+00
34    -1.3094e+00         -1.7611e-01       8.0772e-01          1.6131e+00
35    -1.3095e+00         -1.7556e-01       8.0703e-01          1.6134e+00

 Discrete inf-sup convergence
 k    Infsup-Estimate
 1
 2
 3     1.4305e+00
 4     1.4173e+00
```

```
  5     1.2643e+00
  6     1.1717e+00
  7     1.0011e+00
  8     9.9239e-01
  9     9.6510e-01
 10     9.4439e-01
 11     8.9298e-01
 12     7.6726e-01
 13     7.0983e-01
 14     6.1294e-01
 15     6.1224e-01
 16     5.4051e-01
 17     5.4045e-01
 18     5.1805e-01
 19     5.1797e-01
 20     5.0893e-01
 21     5.0596e-01
 22     4.9605e-01
 23     4.9127e-01
 24     4.8661e-01
 25     4.8203e-01
 26     4.7987e-01
 27     4.7343e-01
 28     4.7311e-01
 29     4.6893e-01
 30     4.6892e-01
 31     4.6535e-01
 32     4.6529e-01
 33     4.6329e-01
 34     4.6315e-01
 35     4.6233e-01

Linear solver statistics:
initial guess is random
ndof is 197634
```

B.3 Convection-diffusion equations test problem

This sample run is produced using IFISS toolbox version 3.3 in MATLAB since the results in chapter 4 of this thesis correspond to computations carried out in this version. Also, `CD_GMRES`, `CD_BICGSTAB(`ℓ`)`, and `CD_TFQMR` are not yet incorporated in IFISS to be called internally from the script `solve_cd` which sets up the FEM and linear algebra logistics. In fact these functions are currently called externally, that is, from the MATLAB prompt. However, the sample runs presented below have been formatted to reflect the final version where these functions will be called internally. Also, in an actual implementation, a posteriori error estimates will be computed periodically, unlike here, where it is computed at each iteration. So, the time taken by the employed solvers for computing the final solution is not included here. The data generated by the runs here corresponds to the convection-diffusion test problem in chapter 4 with AMG preconditioning on a uniform grid with $h = 1/128$.

```
>> cd_testproblem
specification of reference convection-diffusion problem.
choose specific example
     1  Constant vertical wind
     2  Vertical wind, characteristic layers
     3  Constant wind @ 30 degree angle
     4  Recirculating wind
 : 4
        1 file(s) copied.
        1 file(s) copied.

Grid generation for unit square  domain.
grid parameter: 3 for underlying 8x8 grid (default is 16x16) : 8
uniform/stretched grid (1/2) (default is uniform) : 1
[0,1] or [-1,1] square (enter 1/2) (default is [-1,1]) : 2
setting up Q1 convection-diffusion matrices...  done
system matrices saved in square_cd_nobc.mat ...
viscosity parameter (default 1/64) : 1/64
plotting element data... done
maximum element Peclet number is 4.980393e-001
SUPG parameter (default is optimal) :
SUPG setting up Q1 SUPG stabilisation matrix...  not needed!
system saved in square_cd.mat ...

% GMRES run
GMRES/Bicgstab(l)/TFQMR 1/2/3 (default GMRES) : 1
Call to GMRES_CD with error control ...
discrete convection-diffusion system ...
```

```
maximum number of iterations? (default 100) : 100
preconditioner:
   0  none
   1  diagonal
   2  incomplete LU
   3  geometric multigrid
   4  algebraic multigrid
default is AMG  :  4
compute / load AMG data? 1/2 (default 1) : 1
AMG grid coarsening ...  24 grid levels constructed.
setup done.
plot AMG grid sequence? yes/no 1/2 (default no) : 2
PDJ/PGS/LGS/ILU smoother? 1/2/3/4 (point damped Jacobi) : 1
point damped Jacobi smoothing ..

GMRES iteration ...
Bingo! optimal convergence in  5 iterations
final estimated error is 7.8851e-001
maximum eigenvalue of generalized eigenvalue problem is = 1.3596e+07

 k    Estimated-Error     Algebraic-Bound        Residual-Error

 0      7.7377e+01          1.2089e+04            3.2787e+00
 1      4.7025e+00          6.6979e+04            1.8165e-01
 2      8.3137e-01          3.2621e+02            8.8471e-03
 3      7.8937e-01          3.8110e+00            1.0335e-03
 4      7.8858e-01          1.0054e+00            2.7267e-04
 5      7.8851e-01          3.2551e-01            8.8279e-05    Stop here!

% BICGSTAB(2) run
GMRES/Bicgstab(l)/TFQMR 1/2/3 (default GMRES) : 2
Ell (default 2) : 2
Call to BICGSTAB(l)_CD with error control ...
discrete convection-diffusion system ...
maximum number of iterations? (default 100) : 100
preconditioner:
   0  none
   1  diagonal
   2  incomplete LU
   3  geometric multigrid
   4  algebraic multigrid
default is AMG  :  4
compute / load AMG data? 1/2 (default 1) : 1
AMG grid coarsening ...  24 grid levels constructed.
setup done.
plot AMG grid sequence? yes/no 1/2 (default no) : 2
PDJ/PGS/LGS/ILU smoother? 1/2/3/4 (point damped Jacobi) : 1
point damped Jacobi smoothing ..
```

```
BICGSTAB(l) iteration ...
Bingo! optimal convergence in  4 iterations
final estimated error is 7.8851e-001
maximum eigenvalue of generalized eigenvalue problem is = 1.3596e+07

 k    Estimated-Error     Algebraic-Bound        Residual-Error

 0      9.5841e+03          1.0627e+06             2.8820e+02
 1      4.2795e+02          4.6336e+04             1.2567e+01
 2      5.1259e-00          5.4746e+02             1.4847e-01
 3      8.1381e-01          2.1679e+01             5.8795e-03
 4      7.8850e-01          1.6291e-01             4.4181e-05    Stop here!

% TFQMR run
GMRES/Bicgstab(l)/TFQMR 1/2/3 (default GMRES) : 3
Call to TFQMR_CD with error control ...
discrete convection-diffusion system ...
maximum number of iterations? (default 100) : 100
preconditioner:
   0  none
   1  diagonal
   2  incomplete LU
   3  geometric multigrid
   4  algebraic multigrid
default is AMG  :  4
compute / load AMG data? 1/2 (default 1) : 1
AMG grid coarsening ...  24 grid levels constructed.
setup done.
plot AMG grid sequence? yes/no 1/2 (default no) : 2
PDJ/PGS/LGS/ILU smoother? 1/2/3/4 (point damped Jacobi) : 1
point damped Jacobi smoothing ..

TFQMR iteration ...
Bingo! optimal convergence in  5 iterations
final estimated error is 7.8856e-001
maximum eigenvalue of generalized eigenvalue problem is = 1.3596e+07

 k    Estimated-Error     Algebraic-Bound        Residual-Error

 0      7.7189e+01          1.2071e+04             3.2739e+00
 1      4.6939e+00          6.6944e+02             1.8156e-01
 2      8.4801e-01          4.2719e+01             1.1586e-02
 3      7.8953e-01          4.2103e+00             1.1419e-03
 4      7.8863e-01          1.1984e+00             3.2501e-04
 5      7.8856e-01          6.9363e-01             1.8812e-04    Stop here!

Linear solver statistics:
initial guess is random       ndof is 66049
```

B.4 Navier–Stokes equations test problem

This sample run is produced using IFISS toolbox version 3.3 in MATLAB since the results in chapter 5 of this thesis correspond to computations carried out in this version. Also, `Navier_GMRES` is not yet incorporated in IFISS to be called internally from the script `solve_step_navier` which sets up the FEM and linear algebra logistics. In fact `Navier_GMRES` is currently called externally, that is, from the MATLAB prompt. But the sample run presented below has been formatted to reflect the final version where this function will be called internally. Also, in an actual implementation, a posteriori error estimates will be computed periodically, unlike here, where it is computed at each iteration. So, the time taken by the employed solvers for computing the final solution is not included here. The data generated by the run here corresponds to the Navier–Stokes test problem in chapter 5 on a 64×192 grid.

```
>> navier_testproblem
specification of reference Navier-Stokes problem.
choose specific example (default is cavity)
     1  Channel domain
     2  Flow over a backward facing step
     3  Lid driven cavity
     4  Flow over a plate
     5  Flow over an obstacle
 : 2
        1 file(s) copied.
        1 file(s) copied.
horizontal dimensions [-1,L]: L? (default L=5) : 5

Grid generation for backward-facing step domain.
grid parameter: 3 for underlying 8x24 grid (default is 4) : 6
grid stretch factor (default is 1) : 1
   Grid generation for x-channel ...done.
   Grid generation for x-channel ...done.
   Merger of two x-channel grids
   zip distance is 0.0000e+000 ... it should be close to zero!
All done.

Q1-Q1/Q1-P0/Q2-Q1/Q2-P1: 1/2/3/4? (default Q1-P0) : 2
setting up Q1-P0 matrices...  done
system matrices saved in step_stokes_nobc.mat ...
Incompressible flow problem on step domain ...
viscosity parameter (default 1/50) : 1/50
Picard/Newton/hybrid linearization 1/2/3 (default hybrid) : 2
number of Newton iterations (default 20) : 20
stokes system ...
```

```
nubeta is set to 1.250000e+001
setting up Q1 convection matrix...  done.


Newton iteration number 1
setting up Q1 Newton Jacobian matrices...  done.
setting up Q1 convection matrix...  done.

Linear iteration...
inflow/outflow (step) problem ...
solving Jacobian system generated by solution from last Newton step

setting up Q1 Newton Jacobian matrices...  done.
GMRES/Bicgstab(l)/IDR(s)  1/2/3 (default GMRES) : 1
maximum number of iterations? (default 100) : 100
preconditioner:
   0  none
   1  unscaled least-squares commutator (BFBt)
   2  pressure convection-diffusion (Fp)
   3  least-squares commutator (LSC)
   4  modified pressure convection-diffusion (Fp*)
   5  boundary-adjusted least-squares commutator (LSC*)
default is modified pressure convection-diffusion : 4

ideal / AMG iterated  preconditioning? 1/2 (default ideal) : 1
setting up modified Q0 pressure preconditioning matrices...
NonUniform grids are fine.

GMRES iteration ...
maximum eigenvalue of generalized eigenvalue problem = 2.5802e+05
minimum eigenvalue of generalized eigenvalue problem = 2.1691e-01
stronger bound = 5.5400e+05

Bingo!
optimal convergence in 23 iterations

nonlinear relative residual = 7.0434e-04
velocity change = 1.2286e+01

estimated error = 2.0564e-01
energy norm of linearized part = 1.5702e+01


Newton iteration number 2
setting up Q1 Newton Jacobian matrices...  done.
setting up Q1 convection matrix...  done.

Linear iteration...
inflow/outflow (step) problem ...
solving Jacobian system generated by solution from last Newton step
```

```
setting up Q1 Newton Jacobian matrices...  done.
GMRES/Bicgstab(l)/IDR(s)  1/2/3 (default GMRES) : 1
maximum number of iterations? (default 100) : 100
preconditioner:
   0  none
   1  unscaled least-squares commutator (BFBt)
   2  pressure convection-diffusion (Fp)
   3  least-squares commutator (LSC)
   4  modified pressure convection-diffusion (Fp*)
   5  boundary-adjusted least-squares commutator (LSC*)
default is modified pressure convection-diffusion : 4

ideal / AMG iterated  preconditioning? 1/2 (default ideal) : 1
setting up modified Q0 pressure preconditioning matrices...
NonUniform grids are fine.

GMRES iteration ...
maximum eigenvalue of generalized eigenvalue problem = 4.400e+05
minimum eigenvalue of generalized eigenvalue problem = 2.1690e-01
stronger bound = 9.4470e+05

Bingo!
optimal convergence in 33 iterations

nonlinear relative residual = 3.1387e-05
velocity change = 2.404e+00

estimated error = 1.7400e-01
energy norm of linearized part = 3.5289e-01


Newton iteration number 3
setting up Q1 Newton Jacobian matrices...  done.
setting up Q1 convection matrix...  done.

Linear iteration...
inflow/outflow (step) problem ...
solving Jacobian system generated by solution from last Newton step

setting up Q1 Newton Jacobian matrices...  done.
GMRES/Bicgstab(l)/IDR(s)  1/2/3 (default GMRES) : 1
maximum number of iterations? (default 100) : 100
preconditioner:
   0  none
   1  unscaled least-squares commutator (BFBt)
   2  pressure convection-diffusion (Fp)
   3  least-squares commutator (LSC)
   4  modified pressure convection-diffusion (Fp*)
   5  boundary-adjusted least-squares commutator (LSC*)
default is modified pressure convection-diffusion : 4
```

```
ideal / AMG iterated  preconditioning? 1/2 (default ideal) : 1
setting up modified Q0 pressure preconditioning matrices...
NonUniform grids are fine.

GMRES iteration ...
maximum eigenvalue of generalized eigenvalue problem = 4.0478e+05
minimum eigenvalue of generalized eigenvalue problem = 2.1691e-01
stronger bound = 8.6912e+05

Bingo!
optimal convergence in 36 iterations

nonlinear relative residual = 4.4474e-08
velocity change = 1.2167e-01

estimated error = 1.7432e-01
energy norm of linearized part = 1.5442e-02

finished!
nonlinear convergence test (energy norm <= estimated error) satisfied
```

Appendix C

CPUTIME comparisons of some test problems in chapter 2

A comparison of the cputimes when using preconditioned MINRES with the balanced stopping test (as opposed to not using a balanced stopping test) for solving symmetric positive-definite linear systems of chapter 2 are given here. Since the balanced stopping methodology is more useful to employ when solving discrete systems with stochastic PDE origins, comparisons of cputimes of the MINRES solver are presented here only for discrete systems arising from stochastic diffusion equations (for which a 'tight' a posteriori approximation error estimator is available).

The cputimes are listed in $(\cdot)$ along with the corresponding MINRES (with mean-based preconditioning) iteration counts in Tables C.1 to C.4. As in chapter 2, k_{tol1}, k_{tol2} denote the MINRES iterations needed to satisfy a fixed absolute residual $\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}$ reduction tolerance of `1e-6` and `1e-9` respectively. MINRES iterations needed to satisfy the balanced stopping test are denoted by k^*.

Note that the cputime taken to satisfy the balanced stopping test (for any problem, on any grid) is 'far greater' than the cputimes needed to satisfy the corresponding fixed absolute residual $\|\mathbf{r}^{(k)}\|_{\mathcal{M}^{-1}}$ reduction tolerance of `1e-6` and `1e-9`. This is the case here because the a posteriori error estimate (which constitutes the bulk of cputime in MINRES iteration with balanced stopping test) is computed at every iteration. In practice, this would be computed periodically (say after every 5 or 10 iterations) and the cputime corresponding to k^* will be drastically reduced. However, it will still be more (acceptably or unacceptably more depending upon a practitioners choice) than the corresponding cputimes for k_{tol1}, k_{tol2} MINRES iterations. In any case, the balanced stopping test does rule out premature stopping of the MINRES solver.

Table C.1: Iteration counts (cputimes in seconds) for diffusion test problem 1 with $\sigma = 0.3$, $m = 5$, and $p = 3$.

h	k_{tol1}	k_{tol2}	k^*	#dof
1/4	14 (0.296)	19 (0.561)	6 (66.940)	2744
1/8	14 (0.281)	20 (0.749)	7 (84.397)	12600
1/16	15 (0.998)	20 (1.997)	8 (143.75)	53816
1/32	16 (7.816)	21 (10.358)	9 (428.61)	222264

Table C.2: Iteration counts (cputimes in seconds) for diffusion test problem 1 with $\sigma = 0.5$, $m = 5$, and $p = 3$.

h	k_{tol1}	k_{tol2}	k^*	#dof
1/4	30 (0.546)	43 (0.655)	11 (116.49)	2744
1/8	34 (1.138)	49 (1.732)	14 (165.58)	12600
1/16	36 (3.650)	52 (4.992)	16 (286.29)	53816
1/32	38 (18.486)	53 (25.834)	17 (816.32)	222264

Table C.3: Iteration counts (cputimes in seconds) for diffusion test problem 2 with slow decay, $m = 5$, and $p = 3$.

h	k_{tol1}	k_{tol2}	k^*	#dof
1/4	13 (0.250)	18 (0.328)	5 (52.526)	2744
1/8	14 (0.452)	20 (0.686)	6 (72.696)	12600
1/16	15 (1.435)	21 (2.044)	8 (141.49)	53816
1/32	16 (8.128)	21 (10.280)	9 (420.25)	222264

Table C.4: Iteration counts (cputimes in seconds) for diffusion test problem 2 with fast decay, $m = 5$, and $p = 3$.

h	k_{tol1}	k_{tol2}	k^*	#dof
1/4	17 (0.328)	25 (0.343)	6 (63.461)	2744
1/8	20 (0.733)	27 (0.952)	8 (109.06)	12600
1/16	21 (2.044)	29 (2.496)	10 (177.01)	53816
1/32	22 (11.185)	30 (14.898)	12 (555.55)	222264

Appendix D

Eigenvalue behaviour of perturbed convection-diffusion operator

Apart from devising optimal balanced black-box stopping tests in iterative solvers, some interesting computational experiments were carried out, which investigated the behaviour of eigenvalues of the discrete convection-diffusion operator when the discrete diffusion or/and convection operators are slightly perturbed. The motivation for such an investigation stems from the recent work by [Elman and Silvester, 2017] where the authors use stochastic collocation with perturbation analysis to provide insight into stability issues associated with unsteady flow problems. The authors argue that although their approach is similar to pseudo-spectral analysis, it is less expensive than the pseudo-spectral approach.

The material presented here is based on deterministic steady-state convection-diffusion equations. Although there are no 'stability' issues in general associated with steady-state problems, nevertheless it is an important and interesting starting point to begin an investigation of behaviour of eigenvalues to perturbations.

D.1 Convection-diffusion eigenvalue problem

The steady-state deterministic convection-diffusion eigenvalue problem is to find the eigenvector, eigenvalue pairs $u(\vec{x}) : D \rightarrow \mathbb{R}$ and $\lambda \in \mathbb{R}$ satisfying

$$\begin{aligned} -\nabla \cdot \epsilon \nabla u(\vec{x}) + w(\vec{x}) \cdot \nabla u(\vec{x}) &= \lambda u(\vec{x}), \quad \vec{x} \in D \subset \mathbb{R}^d \, (d = 2, 3), \\ u(\vec{x}) &= 0, \quad \vec{x} \in \partial D_D, \\ \epsilon \nabla u(\vec{x}) \cdot \vec{n} &= 0, \quad \vec{x} \in \partial D_N = \partial D \backslash \partial D_D. \end{aligned} \tag{D.1}$$

Here ϵ is the diffusion coefficient, w is the wind, D is the spatial domain, and ∂D_D, ∂D_N are the Dirichlet and Neumann parts respectively of the spatial boundary ∂D. The

vector $\vec{n}$ denotes the outward normal to ∂D, and λ is an eigenvalue of the continuous steady-state convection-diffusion operator.

Converting (D.1) into a weak form followed by Galerkin FEM formulation results in the discrete eigenvalue problem

$$F\mathbf{x} = \lambda_{\text{discrete}} Q\mathbf{x}. \tag{D.2}$$

Here F is the convection-diffusion matrix as in chapter 4, Q is the mass matrix formed from the FEM basis functions and $\lambda_{\text{discrete}}$ is an eigenvalue of the discrete convection-diffusion operator F (scaled by the mass matrix) and $\mathbf{x}$ denotes the corresponding eigenvector.

For the material presented here, $F \in \mathbb{R}^{n\times n}$ will be assumed to be nonsingular and the diffusion coefficient ϵ to be a constant. As seen in chapter 4, the linear system F has the following form for lower order finite elements

$$F \;=\; \epsilon A \;+\; N \;+\; S. \tag{D.3}$$

Here A is a symmetric (or the diffusion stiffness matrix) positive-definite matrix, N is a skew symmetric (or the convection matrix) and S is the stabilization matrix.

Instead of considering (D.2), a slightly perturbed form of (D.2) is considered here, that is

$$(\epsilon A + (N + N_E(\boldsymbol{\xi})) + S)\mathbf{x} = \hat{\lambda}_{\text{discrete}} Q\mathbf{x}, \tag{D.4}$$

where $N_E(\boldsymbol{\xi})$ is a perturbation in the convection matrix N and $\hat{\lambda}_{\text{discrete}}$ is a perturbed eigenvalue. Note that the entries of the perturbation matrix $N_E(\boldsymbol{\xi})$ are dependent on a finite number of parameters $(\xi_1, \ldots, \xi_n)^T := \boldsymbol{\xi}$.

Following research investigations can be carried out from (D.4).

- Examine the perturbations in the eigenvalues when there is no diffusion.
- Examine the perturbations in the eigenvalues when there is no diffusion and no stabilization.
- Examine the perturbations in the eigenvalues when there is no stabilization.
- Examine (D.4) itself.

Results for only the first case are presented here.

D.2 Computational results

Results are presented here for the test problem considered in chapter 4 which can be set up by selecting example 4 in the driver `cd_testproblem` in IFISS. The FEM matrices are set up using Q_1 approximation on a 16×16 uniform grid.

In order to observe the behaviour of a particular eigenvalue $\lambda^*_{\text{discrete}}$ (say) with respect to perturbations, nearby eigenvalue problems (D.4) are solved using a finite number (here 25) of 'collocation' points (using spinterp [Klimke and Wohlmuth, 2005] package in MATLAB) $\boldsymbol{\xi}^{(1)}, \boldsymbol{\xi}^{(2)}, \ldots, \boldsymbol{\xi}^{(25)}$. The parameters ξ_i's here are uniform random variables defined on interval $[-1, 1]$. Next, a polynomial interpolation on 500 points is carried out to obtain a polynomial interpolant for the eigenvalue $\lambda^*_{\text{discrete}}$. Note that different collocation points and varied random variables (Gaussian for example) defined on different intervals can also be chosen but is not done here.

The above method for observing the behaviour to perturbations can be done for each of the eigenvalues. However, the results for only a few of them are presented here in Figures D.1 to D.5. Note that many eigenvalues show similar behaviour to perturbations, so presented here are only the distinct behaviours.

Each pair of figures (from left to right) shows first the eigenvalue to shadow (which is marked with a black cross on the left-hand-side plot) and then the 'perturbation surface' of that eigenvalue is obtained (on the corresponding right-hand-side plot) using the procedure discussed (in the second paragraph) above. Also, the red cross in the 'perturbation surface' plot denotes the true computed eigenvalue.

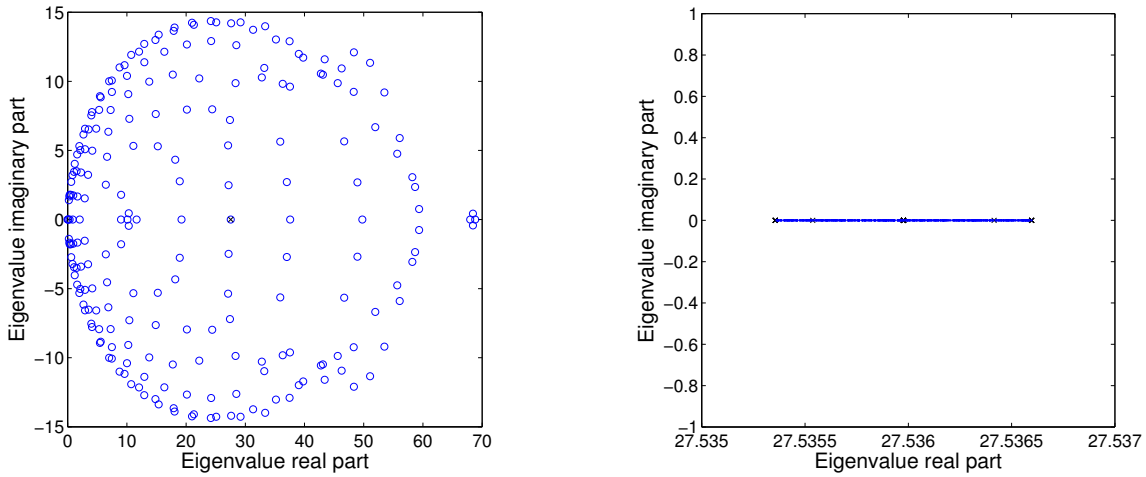

Figure D.1: Eigenvalue to shadow (left) and its perturbed values (right) for CD test problem on a 16×16 uniform grid.

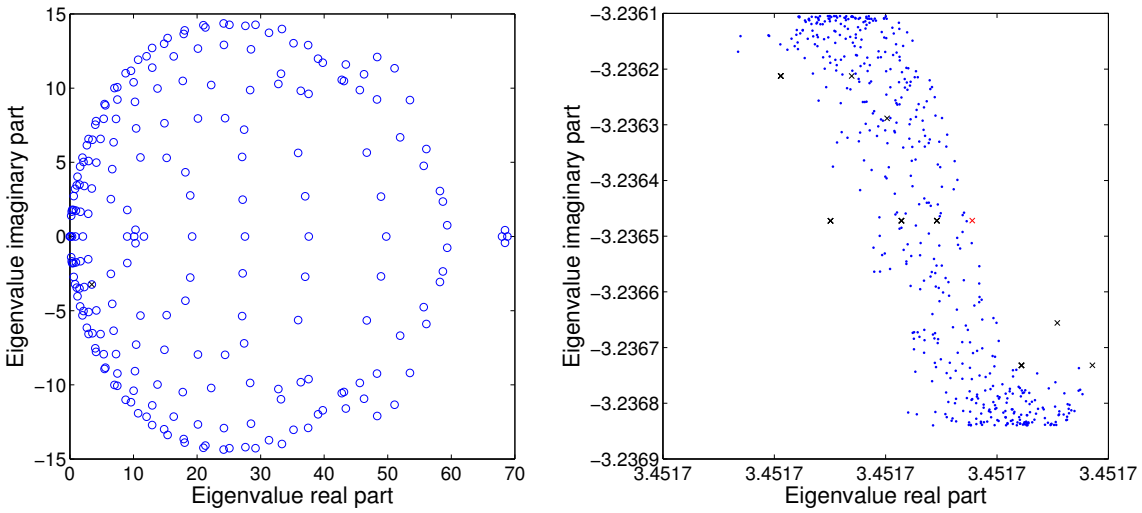

Figure D.2: Eigenvalue to shadow (left) and its perturbed values (right) for CD test problem on a 16×16 uniform grid.

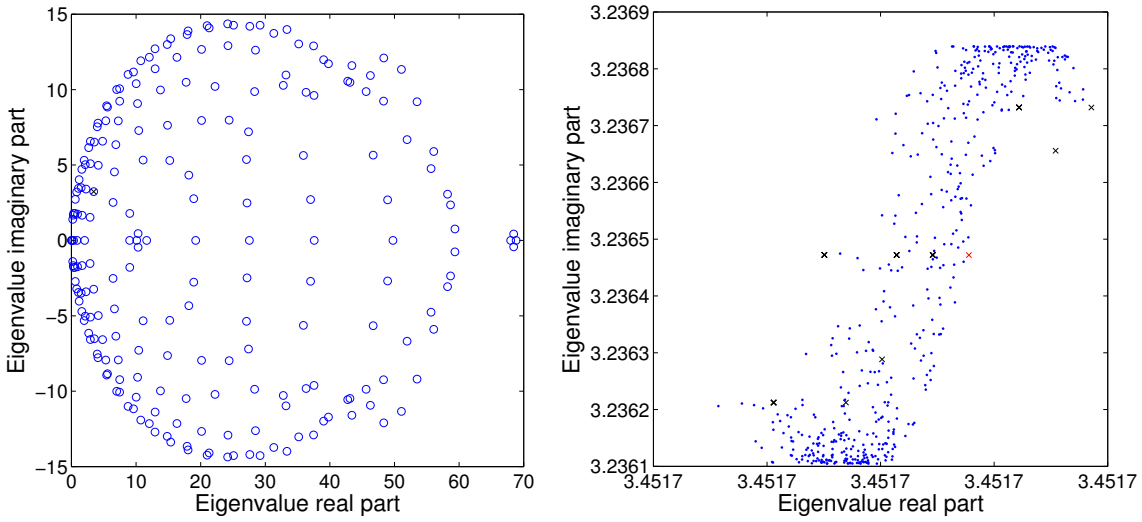

Figure D.3: Eigenvalue to shadow (left) and its perturbed values (right) for CD test problem on a 16×16 uniform grid.

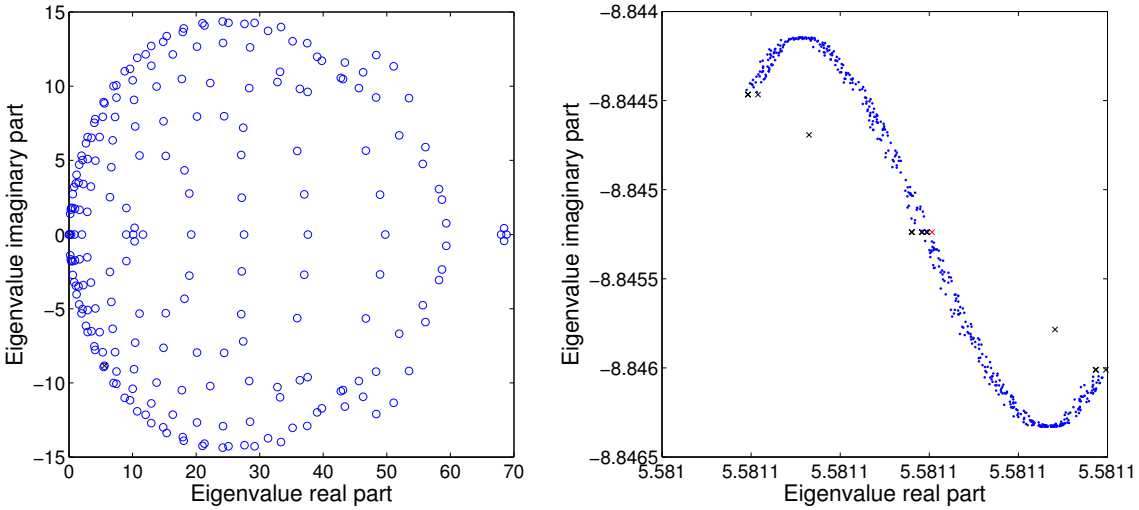

Figure D.4: Eigenvalue to shadow (left) and its perturbed values (right) for CD test problem on a 16×16 uniform grid.

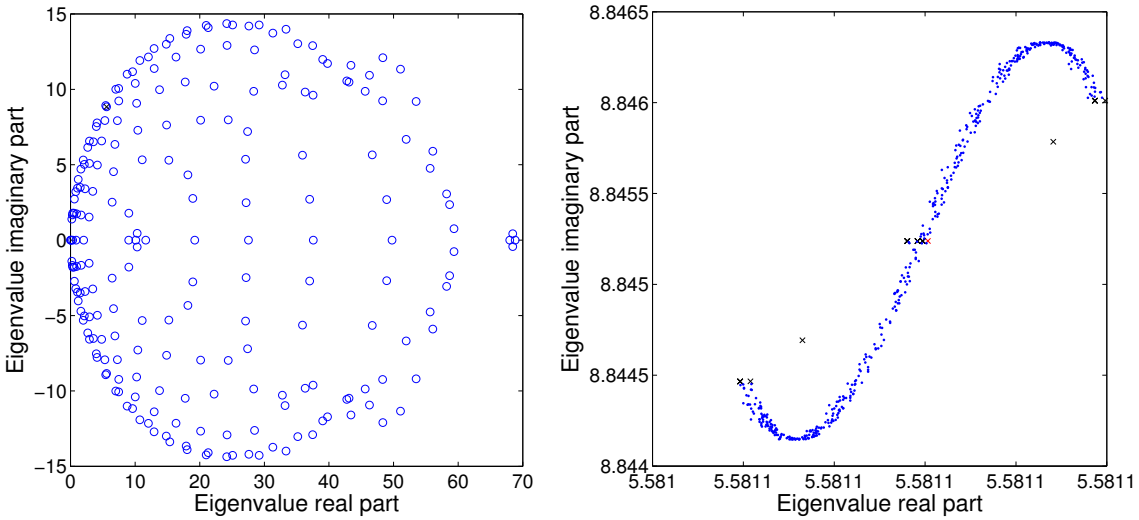

Figure D.5: Eigenvalue to shadow (left) and its perturbed values (right) for CD test problem on a 16×16 uniform grid.

D.3 Computational insights

From the plots, it would be interesting to research the reasons for the following observations from computations.

- Any perturbation in a real eigenvalue keeps it real.
- A complex eigenvalue with both nonzero real and imaginary parts shows a two dimensional surface in the plots.
- The conjugate eigenvalues behave in a similar manner when they are perturbed the only difference being the opposite orientation of their perturbed surfaces.