SJÄLVSTÄNDIGA ARBETEN I MATEMATIK

MATEMATISKA INSTITUTIONEN, STOCKHOLMS UNIVERSITET

Fast Iterative Solution of Large Scale Statistical Inverse Problem

av

Gazi Alam

2013 - No 6

Fast Iterative Solution of Large Scale Statistical Inverse Problem

Gazi Alam

Självständigt arbete i matematik 30 högskolepoäng, Avancerad nivå

Handledare: Boris Shapiro

2013

Stockholm University

Master Thesis

Fast Iterative Solution of Large Scale Statistical Inverse Problem

Author:
Gazi Mahmud Alam

Supervisors:
Dr. Martin Stoll
Prof. Peter Benner and
Prof. Boris Shapiro

A thesis submitted in fulfilment of the requirements for the degree of Master of Science

in

Applied Mathematics

April 2013

Abstract

We consider a large scale statistical inverse problem governed by a three dimensional parabolic partial differential equation within the framework of Bayesian inference with Gaussian noise and prior probability densities. The problem is formulated as a PDE constrained optimization problem. In addition to spectrally neutral prior, we consider 2nd and 4th order Gaussian smoothness prior with both Dirichlet and Neumann boundary conditions. In this thesis we apply a preconditioned Krylov subspace method focusing on the fast solution of the linear systems in saddle point form. The preconditioner is of block diagonal form that involves the effective approximation of the Schur complement. We present the numerical experiments illustrating the performance of the preconditioners and the effects of the regularization parameter for both noise and prior terms.

Acknowledgements

First of all, I express my sincere gratitude to Dr. Martin Stoll for the continuous support of this thesis, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis.

I would like to deliberate my cordial thanks to Professor Peter Benner for his remarkable contribution and guidance in this thesis and for giving me the opportunity and the financial support to do my MS thesis in Max-Planck Institute for Dynamics of Complex Technical Systems, Magdeburg.

I express my hearty gratitude and gratefulness to my teacher and advisor Prof. Boris Shapiro for his valuable reference, suggestions, essential remarks and kind cooperation he has done for me during the course work and completing this thesis.

I acknowledge the financial, academic and technical support of Max-Planck Institute for Dynamics of Complex Technical Systems, Magdeburg. The library facilities and computer facilities of the institute have been indispensable.

I am thankful to Saifullah, Younus, Giovanni De Luca, Jessica Bosch, Zoran Tomljanović, Roky, Monir Uddin, Heiko Weichelt, Dr. Jens Saak for the course Scientific Computing, and above all Martin Köhler for his kindness, friendship and technical help during my stay in Magdeburg.

I would like to express my gratitude to Professor Yishao Zhou who taught me several courses. I remember her generosity and encouragement that helped me a lot in this thesis. I also like to thank Professor Pavel Kurasov for his suggestions and kind cooperation during my study period in the department of Mathematics, Stockholm University. I am thankful to all my teachers in the Department of Mathematics at Stockholm University and the Department of Numerical Analysis and Computer Science at KTH for their co-operation during the course work.

I thank my friend Anu, Najim and colleagues of BUBT for their supports and friendship that I needed.

Last but not least, I would like to thank Dr. Sanzida Sharmin Flora for her mental support and great patience for all times in my life. Without her help this study would never have been possible. Her constant and continuous co-operation starting from my life in Sweden to the end of this work proves her love, support and sacrifice for me. My parents, brother and his wife, uncle, my sister Sabina, Lina and her husband have given me their unequivocal support throughout, as always, for which my mere expression of thanks likewise does not suffice.

Contents

List of Figures

List of Tables

List of Algorithms

This thesis is dedicated to my parents

Chapter 1

Introduction

This chapter includes the scientific overview and plan of the thesis.

1.1 Scientific Overview

Inverse problems [1, 2] are commonplace in many science and engineering applications. Such as geophysics, radar, optics, biology, acoustics, communication theory, signal processing and medical and other image processing. Also a great number of problems from various branches of mathematics such as computational algebra, differential equations, integral equations, functional analysis can be classified as inverse problem. Inverse problem can be define as considering the model

$$y = A(x) \tag{1.1}$$

where, $x \in \mathbb{R}^n$ is the model parameter, $y \in \mathbb{R}^m$ is the observed data and A is the operator describing the relationship between the model parameter x and the observed data y. If the observed data y is evaluated given the model parameter x the problem is called **forward problem**. On the other hand if the model parameter x is evaluated given the observed data y the problem is called **inverse problem**.

The solution of inverse problems can describe the important properties like density, velocity of wave propagation, elasticity parameters, conductivity, dielectric permeability, magnetic permeability and properties and locations of inhomogeneities in inaccessible

areas. The development of computing power and the improvement of the numerical techniques made it easier to simulate the problems of increasing complexity. But in many applications in science and engineering problems model parameters (the causes for desired state or observed state) need to be reconstructed. This fact guided the rapid development of the research in inverse problems. A number of theories and algorithm have been developed to solve these inverse problems. A central feature of the inverse problem is that they are ill-posed. In 1902 Jacques Hadamard [3] formulated three conditions for the well-posedness of mathematical models of physical phenomena. The conditions are namely:

- A solution exists
- The solution is unique
- The solution depends continuously on the problem data

Problems involving models that satisfy all of these conditions are called well-posed. On the other hand, if one or more conditions are not satisfied, the problem is called to be ill-posed. In general inverse problems are ill-posed with solutions that depend on the data [4]. Due to the ill-posedness of the inverse problem, further regularization [5] is required to stabilize the computational solutions. The regularization methods transform the ill-posed problem into a family of well-posed problems indexed by the regularization parameter. In real-world problems another important aspect of the inverse problem is that the measurements contain noise. Then the model (1.1) can be considered as

$$y = A(x) + \eta \tag{1.2}$$

where $\eta \in \mathbb{R}^m$ represents both observational noise and noise due to the model. Because of the noise, y is not the image of A. So, simply inverting A on the data y will not be possible. Moreover, the noise η may not be known to us, often statistical properties of noise are known. So we can not subtract η from the observed data y to obtain the image of A. Probabilistic approach enable us to overcome these difficulties [4]. In addition, the statistical inverse problem has many advantages compared to classical deterministic methods [6]. In the theory of statistical inverse problems these are reformulated as problems of statistical inference through Bayesian statistics. Recently many methods using statistical inference to solve inverse problems are proposed including Bayesian

inference method [7]. In Bayesian statistics all the variables are considered as random variables. The randomness casts in all the variables is coded to the probability densities of the variables. These densities are related to the unknowns and with the data of the problem from which one looks for the characteristic values: average value, value of the largest probability, dispersion, correlations etc. From the statistical view point the solution of an inverse problem is the probability distribution of the quantity of interest called *posterior distribution* with all the available information has been incorporated in the model that describes the degree of confidence about the quantity after measurement has been done [7]. The mean of the *posterior probability distribution* is given by the *maximum a posteriori* (MAP) point. The inverse problem leads us to an optimization problem [7].

We consider a statistical inverse problem with high-dimensional parameter spaces within the framework of Bayesian inference with Gaussian noise and prior probability densities. The Bayesian formulation of linear statistical inverse problem with Gaussian noise and prior is related to an approximately related least squares minimization problem [8]. The parameter to observable map is chosen to be the discretized parabolic PDE. The main objective of the thesis is to formulate the statistical inverse problem as PDE constrained optimization problem for different cases of the prior matrix such as mass matrix, 2nd order and 4th order Gaussian smoothness priors which leads to saddle point systems. Also propose block diagonal preconditioners that requires robust Scur complement approximation to implement the preconditioned MINRES algorithm.

1.2 Plan of the Thesis

To achieve the objectives of this thesis explained in the previous section the plan of work is as follows. The work in the thesis is spread over to six chapters: Chapter one gives the scientific overview and the plan of the thesis. Chapter two contains some basic concepts and a review of some fundamental definitions, theorems and results from linear algebra and saddle point systems. In chapter three we introduce the statistical inverse problem, by discussing the Bayesian framework of statistical inverse problem with Gaussian noise and prior. Then we formulate the large scale statistical inverse problem governed by the parabolic PDE. The discretization via finite element is also discussed here. The saddle point system is derived for prior matrix as mass matrix, 2nd order and 4th order Gaussian smoothness prior. Chapter four is devoted to the implementation of the preconditioned MINRES algorithm. Here five efficient Schur complement approximation is proposed for

the system derived for the different prior matrices. For the case of spectrally neutral prior it is shown that the eigenvalues of $(\hat{S}^{-1}S) \in [\frac{1}{2}, 1)$, where $\hat{S}$ is the approximation of Schur complement S. In the cases of smoothness prior we first make an approximation $\tilde{S}_i$ $(i = 1, 2, 3, 4)$ to the Schur complement which in turn can be robustly approximated by a matrix $\hat{S}_i$. In Chapter five we present the numerical results for different regularization parameter and efficiency of the algorithm is presented by tables showing the number of iterations and time taken by the algorithm. In final, Chapter six contains the conclusions of the thesis.

Chapter 2

Preliminary Concepts

In this chapter we present some basic definitions, notation and some fundamental theorems of linear algebra, matrix theory and saddle point systems. For the sake of brevity discussion and the proofs of the theorems are omitted, since details of every topic and theorem are available in the listed references.

2.1 Linear Algebra and Matrix Theory

Definition 2.1. A matrix with special structure, that has few nonzero entries is called *sparse matrix*. Usually standard discretization of PDEs lead to large and sparse system. A sparse matrix is a matrix that motivates special techniques to take the benefits of the large number of zero elements and their locations. Details about sparse matrices, their properties, representations, and operations can be found in [9].

Definition 2.2. [10] Let X be a vector space. A real valued function $\|.\| : X \to \mathbb{R}$ is said to be a norm on X if it satisfies the following properties:

1. $\|x\| \geq 0$ and $\|x\| = 0$ if and only if $x = 0$;
2. $\|x + y\| \leq \|x\| + \|y\|$;
3. $\|\alpha x\| = |\alpha|\|x\|$

for any $x, y \in X$ and $\alpha \in \mathbb{R}$. Let $x \in \mathbb{C}^n$, then the vector p norm of x is defined as

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}, \qquad \text{for} \qquad 1 \leq p < \infty \tag{2.1}$$

In particular, when $p = 1, 2$, and ∞, we have

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|,$$

$$\|x\|_2 = \left(\sum_{i=1}^{n} |x_i|^2\right)^{\frac{1}{2}},$$

$$\|x\|_\infty = \max_{1 \leq i \leq n} |x_i|.$$

Definition 2.3. [11] Given a nonsingular symmetric positive definite matrix $A \in \mathbb{C}^{n \times n}$, the A-norm (elliptic norm) generated by the A- inner product on $\mathbb{C}^{n \times 1}$ is

$$\|x\|_A = < x, x >_A = < Ax, x >$$

Definition 2.4. A projector or projection matrix P is a square matrix that satisfies

$$P = P^2.$$

Such a matrix is also known as *idempotent matrix* [10].

Definition 2.5. [9] Let $A \in \mathbb{R}^{n \times n}$ and $v \neq 0 \in \mathbb{R}^n$ then,

$$\mathcal{K}_k(A, v) \equiv span\{v, Av, A^2v, \ldots A^{k-2}v, A^{k-1}v\} \tag{2.2}$$

is called the Krylov subspace associated to A and v.

Theorem 2.6. *[10, 12] Let the columns of $V_{k+1} = [v_1, v_2 \ldots, v_{k+1}] \in \mathbb{R}^{n \times (k+1)}$ form an orthogonal basis for $\mathcal{K}_k(A, v_1)$, then there exists an $(k+1) \times k$ unreduced upper Hessenberg matrix*

$$\hat{H}_k = \begin{pmatrix} h_{11} & h_{12} & \ldots & h_{1k} \\ h_{21} & h_{22} & \ldots & h_{2k} \\ & \ddots & \ddots & \vdots \\ & & h_{k,k-1} & h_{k,k} \\ & & & h_{k+1,k} \end{pmatrix}, \tag{2.3}$$

such that

$$AV_k = V_{k+1}\hat{H}_k. \tag{2.4}$$

Conversely, a matrix V_{k+1} with orthonormal columns satisfies a relation of the form in (2.5) only if the columns of V_{k+1} form the basis for $\mathcal{K}_k(A, v_1)$.

Definition 2.7. Let the column of $V_{k+1} = [V_k, v_{k+1}] \in \mathbb{R}^{n\times(k+1)}$ form an orthogonal basis. If there exists a Hessenberg matrix $\hat{H}_k \in \mathbb{R}^{k+1\times k}$ of the form (2.1) so that

$$AV_k = V_{k+1}\hat{H}_k. \tag{2.5}$$

then (2.5) is called (unreduced) Arnoldi decomposition of order k.

By a suitable partition of $\hat{H}_k$, we can write (2.5) as

$$AV_k = \begin{bmatrix} V_k & v_{k+1} \end{bmatrix} \begin{bmatrix} H_k \\ h_{k+1,k}e_k^T \end{bmatrix} = V_kH_k + h_{k+1,k}v_{k+1}e_k^T \tag{2.6}$$

where,

$$H_k = \begin{pmatrix} h_{11} & h_{12} & \dots & h_{1k} \\ h_{21} & h_{22} & \dots & h_{2k} \\ & \ddots & \ddots & \vdots \\ & & & h_{k,k} \end{pmatrix}. \tag{2.7}$$

By the orthogonality property of v_{k+1}, (2.6) yields (for details see [10], [13])

$$H_k = V_k^T AV_k. \tag{2.8}$$

Here, H_k is a projection of A onto the Krylov subspaces $\mathcal{K}_k(A, v)$.

Definition 2.8. [12] Let $A \in \mathbb{R}^{n\times n}$ and let the columns of $V_k \in \mathbb{R}^{n\times k}$ be orthonormal. The $k \times k$ matrix $H_k = V_k^T AV_k$ is called *Rayleigh quotient*, an eigenvalue λ of H_k is

called *Ritz value*, and if v is an eigenvector of H_k associated with λ, then $V_k v$ is called *Ritz vector* belonging to λ.

2.2 Basic Concepts of Saddle Point System

In this thesis we will formulate and attempt to solve a problem in the saddle point form. This section is devoted to discuss a few properties of saddle point systems. The general saddle point system is defined as:

$$\begin{bmatrix} A & B_1^T \\ B_2 & -C \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix} \tag{2.9}$$

where, $A \in \mathbb{R}^{n\times n}$, $B_1, B_2 \in \mathbb{R}^{m\times n}$, $C \in \mathbb{R}^{m\times m}$ with $n \geq m$.

For a generalized saddle point system the constituent blocks A, B_1, B_2 and C satisfy one or more of the following conditions [14]:

C1 A is symmetric: $A = A^T$

C2 the symmetric part of A, $H \equiv \frac{1}{2}(A + A^T)$

C3 $B_1 = B_2 = B$

C4 C is symmetric ($C = C^T$) and positive semidefinite

C5 $C = 0$(the zero matrix)

Note that $C5$ implies $C4$. Large scale saddle point systems arise in many areas of computational science and engineering (see [14] for a list of application area).

Definition 2.9. [15, 16] Let

$$\mathcal{A} = \begin{bmatrix} A & B \\ C & D \end{bmatrix} \tag{2.10}$$

be an $(m+n)\times(m+n)$ block matrix, where A, B, C, D are matrices of size $m\times m$, $m\times n$, $n\times m$, $n\times n$. Then the Schur complement of the block D of the matrix $\mathcal{A}$ is an $m\times m$ matrix given by

$$A - BD^{-1}C \qquad \text{provided } D^{-1} \text{ exists,} \tag{2.11}$$

and the Schur complement of the block A of the matrix $\mathcal{A}$ is an $n \times n$ matrix given by

$$D - CA^{-1}B \qquad \text{provided } D^{-1} \text{ exists.} \tag{2.12}$$

In the next part we discuss some algebraic properties of the saddle point system following [14].

2.2.1 Block factorizations and the Schur complement

If A is nonsingular (when A is positive definite on the kernel of B which needs to have full rank) the saddle point matrix $\mathcal{A}$ can be factorize into following block triangular factorization:

$$\mathcal{A} = \begin{bmatrix} A & B_1^T \\ B_2 & -C \end{bmatrix} = \begin{bmatrix} I & 0 \\ B_2A^{-1} & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & S \end{bmatrix} \begin{bmatrix} I & A^{-1} \\ 0 & I \end{bmatrix} \tag{2.13}$$

where S is the Schur complement of A. The factorization (2.13) is very important because a number of important properties of saddle point system can be derived on the basis of (2.13). There are also two other equivalent factorizations:

$$\mathcal{A} = \begin{bmatrix} A & 0 \\ B_2 & S \end{bmatrix} \begin{bmatrix} I & A^{-1}B_1^T \\ 0 & I \end{bmatrix} \tag{2.14}$$

and

$$\mathcal{A} = \begin{bmatrix} I & 0 \\ B_2A^{-1} & I \end{bmatrix} \begin{bmatrix} A & B_1^T \\ 0 & S \end{bmatrix}. \tag{2.15}$$

In many applications the matrix A is singular [17]. In that case augmented Lagrangian techniques [18–20] can be used to replace the original saddle point system with an equivalent system having the same solution, but now the $(1,1)$ block is nonsingular.

2.2.2 Solvability conditions

Consider A to be nonsingular, then the block decompositions (2.13)-(2.15) imply that $\mathcal{A}$ is nonsingular if and only if S is nonsingular. In order to comment on the invertibility of $S = -(C + B_2A^{-1}B_1^T)$ it is required to place some restrictritions on the matrices A, B_1, B_2 and C.

2.2.3 Symmetric case

First we consider the standard saddle point system (2.9) with A symmetric positive definite, $B_1 = B_2 = B$ and $C = 0$. The Schur complement $S = -BA^{-1}B^T$, is symmetric negative definite. So, S and hence $\mathcal{A}$ is invertible if and only if B^T has full column rank (i.e. if and only if $rank(B) = m$,) since S is symmetric and negative definite. Then the saddle point problem (2.9) has unique solution. For the case $C \neq 0$ see [14]. The above discussion can be summarize by the following theorem.

Theorem 2.10. *Let A is symmetric positive definite, $B_1 = B_2 = B$, and C is symmetric positive semidefinite. If $ker(C) \cap ker(B^T) = \{0\}$, then the saddle point matrix $\mathcal{A}$ is nonsingular. That is, $\mathcal{A}$ is invertible if B has full rank.*

If A is indefinite then $\mathcal{A}$ may be singular, even if B has full rank. However, $\mathcal{A}$ will be invertible if A is definite on $ker(B)$ [14]. For the case of A being symmetric positive semidefinite, the following theorem holds.

Theorem 2.11. *Consider that A is symmetric positive semidefinite, $B_1 = B_2 = B$ has full rank, and $C = 0$. Then a necessary and sufficient condition for the saddle point matrix $\mathcal{A}$ to be nonsingular is $ker(A) \cap ker(B) = \{0\}$.*

Proof: See [14].

The proof of the above theorem shows that for $\mathcal{A}$ to be nonsingular the rank of A must be at least $n - m$.

2.2.4 The inverse of a saddle point matrix

As we discussed earlier a saddle point matrix $\mathcal{A}$ is invertible if and only if $S = -(C + B_2A^{-1}B_1^T)$ is nonsingular, the inverse of $\mathcal{A}$ is given by,

$$\mathcal{A}^{-1} = \begin{bmatrix} A & B_1^T \\ B_2 & -C \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B_1^T S^{-1} B_2 A^{-1} & -A^{-1}B_1^T S^{-1} \\ -S^{-1}B_2A^{-1} & S^{-1} \end{bmatrix}. \tag{2.16}$$

In the case of A is singular but C is nonsingular a similar expression can be derived assuming the matrix $A + B_1^T C^{-1} B_2$, Schur complement of C in $\mathcal{A}$, is nonsingular.

2.2.5 Spectral properties of saddle point matrices

In this subsection we discuss a few facts on the spectral properties of saddle point matrices relevant in solving the system by iterative methods.

Eigenvalues: The symmetric case

Let us consider that A is symmetric positive definite, $B_1 = B_2 = B$ has full rank, and C is symmetric positive semidefinite (possibly zero). Then from (2.13) we get

$$\begin{bmatrix} I & 0 \\ -BA^{-1} & I \end{bmatrix} \begin{bmatrix} A & B^T \\ B & -C \end{bmatrix} \begin{bmatrix} I & -A^{-1}B^T \\ 0 & I \end{bmatrix} = \begin{bmatrix} A & 0 \\ 0 & S \end{bmatrix} \tag{2.17}$$

where, $S = -(C + BA^{-1}B^T)$ is symmetric negative definite. Hence $\mathcal{A}$ is congruent to the block diagonal matrix $\begin{bmatrix} A & 0 \\ 0 & S \end{bmatrix}$. Thus, according to Sylvester's law of inertia the saddle point matrix $\mathcal{A}$ is indefinite with n positive and m negative eigenvalues [21]. If B is rank deficient, Sylverter's law of inertia is also true, as long as S remains negative definite. For the case of S is rank deficient, suppose that $rank(S) = m - r$, $\mathcal{A}$ has n positive, $m - r$ negative and r zero eigenvalues. If A is considered positive semidefinite, then this result holds, provided the condition $ker(A) \cap ker(B) = \{0\}$ is satisfied. Generally, unless m is very small, the matrix $\mathcal{A}$ is highly indefinite, means that it has many eigenvalues of both signs.

The following theorem [22] gives the eigenvalue bounds for the saddle point matrix with $B_1 = B_2 = B$ and $C = 0$.

Theorem 2.12. *Assume A is symmetric positive definite, $B_1 = B_2 = B$ has full rank and $C = 0$. Let μ_1 and μ_n denote the largest and smallest eigenvalues of A, and let σ_1 and σ_m denote the largest and smallest singular values of B. Also let $\sigma(\mathcal{A})$ denote the spectrum of $\mathcal{A}$. Then*

$$\sigma(\mathcal{A}) \subset I^- \cup I^+$$

where,

$$I^- = \left[\frac{1}{2}\left(\mu_n - \sqrt{\mu_n^2 + 4\sigma_1^2}\right), \left(\mu_1 - \sqrt{\mu_n^2 + 4\sigma_m^2}\right)\right]$$

and

$$I^+ = \left[\mu_n, \frac{1}{2}\left(\mu_1 - \sqrt{\mu_1^2 + 4\sigma_1^2}\right)\right].$$

2.2.6 Preconditioner and preconditioning in linear system

Saddle point systems that arise in practice can be very poorly conditioned. So in order to develop and apply solution algorithms extra care should be taken. For the sake of simplicity consider a standard saddle point problem where $A = A^T$ is positive definite, $B_1 = B_2 = B$ has full rank, and $C = 0$. Here $\mathcal{A}$ is symmetric and the spectral condition number of the system is given by

$$\kappa(\mathcal{A}) = \frac{max|\lambda(\mathcal{A})|}{min|\lambda(\mathcal{A})|}. \tag{2.18}$$

If we keep $\lambda_{max}(A)$ and $\sigma_{max}(B)$ constant, then from Theorem 2.12 it can be said that the condition number of $\mathcal{A}$ grows unboundedly as either $\mu_n = \lambda_{min}(A)$ or $\sigma_m = \sigma_{min}(B)$ goes to zero. This growth of the condition number of $\mathcal{A}$ indicates that the rate of convergence of the Krylov subspace methods deteriorates when the problem size increases. Preconditioning is applied to get rid of this problem.

In order to discuss a preconditioner and preconditioning, let us consider that we want to solve the linear system:

$$Ax = b. \tag{2.19}$$

Definition 2.13. [23, 24] A preconditioner P of a matrix A is a matrix such that $P^{-1}A$ has a smaller condition number than A. Preconditioners are used in iterative methods to solve a linear system $Ax = b$ for x because the rate of convergence for most iterative linear solvers increases as the condition number of a matrix decreases as a result of preconditioning. In preconditioning techniques instead of solving the original linear system $Ax = b$, we solve either the right preconditioned system:

$$AP^{-1}Px = b \tag{2.20}$$

by solving

$$AP^{-1}y = b \qquad \text{for } y \tag{2.21}$$

and

$$Px = y \qquad \text{for } x \tag{2.22}$$

or the left preconditioned system

$$P^{-1}(Ax - b) = 0. \tag{2.23}$$

Preconditioning attempts to improve the spectral properties of the system matrix [14]. The convergence of MINRES depends only on the eigenvalues of the generalized eigenvalue problem $Ax = \lambda Px$ [25]. Since the preconditioned matrix has exactly three distinct eigenvalues, the preconditioned MINRES (minimal residual method) will terminate after

three iterations irrespective of the size of the discrete problem. Preconditioned iterative solvers are applied for many problems, such as 3D PDE discretizations, where direct solvers usually don't work. Iterative solvers can be used as matrix-free methods, i.e. become the only choice if the coefficient matrix A cannot be not stored explicitly, but is accessed by evaluating matrix-vector products.

Chapter 3

Statistical Inverse Problem

In this chapter we present the Bayesian framework for statistical inverse problems for the general case of of Bayes' theorem and continue with the linear parameter-to-observable map with Gaussian noise and prior densities. Then we formulate our large scale statistical inverse problem governed by the discretization of a three dimensional parabolic partial differential equation within the framework of Bayesian inference with Gaussian noise and prior probability densities. Then we discretized the problem using the finite element method. In the first section we present the Bayesian framework for statistical inverse problems and in the 2nd section we present the Gaussian smoothness priors following [8] and [7].

3.1 Bayesian Framework for Statistical Inverse Problems

The Bayesian framework is a methodology to associate prior assumptions in a statistical way. A prior probability density describes the potential values that the parameter can take. A posterior probability density describes how these potential values are affected by the measurements. The main concern of the ill-posed inverse problems is the non-uniqueness. Multiple values of the parameters may be consistent with the observations. The least squares minimization techniques for ill-posed problems need regularization for selecting the one solution that has largest regularity among the multiple parameter values which results in a single deterministic estimate of the unknown parameters [8]. On the other hand Bayesian estimation of the unknown, is a probability density that suggests the credibility of any given point estimation. In Bayesian estimation of statistical inverse

problem all the parameters are considered as random variables and hence the parameter -to- observable map is written as $g\colon \mathbb{R}^{n\times k} \to \mathbb{R}^n$ as

$$Y = g(X, E),$$

where X, Y, and E are random variables. The variable $\boldsymbol{x} \in \mathbb{R}^n$ is the vector of the model parameters to be recovered and is the realization of the random variable X, $\boldsymbol{e} \in \mathbb{R}^k$ is the vector of errors caused by both model error and observation noise and the realization of the random variable E and $\boldsymbol{y} \in \mathbb{R}^m$ is a vector of observables with $\boldsymbol{y}_{obs}$ the actual observation values and realization of the random variable Y.

Let us assume that we know the joint probability density of X and Y which is denoted by $\pi(\boldsymbol{x}, \boldsymbol{y})$. Then the probability density function $\pi_{prior}\colon \mathbb{R}^n \to \mathbb{R}$ which describes the additional information about the parameters X, is defined by the marginal density of the unknown X, i.e.,

$$\pi_{prior}(\boldsymbol{x}) = \int_{\mathbb{R}^m} \pi(\boldsymbol{x}, \boldsymbol{y}) dy.$$

On the other hand, if we would know the value of the unknown $X = \boldsymbol{x}$, then the conditional probability density of Y given this information, would be

$$\pi(\boldsymbol{y}|\boldsymbol{x}) = \frac{\pi(\boldsymbol{x}, \boldsymbol{y})}{\pi_{prior}(\boldsymbol{x})}, \quad \text{provided that } \pi_{prior}(\boldsymbol{x}) \neq 0.$$

This conditional probability is called the likelihood function since it describes the likelihood of different measurement outcomes with $X = \boldsymbol{x}$ given. That is the likelihood function $\pi(\boldsymbol{y} \mid \boldsymbol{x})$ describes the relationship between the observables $\boldsymbol{y}$ and the unknown parameter $\boldsymbol{x}$,

Again we assume that the measured data $Y = \boldsymbol{y}_{obs}$ is given. Then the conditional probability distribution

$$\pi(\boldsymbol{x}|\boldsymbol{y}_{obs}) = \frac{\pi(\boldsymbol{x}, \boldsymbol{y}_{obs})}{\pi(\boldsymbol{y}_{obs})}$$

is the posterior probability density $\pi_{post}\colon \mathbb{R}^n \to \mathbb{R}$ on the model parameter X,

where

$$\pi(\boldsymbol{y}_{obs}) = \int_{\mathbb{R}^n} \pi(\boldsymbol{x}, \boldsymbol{y}_{obs}) dx \neq 0.$$

and let $\pi_{noise}\colon \mathbb{R}^n \to \mathbb{R}$ describes the modeling error and the observation noise.

In the Bayesian framework, the inverse problem is expressed as:
Given data $Y = \boldsymbol{y}_{obs}$*, find the conditional probability distribution* $\pi(\boldsymbol{x}|\boldsymbol{y}_{obs})$ *of the variable* X*.*

We state Bayes' theorem for inverse problems as:

Theorem 3.1 (Bayes' Theorem)**.** *[7] Assume that the random variable* $X \in \mathbb{R}^n$ *has a known prior probability density* $\pi_{prior}(\boldsymbol{x})$ *and the data consists of observed value* $\boldsymbol{y}_{obs}$ *of an observable random variable* $Y \in \mathbb{R}^k$ *such that* $\pi(\boldsymbol{y}_{obs}) > 0$*. Then the posterior probability distribution of* X*, given the data* $\boldsymbol{y}_{obs}$*, is given by*

$$\pi_{post}(\boldsymbol{x}) = \pi(\boldsymbol{x}|\boldsymbol{y}_{obs}) = \frac{\pi_{prior}(\boldsymbol{x})\pi(\boldsymbol{y}_{obs} \mid \boldsymbol{x})}{\pi(\boldsymbol{y}_{obs})}$$

where the marginal density

$$\pi(\boldsymbol{y}_{obs}) = \int_{\mathbb{R}^n} \pi(\boldsymbol{x}, \boldsymbol{y}_{obs}) dx = \int_{\mathbb{R}^n} \pi(\boldsymbol{y}_{obs}|\boldsymbol{x}) \pi_{prior}(\boldsymbol{x}) dx$$

plays the role of a norming constant and is usually of little importance [7].

Thus applying Bayes' theorem the posterior probability density $\pi_{post}\colon \mathbb{R}^n \to \mathbb{R}$ on the model parameter X is obtained as

$$\pi_{post}(\boldsymbol{x}) \propto \pi_{prior}(\boldsymbol{x})\pi(\boldsymbol{y}_{obs} \mid \boldsymbol{x}).$$

That is, the posterior probability density on the parameter X is proportional to the product of the prior probability on the parameter X and the conditional probability of the observable Y given the parameter X.

Often in classical inverse problems the noise is modeled as additive and mutually independent with the unknown X. Thus if we consider the additive noise the parameter-to-observable map is

$$Y = f(X) + E \tag{3.1}$$

where $X \in \mathbb{R}^n, Y, E \in \mathbb{R}^m$ with X and E are mutually independent random variables. Suppose that we know the probability distribution of the noise E is $\pi_{noise}(\boldsymbol{e})$. Since X and E are mutually independent to each other, if we fix $X = \boldsymbol{x}$ the probability distribution of E does not change when conditioned on $X = \boldsymbol{x}$ [26]. That is,

$$\pi(\boldsymbol{e}|\boldsymbol{x}) = \pi(\boldsymbol{e}) = \pi_{noise}(\boldsymbol{e}). \tag{3.2}$$

On the other hand if $X = \boldsymbol{x}$ is fixed we can say that Y conditioned on $X = \boldsymbol{x}$ is distributed like E, the probability density being translated by $f(\boldsymbol{x})$, that is the likelihood function is

$$\pi(\boldsymbol{y}_{obs}|\boldsymbol{x}) = \pi_{noise}(\boldsymbol{e}) = \pi_{noise}(\boldsymbol{y}_{obs} - f(\boldsymbol{x})).$$

Where $f\colon \mathbb{R}^n \to \mathbb{R}^m$ and $\boldsymbol{e} \in \mathbb{R}^m$ describes both the modeling error of f and observation noise. Which implies $E = Y - f(X)$. We consider that both X and E are mutually independent. In case of X and E are not mutually independent we need to know the conditional density of the noise, for detail see [7].

Hence from Bayes' theorem we can write

$$\pi_{post}(\boldsymbol{x}) \propto \pi_{prior}(\boldsymbol{x})\pi_{noise}(\boldsymbol{y}_{obs} - f(\boldsymbol{x})).$$

In statistical inverse problems, the most important and challenging step is to construct the prior density. Actually, it depends on the nature of the prior information. In most cases our prior knowledge of the unknown is qualitative in nature. Then the challenge is to transform the qualitative information into quantitative information from which the prior density is encoded. The most commonly used probability densities in Statistical Inverse problems are Gaussian prior densities, since they are easy to construct but lead to an explicit estimator.

The Gaussian n-variate random variable is defined now.

Definition 3.2. [7] Let $x_0 \in \mathbb{R}^n$and $\Gamma \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix, denoted by $\Gamma > 0$ in the sequel. A Gaussian n-variate random variable X with mean x_0 and covariance Γ is a random variable with probability density

$$\pi(x) = \left(\frac{1}{2\pi|\Gamma|}\right)^{n/2} exp\left(-\frac{1}{2}(x-x_0)^T\Gamma^{-1}(x-x_0)\right)$$

where, $|\Gamma| = det(\Gamma)$. We use the notation $X \sim \mathcal{N}(x_0, \Gamma)$ to mean that X is a Gaussian random variable with mean x_0 and covariance Γ.

Thus if the prior probability density of X and the probability density of the error E are both Gaussian then the prior and noise probability density function can be written as

$$\pi_{prior}(\boldsymbol{x}) \propto exp\left(-\frac{1}{2}(\boldsymbol{x}-\bar{\boldsymbol{x}}_{prior})^T\Gamma_{prior}^{-1}(\boldsymbol{x}-\bar{\boldsymbol{x}}_{prior})\right)$$

$$\pi_{noise}(\boldsymbol{e}) \propto exp\left(-\frac{1}{2}(\boldsymbol{e}-\bar{\boldsymbol{e}})^T\Gamma_{noise}^{-1}(\boldsymbol{e}-\bar{\boldsymbol{e}})\right)$$

in which $\bar{\boldsymbol{x}}_{prior} \in \mathbb{R}^n$ is the mean of the model parameter prior pdf, $\bar{\boldsymbol{e}} \in \mathbb{R}^m$ is the mean of the noise pdf, $\Gamma_{prior} \in \mathbb{R}^{n\times n}$ is the covariance matrix of the prior pdf, and $\Gamma_{noise} \in \mathbb{R}^{m\times m}$ is the covariance of the noise pdf [8]. For Gaussian noise and prior Bayes' theorem can be written as

$$\pi_{post}(\boldsymbol{x}) \propto exp\left(-\frac{1}{2}\|\boldsymbol{x}-\bar{\boldsymbol{x}}_{prior}\|^2_{\Gamma^{-1}_{prior}} - \frac{1}{2}\|\boldsymbol{y}_{obs}-f(\boldsymbol{x})-\bar{\boldsymbol{e}}\|^2_{\Gamma^{-1}_{noise}}\right).$$

If $f(x)$ is non-linear then the posterior probability density may not be Gaussian even though the prior and noise probability density are Gaussian. We choose the parameter-to-observable map to be linear i.e.

$$f(X) = \mathbf{A}X$$

Here, $\mathbf{A} \in \mathbb{R}^{m\times n}$ is the linear operator that maps the parameter $\boldsymbol{x}$ to the observable $\boldsymbol{y}$ through the solution of a large-scale discretized PDE. In this case $\pi_{post}(\boldsymbol{x})$ is Gaussian with mean $\bar{\boldsymbol{x}}_{post} \in \mathbb{R}^n$ since the parameter-to-observable map is linear.

The mean $\bar{\boldsymbol{x}}_{post} \in \mathbb{R}^n$ is given by the *maximum a posteriori* (MAP) point, i.e.

$$\bar{\boldsymbol{x}}_{post} = \boldsymbol{x}_{MAP} = \arg\max_{x \in \mathbb{R}^n} \pi_{post}(\boldsymbol{x})$$

provided that such maximization exists. Please note that even if the maximizer exists, it may not be unique. The possible non-existence and non-uniqueness show that the single-estimator based approaches to inverse problems may not be satisfactory [7]. To find a MAP estimate the solution of an optimization problem is required.

Another popular point estimate is the *conditional mean* (CM) of the unknown X conditioned on the data $\boldsymbol{y}$, defined as

$$\boldsymbol{x}_{CM} = E\{\boldsymbol{x}|\boldsymbol{y}\} = \int_{\mathbb{R}^n} x\pi(\boldsymbol{x}|\boldsymbol{y})dx,$$

provided that the integral converges [7].

In the case of purely Gaussian random variables the center point $\bar{\boldsymbol{x}}_{post}$ is simultaneously the maximum a posteriori estimate and the conditional mean [7], that is,

$$\bar{\boldsymbol{x}}_{post} = \boldsymbol{x}_{CM} = \boldsymbol{x}_{MAP} \tag{3.3}$$

Thus finding the MAP point is equivalent to solving the weighted least squares optimization problem [8] i.e.

$$\bar{\boldsymbol{x}}_{post} = \arg\min_{x \in \mathbb{R}^n} \left(\frac{1}{2} \|\boldsymbol{x} - \bar{\boldsymbol{x}}_{prior}\|^2_{\Gamma^{-1}_{prior}} + \frac{1}{2} \|\boldsymbol{y}_{obs} - A\boldsymbol{x} - \bar{\boldsymbol{e}}\|^2_{\Gamma^{-1}_{noise}} \right). \tag{3.4}$$

This equation is equivalent to solving the regularized deterministic inverse problem where Γ^{-1}_{prior} works as the regularization operator and Γ^{-1}_{noise} is a weighting of the data misfit term.

The covariance matrix of the posterior probability density function model parameter $\pi_{post}(\boldsymbol{x}) \in \mathbb{R}^{n \times n}$ is given by the inverse of the Hessian matrix of the least squares objective function i. e.,

$$\Gamma_{post} = \left(\mathbf{A}^T \Gamma^{-1}_{noise} \mathbf{A} + \Gamma^{-1}_{prior} \right)^{-1} \tag{3.5}$$

Since prior probability density, noise probability density and the posterior probability densities are Gaussian we can write

$$\begin{aligned} \pi_{prior}(\boldsymbol{x}) &= \mathcal{N}(\bar{\boldsymbol{x}}_{prior}, \Gamma_{prior}), \\ \pi_{noise}(\boldsymbol{e}) &= \mathcal{N}(\bar{\boldsymbol{e}}, \Gamma_{noise}), \\ \pi_{post}(\boldsymbol{x}) &= \mathcal{N}(\bar{\boldsymbol{x}}_{post}, \Gamma_{post}). \end{aligned}$$

3.2 Gaussian Smoothness Priors

In this section we are going to discuss the Gaussian smoothness priors in the light of [7]. Let us consider that we are interested in solving (3.1) by a classical regularization method. Suppose further that $x \in \mathbb{R}^n$ represents the discretized values of some function $f\colon D \subset \mathbb{R}^n \to \mathbb{R}$, which we know a priori to be twice continuously differentiable over D. If we want to express the above information as a constraint, we introduce the generalized Tikhonov functional

$$\boldsymbol{T}(x) = \|Ax - y\|^2 + \alpha\|Lx\|^2 \tag{3.6}$$

where $\alpha > 0$ is the regularization parameter and the Tikhonov matrix $L\colon \mathbb{R}^n \mapsto \mathbb{R}$ is a discrete approximation of a differential operator in $\mathbb{R}^n$. Hence the posterior potential $V(x|y)$ defined by

$$\pi(x|y) \propto exp(-V(x|y))$$

where

$$V(x|y) = \frac{1}{2\sigma^2}\|Ax - y\|^2 + \frac{\alpha}{2\sigma^2}\|Lx\|^2 =: \frac{1}{2\sigma^2}\boldsymbol{T}(x) \tag{3.7}$$

We assume that the data are corrupted by white noise with variance σ^2. Then minimizing $\boldsymbol{T}(x)$ is equivalent to maximizing the conditional density $x \mapsto exp(-V(x|y))$. Hence the choice for the prior distribution is

$$\pi_{prior}(\boldsymbol{x}) \propto exp(-\frac{1}{2\gamma^2}\|Lx\|^2) \; with \; \gamma^2 = \frac{\sigma^2}{\alpha} \tag{3.8}$$

If L is used as the gradient operator then (3.8) gives us

$$\pi_{prior}(\boldsymbol{x}) \propto exp(-\frac{1}{2\gamma^2}\|\nabla x\|^2) \; with \; \gamma^2 = \frac{\sigma^2}{\alpha} \tag{3.9}$$

The prior given in (3.9) is called *2nd order smoothness prior*.

Again, if L is used as the Laplacian operator then (3.8) gives us

$$\pi_{prior}(\boldsymbol{x}) \propto exp(-\frac{1}{2\gamma^2}\|\Delta x\|^2) \; with \; \gamma^2 = \frac{\sigma^2}{\alpha} \tag{3.10}$$

The prior given in (3.10) is called *4th order smoothness prior*. We will discuss both Dirichlet and Neumann boundary conditions, as indicated in [7].

3.3 Problem Description

In this section we formulate a large scale statistical inverse problem governed by a parabolic PDE. Here we introduce the operator of the parameter-to-observable map by discretization of the parabolic PDE

$$u_t - \Delta u = 0 \qquad in \; \Omega \times (0,T) \tag{3.11a}$$

$$u = 0 \qquad on \; \partial\Omega \times (0,T) \tag{3.11b}$$

$$u = u_0(x) \qquad on \; \Omega \times \{t = 0\} \tag{3.11c}$$

Here $u_0(x)$ describes a temperature profile at time $t = 0$. On the boundary $\partial\Omega$ of the domain Ω the temperature is kept at 0. The forward problem is to find the temperature at $t = T$. The inverse problem consists in finding the initial temperature given the temperature profile over time and prior information on the initial temperature.

As discussed in the earlier section the Bayesian formulation of a linear statistical inverse problem with Gaussian noise and prior is related to an approximately-weighted

least squares minimization problem [8]. We choose to define our noise and prior pdf by discretizing the infinite dimensional functional

$$J(t,x) \quad = \quad \frac{\beta_{noise}}{2}\int_\Omega\int_0^T (u-u_{obs})^2 dxdt + \frac{\beta_{prior}}{2}\int_\Omega u_0^2 dx \tag{3.12}$$

where $\beta_{noise}, \beta_{prior} > 0$ are the regularization parameters and $u(x,t)$ satisfies the time dependent parabolic PDE (3.11a). Here $\Omega \in \mathbb{R}^d$, T is the final time and we have given the observed state $u_{obs}(x,t)$. This as a classical PDE- constrained optimization problem as given in [27]. The goal of the optimization process is to drive the state variable $u(x,t)$ as close as possible to the observed state using the control $u_0(x)$. The discretization of the infinite dimensional functional (3.12) for the posterior mean (3.4) gives,

$$\bar{\boldsymbol{u}}_{0,post} = \arg\min_{\boldsymbol{u}_0} \left(\frac{1}{2}(\boldsymbol{u}-\boldsymbol{u}_{obs})^T\Gamma_{noise}^{-1}(\boldsymbol{u}-\boldsymbol{u}_{obs}) + \frac{1}{2}\boldsymbol{u}_0^T\Gamma_{prior}^{-1}\boldsymbol{u}_0\right) \tag{3.13}$$

where $\Gamma_{prior} \in \mathbb{R}^{n\times n}$ is the covariance matrix of the prior pdf and $\Gamma_{noise} \in \mathbb{R}^{m\times m}$ is the covariance of the noise pdf and $\boldsymbol{u}$ satisfies the discretization of the parabolic PDE (3.11).

The adjoint PDE is derived [28] as:

$$-p_t - \Delta p + \beta_{noise}(u-u_{obs}) = 0 \qquad in\ \Omega\times(0,T) \tag{3.14a}$$

$$p = 0 \qquad on\ \partial\Omega\times(0,T) \tag{3.14b}$$

$$p(x,T) = 0 \qquad on\ \Omega \tag{3.14c}$$

3.4 Finite Element Discretization

In this thesis we follow the discretize then optimize strategy [29]. We discretize the objective functional (3.12) and the constraint equation (3.11a) using finite elements [25, 30, 31]. In this case the prior covariance operator is chosen as $L = I$. The time discretization of the PDE (3.11a) using a backward Euler method with time step τ gives

$$\frac{u^k - u^{k-1}}{\tau} - \Delta u^k = 0 \quad in \quad \Omega \qquad for\ k = 1, 2, 3, ..., N_t \tag{3.15a}$$

$$u^k = 0 \quad on \quad \partial\Omega \qquad for\ k = 1, 2, 3, ..., N_t \tag{3.15b}$$

$$u^0 = u_0(x) \quad in \quad \Omega \tag{3.15c}$$

Then the weak formulation of (3.15) is to find $u^k \in \mathcal{H}_0^1(\Omega)$ such that

$$a(u^k, v) = L(v) \tag{3.16}$$

where,

$$a(u^k, v) = \int_\Omega (u^k - \tau \Delta u^k) v dx \tag{3.17}$$

and

$$L(v) = \int_\Omega u^{k-1} v dx \tag{3.18}$$

where $v \in V_h$ be the space of piecewise continuous linear functions. Let $\{\phi_1, \phi_2, \phi_3, \ldots, \phi_{n-1}, \phi_n\}$ be the basis of dimension n. Integrating by parts and using Green's first identity $a(u^k, v)$ becomes

$$a(u^k, v) = \int_\Omega u^k v dx + \tau \int_\Omega \nabla v \cdot \nabla u^k dx. \tag{3.19}$$

For $v \in V_h$ we can write

$$u^k = \sum_{j=1}^{n} u_j^k \phi_j \tag{3.20}$$

and

$$u^{k-1} = \sum_{j=1}^{n} u_j^{k-1} \phi_j \tag{3.21}$$

with

$$v = \phi_i \qquad 1 \leq i \leq n \tag{3.22}$$

where, $u_1^k, u_2^k, u_3^k, \ldots, u_{n-1}^k, u_n^k$ and $u_1^{k-1}, u_2^{k-1}, u_3^{k-1}, \ldots, u_{n-1}^{k-1}, u_n^{k-1}$ the unknown coefficients to be determined.

Then we have

$$Mu^k + \tau K u^k = M u^{k-1} \tag{3.23}$$

where,

$$\begin{aligned} M &= [M_{ij}], \qquad 1 \leq i, j \leq n \\ K &= [K_{ij}], \qquad 1 \leq i, j \leq n \\ u^k &= [u_1^k, u_2^k, u_3^k, \ldots, u_{n-1}^k, u_n^k]^T \\ u^{k-1} &= [u_1^{k-1}, u_2^{k-1}, u_3^{k-1}, \ldots, u_{n-1}^{k-1}, u_n^{k-1}]^T \end{aligned}$$

with

$$M_{ij} = \int_\Omega \phi_i \phi_j dx \tag{3.25a}$$

$$K_{ij} = \int_\Omega \nabla\phi_i \nabla\phi_j dx. \tag{3.25b}$$

Here the matrix M is called the mass matrix and the matrix K is called the stiffness matrix.

Thus for each time step we have a system of the form

$$(M+\tau K)u^k \quad = Mu^{k-1} \qquad for\ k=1,2,3,....,N_t. \tag{3.26}$$

Putting all of the equation (3.26) together, the one-shot discretization for N_t time steps becomes

$$\underbrace{\begin{bmatrix} M+\tau K & & & & \\ -M & M+\tau K & & & \\ & \ddots & \ddots & & \\ & -M & M+\tau K & & \\ & & -M & M+\tau K & \end{bmatrix}}_{\mathcal{K}} \begin{bmatrix} u^1 \\ u^2 \\ \vdots \\ \\ u^{N_t} \end{bmatrix} = \begin{bmatrix} M\boldsymbol{u}_0 \\ 0 \\ \vdots \\ \\ 0 \end{bmatrix} \tag{3.27}$$

In the objective functional the observed state $u_{obs}(x,t)$ is defined for the whole time interval while we use the initial condition as the control $u_0(x)$. We discretize the observed state first with respect to time by trapezoidal rule and then by weak formulation. We discretize the control by weak formulation. Thus the discretize objective functional is obtained as

$$J(u,u_0) \quad = \quad \frac{\tau\beta_{noise}}{2}(\boldsymbol{u}-\boldsymbol{u}_{obs})^T\mathcal{M}_n(\boldsymbol{u}-\boldsymbol{u}_{obs}) + \frac{\beta_{prior}}{2}\boldsymbol{u}_0{}^T\mathcal{W}_p\boldsymbol{u}_0 \tag{3.28}$$

Where,

$$\mathcal{M}_n = \begin{bmatrix} \frac{1}{2}M & & & & \\ & M & & & \\ & & \ddots & & \\ & & & M & \\ & & & & \frac{1}{2}M \end{bmatrix} \quad and \quad \mathcal{W}_p = M. \tag{3.29}$$

Here, $\boldsymbol{u} = \begin{bmatrix}(u^1)^T, & (u^2)^T, & \dots & (u^{N_t})^T\end{bmatrix}^T$, $\boldsymbol{u}_{obs} = \begin{bmatrix}{u_{obs1}}^T, & {u_{obs2}}^T, & \dots & {u_{obsN_t}}^T\end{bmatrix}^T$ and $\boldsymbol{u}_0 = \begin{bmatrix}u_{01}, & u_{02}, & \dots & u_{0n}\end{bmatrix}^T$ are the state and observed state at time step 1 to N_t and the control variable at the initial time step of backward Euler scheme. Here M represents the lumped matrix for our choice of finite elements on Ω.

According to [32] everything derived in this thesis also holds for consistent mass matrices.

Hence the Lagrangian for the functional $J(u, u_0)$ with respect to the constraint (3.27) is given by

$$L(\boldsymbol{u}, \boldsymbol{u}_0, \boldsymbol{p}) = \frac{\tau\beta_{prior}}{2}(\boldsymbol{u} - \boldsymbol{u}_{obs})^T\mathcal{M}_n(\boldsymbol{u} - \boldsymbol{u}_{obs}) + \frac{\beta_{noise}}{2}{\boldsymbol{u}_0}^T\mathcal{W}_p\boldsymbol{u}_0 + p^T(-\mathcal{K}\boldsymbol{u} + d)$$

with, $d = \begin{bmatrix} M\boldsymbol{u}_0 + c \\ c \\ \vdots \\ \\ c \end{bmatrix}$ where, c represents the boundary condition of the PDE, in our case 0.

The first order optimality condition [33] for the system can be written as

$$\begin{bmatrix} \tau\beta_{noise}\mathcal{M}_n & 0 & -\mathcal{K}^T \\ 0 & \beta_{prior}\mathcal{W}_p & \mathcal{M}_1^T \\ -\mathcal{K} & \mathcal{M}_1 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{u}_0 \\ \boldsymbol{p} \end{bmatrix} = \begin{bmatrix} \tau\beta_{noise}\mathcal{M}_n u_{obs} \\ 0 \\ 0 \end{bmatrix} \tag{3.30}$$

where, $\mathcal{M}_1 = \begin{bmatrix} M \\ 0 \\ \vdots \\ 0 \end{bmatrix}$.

The discretization of the problem and solution via first order optimality condition on a Lagrangian leads to a linear system in saddle point form:

$$\underbrace{\begin{bmatrix} A & B^T \\ B & 0 \end{bmatrix}}_{\mathcal{A}} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix} \tag{3.31}$$

where $A \in \mathbb{R}^{n\times n}$ is a symmetric positive definite or positive semi definite and $B \in \mathbb{R}^{m\times n}$, $m < n$ is a matrix of full rank [32].

3.5 Approximation of the Prior Matrix

In the discretized system (3.30) the $(2,2)$ block $\mathcal{W}_p$ represents the prior covariance matrix. If we choose the Tikhonov matrix L as the discrete gradient operator, then the Tikhonov regularization technique gives us 2nd order smoothness prior given in (3.9). Thus the approximation gives us the prior matrix as $\mathcal{W}_p = \Delta_h$, where Δ_h is the discrete Laplacian operator. Now the question is which boundary conditions to choose. We here consider Dirichlet boundary conditions and Neumann boundary conditions.

3.5.1 Saddle point system for 2nd order smoothness prior with Dirichlet boundary condition

The 2nd order smoothness prior with Dirichlet boundary conditions gives us the prior matrix $\mathcal{W}_p = K_D$. Where K_D is the discrete Laplacian with Dirichlet boundary conditions. And in that case the system (3.30) would become:

$$\begin{bmatrix} \tau\beta_{noise}\mathcal{M}_n & 0 & -\mathcal{K}^T \\ 0 & \beta_{prior}K_D & \mathcal{M}_1^T \\ -\mathcal{K} & \mathcal{M}_1 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{u}_0 \\ \boldsymbol{p} \end{bmatrix} = \begin{bmatrix} \tau\beta_{noise}\mathcal{M}_n u_{obs} \\ 0 \\ 0 \end{bmatrix} \tag{3.32}$$

3.5.2 Saddle point system for 2nd order smoothness prior with Neumann boundary condition

Similarly, if we would use the regularization technique by the 2nd order smoothnes prior with Neumann boundary condition then the the prior matrix would be $\mathcal{W}_p = K_N$, where

K_N is the discrete Laplacian with Neumann boundary condition. The system (3.30) then becomes:

$$\begin{bmatrix} \tau\beta_{noise}\mathcal{M}_n & 0 & -\mathcal{K}^T \\ 0 & \beta_{prior}K_N & \mathcal{M}_1^T \\ -\mathcal{K} & \mathcal{M}_1 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{u}_0 \\ \boldsymbol{p} \end{bmatrix} = \begin{bmatrix} \tau\beta_{noise}\mathcal{M}_n u_{obs} \\ 0 \\ 0 \end{bmatrix} \tag{3.33}$$

Alternatively, the choice of the Tikhonov matrix L as the discrete Laplacian gives 4th order smoothness prior given in (3.10). Then the the prior matrix becomes $\mathcal{W}_p = \Delta_h^T \Delta_h$, where Δ_h is the discrete Laplacian. Since Δ_h is symmetric, we have $\mathcal{W}_p = \Delta_h^2$. As in the previous case we choose the Laplacian operator for two cases of boundary conditions.

3.5.3 Saddle point system for 4th order smoothness prior with Dirichlet boundary condition

The use of 4th order smoothness prior with Dirichlet boundary condition gives $\mathcal{W}_p = K_D^2$, where K_D again the discrete Laplacian with Dirichlet boundary conditions. Hence the system (3.30) would become

$$\begin{bmatrix} \tau\beta_{noise}\mathcal{M}_n & 0 & -\mathcal{K}^T \\ 0 & \beta_{prior}K_D^2 & \mathcal{M}_1^T \\ -\mathcal{K} & \mathcal{M}_1 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{u}_0 \\ \boldsymbol{p} \end{bmatrix} = \begin{bmatrix} \tau\beta_{noise}\mathcal{M}_n u_{obs} \\ 0 \\ 0 \end{bmatrix} \tag{3.34}$$

3.5.4 Saddle point system for 4th order smoothness prior with Neumann boundary condition

In a similar manner the use of 4th order smoothness prior with Neumann boundary condition gives a new prior $\mathcal{W}_p = K_N^2$, where K_N is the discrete Laplacian with Neumann boundary condition. Then the system (3.30) given as

$$\begin{bmatrix} \tau\beta_{noise}\mathcal{M}_n & 0 & -\mathcal{K}^T \\ 0 & \beta_{prior}K_N^2 & \mathcal{M}_1^T \\ -\mathcal{K} & \mathcal{M}_1 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{u} \\ \boldsymbol{u}_0 \\ \boldsymbol{p} \end{bmatrix} = \begin{bmatrix} \tau\beta_{noise}\mathcal{M}_n u_{obs} \\ 0 \\ 0 \end{bmatrix} \tag{3.35}$$

In this situation the system matrices given in (3.30) ,(3.32), (3.33) (3.34) and (3.35) are symmetric and indefinite.

We now briefly show definiteness of mass matrix M and stiffness matrix K.

$$\begin{aligned}
\mathbf{v}^T M\mathbf{v} &= \sum_{j=1}^{n}\sum_{i=1}^{n}\mathbf{v}_j\left(\int_\Omega \phi_j\phi_i dx\right)\mathbf{v}_j \\
&= \int_\Omega\left(\sum_{j=1}^{n}\mathbf{v}_j\phi_j\right)\left(\sum_{i=1}^{n}\mathbf{v}_j\phi_j\right)dx \\
&= \int_\Omega v_h.v_h dx \\
&\geq 0
\end{aligned}$$

where $\mathbf{v}$ is a general coefficient vector and $v_h = \sum_{j=1}^{n}\mathbf{v}_j\phi_j \in V_0^h$. Therefore M is positive semidefinite. To proof that M is positive definite let $\mathbf{v}^T M\mathbf{v} = 0$ that implies that

$$\int_\Omega v_h \cdot v_h dx = 0. \tag{3.37}$$

That is, $v_h = 0$ which implies $\mathbf{v} = 0$ but $\mathbf{v}$ is nonzero. So, $\mathbf{v}^T M\mathbf{v} > 0$. Hence the mass matrix M is positive definite.

Again for the stiffness matrix K

$$\begin{aligned}
\mathbf{v}^T K\mathbf{v} &= \sum_{j=1}^{n}\sum_{i=1}^{n}\mathbf{v}_j\left(\int_\Omega \nabla\phi_j\nabla\phi_i dx\right)\mathbf{v}_j \\
&= \int_\Omega\left(\sum_{j=1}^{n}\mathbf{v}_j\nabla\phi_j\right)\left(\sum_{i=1}^{n}\mathbf{v}_j\nabla\phi_j\right)dx \\
&= \int_\Omega \nabla v_h.\nabla v_h dx \\
&\geq 0
\end{aligned}$$

where $v_h = \sum_{j=1}^{n} \mathbf{v}_j \phi_j \in V_0^h$. Therefore K is positive semidefinite. To proof that K is positive definite let $\mathbf{v}^T K \mathbf{v} = 0$ that implies

$$\int_\Omega \nabla v_h . \nabla v_h dx = 0. \tag{3.39}$$

That is, $\mathbf{v}^T K \mathbf{v} = 0$ if and only if $\nabla v_h = 0$, i.e. if and only if v_h is constant in Ω. Since $v_h \in V_0^h$, it is continuous up to the boundary and is zero on $\partial\Omega_D$. Thus $\nabla v_h = 0$ implies $v_h = 0$. Finally, since the test functions are a basis for V_0^h, $v_h = 0$ implies $\mathbf{v} = 0$, but $\mathbf{v}$ is nonzero. So, $\mathbf{v}^T K \mathbf{v} > 0$. Hence the stiffness matrix K is positive definite.

In this thesis we consider iterative solvers to solve the systems (3.30) ,(3.32), (3.33) (3.34) and (3.35) since the direct methods [34] perform very well for 2D problems, on the other hand for the discretized 3D problems they quickly run of memory [35]. As the systems (3.30) ,(3.32), (3.33) (3.34) and (3.35) are symmetric and indefinite the minimal residual method (MINRES) [36] described in Chapter 4 is the method of choice.

Chapter 4

MINRES and the Preconditioner

The formulation of the statistical inverse problem leads us to solve the sparse saddle point system (3.31). A number of direct and iterative solvers are available. The sparse direct solver is based on a direct elimination of equations, direct elimination requires the factorization of an initially very sparse linear system of equations into a lower triangular matrix followed by forward and backward substitution using these triangular systems. The lower triangular matrix factors are typically much denser than the initial assembled sparse matrix, hence the large disk or in-core memory requirements for direct methods. Sparse direct solvers seek to minimize the cost of factorizing the matrix as well as the size of the factor using sophisticated equation reordering strategies.

On the other hand we can use iterative solvers, where the solution is obtained through an iterative process that successively updates an initial guess to a solution that is within an acceptable tolerance of the exact solution. Iterative solvers do not require a matrix factorization and typically iterate towards the solution using a series of sparse matrix-vector multiplications along with a preconditioning step, both of which require less memory and time per iteration than a direct factorization.

Iterative methods for solving saddle point systems are subdivided into two broad classes, one is called segregated and the other is called coupled or all at once method. The segregated method uses the techniques of solving two linear systems of reduced order. Two main types of segregated methods are Schur complement reduction method and the null space method. On the other hand an all at once method uses the whole system matrix and then approximates the solution simultaneously for all variables. The iterative techniques for all at once methods include Krylov subspace methods. In the next section we briefly discuss the Krylov subspace methods following the book [25].

4.1 Krylov Subspace Methods

Suppose we want to solve the system

$$\mathcal{A}u = f \tag{4.1}$$

The k dimensional Krylov subspace generated by the square matrix $\mathcal{A}$ and a nonzero vector x is given by

$$\mathcal{K}_k(\mathcal{A}, x) \equiv span\{x, \mathcal{A}x, \mathcal{A}^2x, \dots \mathcal{A}^{k-2}x, \mathcal{A}^{k-1}x\} \tag{4.2}$$

That is, the kth krylov subspace associated with the pair $(\mathcal{A}, x)$ is a space consisting of the linear combination of the linearly independent vectors $\{x, \mathcal{A}x, \mathcal{A}^2x, \dots \mathcal{A}^{k-2}x, \mathcal{A}^{k-1}x\}$. The solution of (4.1) can be approximated from $\mathcal{K}_k(\mathcal{A}, x)$ for any particular vector x and different increasing values of k. Since $\mathcal{K}_k(\mathcal{A}, x) \equiv span\{x, \mathcal{A}x, \mathcal{A}^2x, \dots \mathcal{A}^{k-2}x, \mathcal{A}^{k-1}x\}$ any member y of $\mathcal{K}_k(\mathcal{A}, x)$ is of the form

$$y = \sum_{j=0}^{k-1} \alpha_j \mathcal{A}^j x. \tag{4.3}$$

Alternatively we can write

$$y = q_{k-1}(\mathcal{A})x \tag{4.4}$$

where, $q_{k-1}(t)$ is a polynomial given by

$$q_{k-1}(t) = \sum_{j=0}^{k-1} \alpha_j t^j \tag{4.5}$$

Thus y is specified by the coefficients of polynomial $q_{k-1}(t)$. Now our concern is how well $u = \mathcal{A}^{-1}f$ can be approximated by a vector from $q_{k-1}(\mathcal{A})x$. If $\mathcal{A}$ is a square matrix, then according to the Cayley Hamilton Theorem

$$p(\mathcal{A}) = \mathcal{A}^n + c_{n-1}\mathcal{A}^{n-1} + \cdots + c_1\mathcal{A} + (-1)^n det(\mathcal{A})I = 0 \tag{4.6}$$

which gives us

$$\mathcal{A}^{-1} = \frac{(-1)^k}{det(\mathcal{A})}(\mathcal{A}^{k-1} + c_{k-2}\mathcal{A}^{k-2} + \cdots + c_1 I) \tag{4.7}$$

That is $\mathcal{A}^{-1} = q_{n-1}(\mathcal{A})$, where q_{n-1} is the polynomial on the right-hand side of (4.7). Hence,

$$u = q_{n-1}(\mathcal{A})f \tag{4.8}$$

So, f is a natural choice of the initial vector x.

More generally, if u^0 is a starting vector and $r^0 = f - \mathcal{A}u^0$ is the starting residual, then the kth solution is given by

$$u^{(k)} = u^{(0)} + \mathcal{K}_k(\mathcal{A}, r^0) \tag{4.9}$$

The most common Krylov subspace methods are the Conjugate gradient, GMRES (generalized minimum residual), and MINRES (minimal residual) methods.

The Arnoldi iteration [37] is an iterative method to find the eigenvalues of non-Hermitian matrices [9] on the other hand the Lanczos iteration is used to find the eigenvalues of Hermitian matrices. The Conjugate gradient method [38] is an iterative method used where the system matrix has to be symmetric and positive definite [9]. The minimal residual (MINRES) method of Paige and Saunders [36] is used for nonsingular symmetric possibly indefinte matrices and on the other hand the generalized minimal residual

(GMRES) of Saad and Schultz [39], [40] is used for generalized nonsingular matrices. In the next section we are going to discuss the MINRES method we have used to solve our problem in the light of the book [25].

4.2 The MINRES method

The formulation of the statistical inverse problem leads us to solve the saddle point system (3.31), where A is symmetric so, $\mathcal{A}$ is symmetric (see for example [25]). $\mathcal{A}$ is symmetric and indefinite, i.e. $\mathcal{A}$ has eigenvalues with both positive and negetive real parts. For symmetric and indefinite system MINRES is a robust method which is derived from Lanczos algorithm.

To solve the system (4.1) let $\mathbf{v}^{(1)}$ be a vector such that the Euclidean norm $\|\mathbf{v}^{(1)}\| = 1$ and let $\mathbf{v}^{(0)} = 0$. An orthogonal basis for $\mathcal{K}_k(\mathcal{A}, \mathbf{v}^{(1)})$ is constructed by the recurrence relation

$$\gamma_{j+1}\mathbf{v}^{(j+1)} = \mathcal{A}\mathbf{v}^{(j)} - \delta_{(j)}\mathbf{v}^{(j)} - \gamma_{(j)}\mathbf{v}^{(j-1)}, \qquad 1 \leq j \leq k \tag{4.10}$$

where $\delta_j = \langle \mathcal{A}\mathbf{v}^{(j)}, \mathbf{v}^{(j)} \rangle$, and γ_{j+1} is chosen so that $\|\mathbf{v}^{(j+1)}\| = 1$. Note that $\delta_j > 0$ for positive definite $\mathcal{A}$, where as the sign of γ_{j+1} is not prescribed.

Let $V_k = [\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots, , \mathbf{v}^{(k)}]$ is the matrix containing $\mathbf{v}^{(j)}$ in its jth column, $j = 1, 2, \ldots, k$ called the Lanczos vectros and let T_k denotes the symmetric tridiagonal matrix

$$tridiag[\gamma_j, \delta_j, \gamma_{j+1}], \qquad 1 \leq j \leq k$$

Then (4.10) is equivalent to

$$\mathcal{A}V_k = V_k T_k + \gamma_{k+1}[\mathbf{0}, \mathbf{0}, \ldots, \mathbf{v}^{(k+1)}] \tag{4.11}$$

using the orthogonality of V_k we have

$$V_k^T \mathcal{A} V_k = T_k \tag{4.12}$$

That is, in the Lanczos algorithm we find an orthonormal set $\mathbf{v}^{(j)}, j = 1, 2, \ldots, k$, which is a basis for the Krylov subspace $\mathcal{K}_k(\mathcal{A}, x)$.

Minimal residual method (MINRES) is an important Krylov subspace method derived form Lanczos algorithm by Paige and Saunders [36]. This method is a robust algorithm for indefinite coefficient matrices as well as symmetric positive definite matrices. The MINRES algorithm works by minimizing the Euclidean norm of the residual $\|\mathbf{r}^{(k)}\|$.

We start by assuming that

$$\mathbf{u}^{(k)} = \mathbf{u}^{(0)} + V_k \mathbf{y}^{(k)} \tag{4.13}$$

where $\mathbf{y}^{(k)} = [y_k^{(1)}, y_k^{(2)}, \ldots, , y_k^{(k)}]$ is a vector of dimension k such that the Euclidean norm of the residual $\|\mathbf{r}^{(k)}\|$ is minimal. Using (4.13), the residual of $\mathbf{u}^{(k)}$ can be written as

$$\mathbf{r}^{(k)} \quad = \quad f - \mathcal{A}\mathbf{u}^{(k)} = \mathbf{r}^{(0)} - \mathcal{A} V_k \mathbf{y}^{(k)}.$$

This gives us,

$$\|\mathbf{r}^{(k)}\| = \|V_{k+1}(\|\mathbf{r}^{(0)}\|\mathbf{e}_1 - \hat{T}_k \mathbf{y}^{(k)})\| \tag{4.14}$$

as $v_1 = \frac{r_0}{\|r_0\|}$, where, $\hat{T}_k \in \mathbb{R}^{k+1 \times k}$ is the tridiagonal matrix T_k with an additional final row $[0, 0, \ldots, , \gamma_{k+1}]$ and $\mathbf{e}_1$ represents the unit vector of dimensions $k + 1$. Thus, the MINRES algorithm can be written as (see [25]):

Algorithm 4.1 THE MINRES METHOD

1: $\mathbf{v}^{(0)} = 0, \mathbf{w}^{(0)} = 0, \mathbf{w}^{(1)} = 0$
2: Choose $\mathbf{u}^{(0)}$, compute $\mathbf{v}^{(1)} = f - \mathcal{A}\mathbf{u}^{(0)}$, set $\gamma_1 = \|\mathbf{v}^{(1)}\|$
3: Set$\eta = \gamma_1$, $s_0 = s_1 = 0$, $c_0 = c_1 = 1$
4: **for** $j = 1$ until convergence **do**
5: $\quad \mathbf{v}^{(j)} = \mathbf{v}^{(j)}/\gamma_j$
6: $\quad \delta_j = \langle \mathcal{A}\mathbf{v}^{(j)}, \mathbf{v}^{(j)} \rangle$
7: $\quad \mathbf{v}^{(j+1)} = \mathcal{A}\mathbf{v}^{(j)} - \delta_j \mathbf{v}^{(j)} - \gamma_j \mathbf{v}^{(j-1)}$ ▷ Lanczos process
8: $\quad \gamma_{j+1} = \|\mathbf{v}^{(j+1)}\|$
9: $\quad \alpha_0 = c_j \delta_j - c_{j-1} s_j \gamma_j$ ▷ Update the QR factorization
10: $\quad \alpha_1 = \sqrt{\alpha_0^2 + \gamma_{j+1}^2}$
11: $\quad \alpha_2 = s_j \delta_j + c_{j-1} c_j \gamma_j$
12: $\quad \alpha_3 = s_j \gamma_j$
13: $\quad c_{j+1} = \alpha_0/\alpha_1$; $s_{j+1} = \gamma_{j+1}/\alpha_1$ ▷ Givens Rotation
14: $\quad \mathbf{w}^{(j+1)} = (\mathbf{v}^{(j)} - \alpha_3 \mathbf{w}^{(j-1)} - \alpha_2 \mathbf{w}^{(j)})/\alpha_1$
15: $\quad \mathbf{u}^{(j)} = (\mathbf{u}^{(j-1)} + c_{j+1} \eta \mathbf{w}^{(j+1)}$
16: $\quad \eta = -s_{j+1} \eta$
17: $\quad \langle$ Test for convergence $\rangle$
18: **end for**

4.2.1 Preconditioning

The MINRES algorithm states that the Euclidean norm $\|\mathbf{r}^{(k)}\|$ is minimal over the whole Krylov subspace. We describe vectors in the translated Krylov subspace in terms of polynomials in $\mathcal{A}$ operating on the initial residual $\mathbf{r}^{(0)}$. The residual vectors defined by the MINRES iterates satisfy $\mathbf{r}^{(k)} = p_k(\mathcal{A})\mathbf{r}^{(0)}$, where $p_k \in \Pi_k$, Π_k is the set of real polynomials of degree less than or equal to k. Also $p_k(0) = 1$ and p_k is optimal in the sense that $\|\mathbf{r}^{(k)}\|$ is minimal. Let the eigenvectors of the matrix $\mathcal{A}$ are $\mathbf{v}_j$, then expanding in terms of the eigenvectors

$$\mathbf{r}^{(0)} = \sum_j \alpha_j \mathbf{v}_j, \qquad \mathcal{A}\mathbf{v}_j = \lambda_j \mathbf{v}_j \tag{4.15}$$

then

$$\mathbf{r}^{(k)} = p_k(\mathcal{A}) \sum_j \alpha_j \mathbf{v}_j = \sum_j \alpha_j p_k(\lambda_j) \mathbf{v}_j \tag{4.16}$$

and

$$\|\mathbf{r}^{(k)}\| = \min_{p_k \in \Pi_k, p_k(0)=1} \|\sum_j \alpha_j p_k(\lambda_j)\mathbf{v}_j\| \tag{4.17}$$

or

$$\begin{aligned}
\|\mathbf{r}^{(k)}\| &= \min_{p_k \in \Pi_k, p_k(0)=1} \|\sum_j \alpha_j p_k(\lambda_j)\mathbf{v}_j\| \\
&= \min_{p_k \in \Pi_k, p_k(0)=1} \left(\sum \alpha_j^2 p_k(\lambda_j)^2 \langle \mathbf{v}_j, \mathbf{v}_j \rangle\right)^{1/2} \\
&\leq \min_{p_k \in \Pi_k, p_k(0)=1} \max_j |p_k(\lambda_j)| \left(\sum \alpha_j^2 \langle \mathbf{v}_j, \mathbf{v}_j \rangle\right)^{1/2}
\end{aligned}$$

hence

$$\|\mathbf{r}^{(k)}\| \leq \min_{p_k \in \Pi_k, p_k(0)=1} \max_j |p_k(\lambda_j)| \|\mathbf{r}^{(0)}\| \tag{4.19}$$

From (4.19) the convergence depends only on the eigenvalues of the matrix. The indefinite matrices have both positive and negative real eigenvalues λ_j. The minimum number of iterations required for convergence for an indefinite system must be more than a positive-definite system with the same number of positive eigenvalues. So for the rapid convergence, preconditioning will be important. We have to choose the preconditioner in such a way that the preconditioner does not destroy the symmetry of the system. Otherwise, iterative methods for non-symmetric systems would have to be implemented for the system. In order to maintain the symmetric in the preconditioned system a symmetric and positive definite preconditioner $\mathcal{P} = HH^T$ is required. Thus the symmetric system preconditioned as

$$H^{-1}\mathcal{A}H^{-T}\boldsymbol{y} = H^{-1}f, \qquad \boldsymbol{y} = H^T u \tag{4.20}$$

has the same solution as the system (4.1) since $H^{-1}\mathcal{A}H^{-T}$ is a congruence transformation on $\mathcal{A}$. Thus by the Sylvester Law of inertia the coefficient matrix has the same number of positive, zero and negative eigenvalues as $\mathcal{A}$.

For any approximate solution $\boldsymbol{u}^{(k)}$ to the system (4.1), the corresponding residual for the preconditioned system is

$$H^{-1}(b - \mathcal{A}\boldsymbol{u}^{(k)}) = H^{-1}\boldsymbol{r}^{(k)} = H^T\boldsymbol{z}^{(k)}, \tag{4.21}$$

where $\boldsymbol{r}^{(k)}$ is the residual for the original system (4.1) and $\boldsymbol{z}^{(k)} = \mathcal{P}^{-1}\boldsymbol{r}^{(k)}$ is the preconditioned residual. Thus we have,

$$\|H^{-1}\boldsymbol{r}^{(k)}\| = \|\boldsymbol{r}^{(k)}\|_{\mathcal{P}^{-1}} \tag{4.22}$$

and the convergence estimate for the preconditioned MINRES method becomes

$$\|\mathbf{r}^{(k)}\|_{\mathcal{P}^{-1}} \leq \min_{p_k \in \Pi_k, p_k(0)=1} \max_j |p_k(\lambda_j)| \|\mathbf{r}^{(0)}\|_{\mathcal{P}^{-1}} \tag{4.23}$$

where the maximum is taken over the eigenvalues of $\mathcal{P}^{-1}\mathcal{A}$. Thus the MINRES algorithm for the system (4.20) in terms of the vectors associated with (4.1) can be written as given in Algorithm 4.2 taken from [25].

Algorithm 4.2 THE PRECONDITIONED MINRES METHOD

1: $\mathbf{v}^{(0)} = 0, \mathbf{w}^{(0)} = 0, \mathbf{w}^{(1)} = 0$
2: Choose $\mathbf{u}^{(0)}$, compute $\mathbf{v}^{(1)} = f - \mathcal{A}\mathbf{u}^{(0)}$
3: Solve $\mathcal{P}\mathbf{z}^{(1)} = \mathbf{v}^{(1)}$, set $\gamma_1 = \sqrt{\langle \mathbf{z}^{(1)}, \mathbf{v}^{(1)} \rangle}$
4: Set$\eta = \gamma_1$, $s_0 = s_1 = 0$, $c_0 = c_1 = 1$
5: **for** $j = 1$ until convergence **do**
6: $\quad \mathbf{z}^{(j)} = \mathbf{z}^{(j)}/\gamma_j$
7: $\quad \delta_j = \langle \mathcal{A}\mathbf{z}^{(j)}, \mathbf{z}^{(j)} \rangle$
8: $\quad \mathbf{v}^{(j+1)} = \mathcal{A}\mathbf{z}^{(j)} - \delta_j \mathbf{v}^{(j)} - \gamma_j \mathbf{v}^{(j-1)}$ ▷ Lanczos process
9: $\quad$ Solve $\mathcal{P}\mathbf{z}^{(j+1)} = \mathbf{v}^{(j+1)}$,
10: $\quad \gamma_{j+1} = \sqrt{\langle \mathbf{z}^{(j+1)}, \mathbf{v}^{(j+1)} \rangle}$
11: $\quad \alpha_0 = c_j \delta_j - c_{j-1} s_j \gamma_j$ ▷ Update the QR factorization
12: $\quad \alpha_1 = \sqrt{\alpha_0^2 + \gamma_{j+1}^2}$
13: $\quad \alpha_2 = s_j \delta_j + c_{j-1} c_j \gamma_j$
14: $\quad \alpha_3 = s_j \gamma_j$
15: $\quad c_{j+1} = \alpha_0/\alpha_1$; $s_{j+1} = \gamma_{j+1}/\alpha_1$ ▷ Givens Rotation
16: $\quad \mathbf{w}^{(j+1)} = (\mathbf{v}^{(j)} - \alpha_3 \mathbf{w}^{(j-1)} - \alpha_2 \mathbf{w}^{(j)})/\alpha_1$
17: $\quad \mathbf{u}^{(j)} = (\mathbf{u}^{(j-1)} + c_{j+1} \eta \mathbf{w}^{(j+1)}$
18: $\quad \eta = -s_{j+1} \eta$
19: $\quad \langle$ Test for convergence $\rangle$
20: **end for**

4.3 Preconditioning Strategies

Generally the iterative solver is only applied together with a suitable preconditioner $\mathcal{P}$ that speeds up the convergence of the solution procedure [14]. The role of the preconditioner, $\mathcal{P}$ is to reduce the number of iterations required for the convergence of the solution and also the preconditioner should work in such a way that the amount of work the computation requires at each iteration does not increase significantly. For a good preconditioner $\mathcal{P}$ it is not necessary that $\mathcal{P}^{-1}$ be an approximate inverse of $\mathcal{A}$. If $\mathcal{P}$ can be selected with less computing effort such that $\mathcal{P}^{-1}$ be an approximate inverse of $\mathcal{A}$, then $\mathcal{P}$ can be considered as a good preconditioner. The main concern in selecting a good preconditioner should be that the preconditioned matrix $\mathcal{P}^{-1}\mathcal{A}$ has a low degree minimal polynomial. The preconditioned matrix $\mathcal{P}^{-1}\mathcal{A}$ has only few distinct eigenvalues. For symmetric problems, the rate of convergence of Krylov subspace methods such as

CG or MINRES depends on the distribution of the eigenvalues of $\mathcal{A}$. The preconditioned matrix $\tilde{\mathcal{A}} = \mathcal{P}^{-1}\mathcal{A}$ (or $\tilde{\mathcal{A}} = \mathcal{A}\mathcal{P}^{-1}$) will have a smaller spectral condition number and/or a small number of eigenvalues clusters [14]. In general for MINRES the preconditioner is often a blockdiagonal matrix. In [41] it is shown that if we consider the preconditioner for $\mathcal{A}$ to be defined by

$$\mathcal{P} = \begin{bmatrix} A & 0 \\ 0 & S \end{bmatrix} \tag{4.24}$$

where S is the Schur complement of A, then the following result holds

Proposition 4.1. *If a saddle point system (3.31) is preconditioned by the preconditioner given in (4.24) then the preconditioned matrix $\tilde{\mathcal{A}} = \mathcal{P}^{-1}\mathcal{A}$ satisfies*

$$\tilde{\mathcal{A}}(\tilde{\mathcal{A}} - I)(\tilde{\mathcal{A}}^2 - \tilde{\mathcal{A}} - I) = 0 \tag{4.25}$$

Proof. The preconditioned matrix $\tilde{\mathcal{A}}$ is given by

$$\begin{aligned} \tilde{\mathcal{A}} &= \mathcal{P}^{-1}\mathcal{A} \\ &= \begin{bmatrix} I & A^{-1}B^T \\ S^{-1}B & 0 \end{bmatrix} \end{aligned}$$

then

$$\left(\tilde{\mathcal{A}} - \frac{1}{2}I\right)^2 = \begin{bmatrix} \frac{1}{4}I + A^{-1}B^T S^{-1}B & 0 \\ 0 & \frac{5}{4}I \end{bmatrix}$$

But $(\mathcal{A}^{-1}B^T S^{-1}B)^2 = I^2 = I$ implies that $\mathcal{A}^{-1}B^T S^{-1}B$ is a projection, so that

$$\left[\left(\tilde{\mathcal{A}} - \frac{1}{2}I\right)^2 - \frac{1}{4}I\right]^2 = \left[\left(\tilde{\mathcal{A}} - \frac{1}{2}I\right)^2 - \frac{1}{4}I\right]$$

That gives us

$$\tilde{\mathcal{A}}(\tilde{\mathcal{A}} - I)(\tilde{\mathcal{A}}^2 - \tilde{\mathcal{A}} - I) = 0 \tag{4.26}$$

□

Since (4.26) can be factorized into distinct linear factors over $\mathbb{R}$, $\tilde{\mathcal{A}}$ is diagonalizable and has at most four distinct eigenvalues $0, 1, \frac{1}{2} \pm \frac{\sqrt{5}}{2}$. If $\tilde{\mathcal{A}}$ is nonsingular then it has 3 nonzero eigenvalues. This implies that any appropriate Krylov subspace method such as MINRES will converge in 3 iterations [41]. The practical implementation of such a method involves exact construction of the Schur complement S, also to find the solution of the systems with S, and the $(1,1)$ block of the system (3.31). Unfortunately, S is dense even though both $(1,1)$ and $(1,2)$ blocks of the system (3.31) are sparse. Hence, we require to approximate S by $\hat{S}$ and apply the preconditioner

$$\hat{\mathcal{P}} = \begin{bmatrix} A & 0 \\ 0 & \hat{S} \end{bmatrix}. \tag{4.27}$$

Also approximation to A by $\hat{A}$ might be needed, see Stokes equation in [25].

The $(1,1)$ block A for the system (3.30) is given by

$$A = \begin{bmatrix} \tau\beta_{noise}\mathcal{M}_n & 0 \\ 0 & \beta_{prior}\mathcal{W}_p \end{bmatrix} \tag{4.28}$$

where $\mathcal{M}_n$ is given in (3.29) and initially $\mathcal{W}_p = M$. Thus the $(1,1)$ block is a block diagonal matrix consisting multiples of lumped the mass matrices. The inverse of A is given by

$$A^{-1} = \begin{bmatrix} \frac{1}{\tau\beta_{noise}}\mathcal{M}_n^{-1} & 0 \\ 0 & \frac{1}{\beta_{prior}}M^{-1} \end{bmatrix}. \tag{4.29}$$

4.4 Schur Complement Approximation

The necessity of the Schur complement approximation was discussed in the previous section. In this section we present a Schur complement approximation for the system (3.30). If $\mathcal{W}_p = M$ equation (3.30) can be written as a saddle point system (3.31) with

$$A = \begin{bmatrix} \tau\beta_{noise}\mathcal{M}_n & 0 \\ 0 & \beta_{prior}M \end{bmatrix} \qquad \text{and} \qquad B = \begin{bmatrix} -\mathcal{K} & \mathcal{M}_1 \end{bmatrix}.$$

The (negative) Schur complement of the system is given by

$$S = \frac{1}{\tau\beta_{noise}}\mathcal{K}\mathcal{M}_n^{-1}\mathcal{K}^T + \frac{1}{\beta_{prior}}\mathcal{M}_1 M^{-1}\mathcal{M}_1^T. \tag{4.30}$$

The 2nd part $\mathcal{M}_1 M^{-1}\mathcal{M}_1^T$ of (4.30) gives us

$$\begin{bmatrix} M \\ 0 \\ \vdots \\ \\ 0 \end{bmatrix} M^{-1} \begin{bmatrix} M & 0 & \dots & 0 \end{bmatrix} = \begin{bmatrix} M & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}$$

which can be written as $\Gamma_1\mathcal{M}_n^{-1}\Gamma_1, where, \Gamma_1 = \begin{bmatrix} \frac{1}{\sqrt{2}}M & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}$.

So, S in (4.30) can be written as

$$S = \frac{1}{\tau\beta_{noise}}\left[\mathcal{K}\mathcal{M}_n^{-1}\mathcal{K}^T + \frac{\tau\beta_{noise}}{\beta_{prior}}\Gamma_1\mathcal{M}_n^{-1}\Gamma_1\right]. \tag{4.31}$$

Now we wish to determine an appropriate approximation of the Schur complement given by (4.31). In order to approximate S in (4.31) let us consider $\phi_1 = \tau\beta_{noise}\mathcal{M}_n$ and $\phi_2 = \Gamma_1$, here both ϕ_1 and ϕ_2 are block diagonal matrices consisting of the mass matrix, hence symmetric positive definite. Define $\mathcal{M}$ as follows

$$\mathcal{M} = \begin{bmatrix} M & & & & \\ & M & & & \\ & & \ddots & & \\ & & & M & \\ & & & & M \end{bmatrix}.$$

Then the Schur complement (4.31) can be written as

$$S = \mathcal{K}\phi_1^{-1}\mathcal{K}^T + \frac{\tau\beta_{noise}}{\beta_{prior}}\phi_2\phi_1^{-1}\phi_2. \tag{4.32}$$

We want to show that $\mathbf{v}^T(\mathcal{K}\Delta + \Delta\mathcal{K}^T)\mathbf{v} > 0$ for all $\mathbf{v} = [\mathbf{v}_1^T, \mathbf{v}_2^T, \ldots, \mathbf{v}_{Nt-1}^T, \mathbf{v}_{Nt}^T]^T$ with $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_{Nt-1}, \mathbf{v}_{Nt} \in \mathbb{R}^n$, and

$$\begin{aligned}
\Delta &= \phi_1^{-1}\phi_2 \\
&= \begin{bmatrix} \frac{1}{2}M & & & & \\ & M & & & \\ & & \ddots & & \\ & & & M & \\ & & & & \frac{1}{2}M \end{bmatrix}^{-1} \begin{bmatrix} \frac{1}{\sqrt{2}}M & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} \\
&= \frac{1}{\tau\beta_{noise}} \begin{bmatrix} \sqrt{2}I & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}.
\end{aligned}$$

The product matrix $\Delta = \phi_1^{-1}\phi_2$ is symmetric and positive definite.

Then,

$$\begin{aligned}
\mathcal{M}\Delta &= \frac{1}{\tau\beta_{noise}} \begin{bmatrix} M & & & & \\ & M & & & \\ & & \ddots & & \\ & & & M & \\ & & & & M \end{bmatrix} \begin{bmatrix} \sqrt{2}I & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix} \\
&= \frac{1}{\tau\beta_{noise}} \begin{bmatrix} \sqrt{2}M & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}
\end{aligned}$$

and

$$\begin{aligned}\Delta\mathcal{M} &= \frac{1}{\tau\beta_{noise}}\begin{bmatrix}\sqrt{2}I & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0\end{bmatrix}\begin{bmatrix}M & & & & \\ & M & & & \\ & & \ddots & & \\ & & & M & \\ & & & & M\end{bmatrix} \\ &= \frac{1}{\tau\beta_{noise}}\begin{bmatrix}\sqrt{2}M & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0\end{bmatrix}.\end{aligned}$$

So, $\Delta\mathcal{M} = \mathcal{M}\Delta$.

Now,

$$\mathcal{K}\Delta + \Delta\mathcal{K}^T = \frac{1}{\tau\beta_{noise}}\begin{bmatrix}2\sqrt{2}(M+\tau K) & -\sqrt{2}M & & \\ -\sqrt{2}M & 0 & \ddots & \\ & \ddots & \ddots & 0 \\ & & 0 & 0\end{bmatrix}$$

and

$$\mathbf{v}^T(\mathcal{K}\Delta + \Delta\mathcal{K}^T)\mathbf{v} = \frac{1}{\tau\beta_{noise}}\left[2\sqrt{2}\tau\mathbf{v}_1^T K\mathbf{v}_1 + \sqrt{2}(\mathbf{v}_1-\mathbf{v}_2)^T M(\mathbf{v}_1-\mathbf{v}_2) + \sqrt{2}\mathbf{v}_1^T M\mathbf{v}_1\right]$$

where we have used the fact that $M\Delta_1 = \Delta_1 M$ with Δ_1 is the $(1,1)$ block of Δ. Since M and K are symmetric and positive definite, all of the above terms are positve. So, we deduce that $\mathbf{v}^T(\mathcal{K}\Delta + \Delta\mathcal{K}^T)\mathbf{v} > 0$, and hence $\mathcal{K}\Delta + \Delta\mathcal{K}^T$ is positive definite. Since $\mathcal{K}\Delta + \Delta\mathcal{K}^T$ is positive definite we are now ready to find the eigenvalue bounds of $\hat{S}^{-1}S$,

where

$$\hat{S} = \left(\mathcal{K} + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\phi_2 \right) \phi_1^{-1} \left(\mathcal{K} + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\phi_2 \right)^T \tag{4.36}$$

and S is given by (4.32).

The eigenvalues of the matrix $\hat{S}^{-1}S$ are bounded by the extreme values of the Rayleigh quotient [42]. Let us consider the Rayleigh quotient R given as

$$R = \frac{\mathbf{p}^T\mathbf{p} + \mathbf{q}^T\mathbf{q}}{\mathbf{p}^T\mathbf{p} + \mathbf{q}^T\mathbf{q} + \mathbf{p}^T\mathbf{q} + \mathbf{q}^T\mathbf{p}} \tag{4.37}$$

where,

$$\mathbf{p} = \frac{1}{\sqrt{\tau\beta_{noise}}}\phi_1^{-\frac{1}{2}}\mathcal{K}^T\mathbf{w} \tag{4.38}$$

and

$$\mathbf{q} = \frac{1}{\sqrt{\beta_{prior}}}\phi_1^{-\frac{1}{2}}\phi_2\mathbf{w} \tag{4.39}$$

then,

$$\mathbf{p}^T\mathbf{q} + \mathbf{q}^T\mathbf{p} = \frac{1}{\sqrt{\tau\beta_{noise}\beta_{prior}}}\mathbf{w}^T\left[\mathcal{K}\phi_1^{-1}\phi_2 + \phi_2\phi_1^{-1}\mathcal{K}^T\right]\mathbf{w} > 0 \tag{4.40}$$

for any vector w since we have $[\mathcal{K}\phi_1^{-1}\phi_2 + \phi_2\phi_1^{-1}\mathcal{K}^T] > 0$. Thus $R < \frac{\mathbf{p}^T\mathbf{p}+\mathbf{q}^T\mathbf{q}}{\mathbf{p}^T\mathbf{p}+\mathbf{q}^T\mathbf{q}} = 1$ as $\mathbf{q} \neq 0$ for any $\mathbf{w}$. Again, for any $\mathbf{p}$ and $\mathbf{q} \neq 0$

$$
\begin{aligned}
& (\mathbf{p}-\mathbf{q})^T(\mathbf{p}-\mathbf{q}) \geq 0 \\
\Leftrightarrow \quad & \mathbf{p}^T\mathbf{p} + \mathbf{q}^T\mathbf{p} - \mathbf{p}^T\mathbf{q} - \mathbf{q}^T\mathbf{p} \geq 0 \\
\Leftrightarrow \quad & \tfrac{1}{2}[\mathbf{p}^T\mathbf{p} + \mathbf{q}^T\mathbf{p} - \mathbf{p}^T\mathbf{q} - \mathbf{q}^T\mathbf{p}] \geq 0 \\
\Leftrightarrow \quad & [\mathbf{p}^T\mathbf{p} + \mathbf{q}^T\mathbf{p}] \geq \tfrac{1}{2}[\mathbf{p}^T\mathbf{p} + \mathbf{q}^T\mathbf{p} + \mathbf{p}^T\mathbf{q} + \mathbf{q}^T\mathbf{p}] \\
\Leftrightarrow \quad & \tfrac{\mathbf{p}^T\mathbf{p}+\mathbf{q}^T\mathbf{q}}{\mathbf{p}^T\mathbf{p}+\mathbf{q}^T\mathbf{q}+\mathbf{p}^T\mathbf{q}+\mathbf{q}^T\mathbf{p}} \geq \tfrac{1}{2} \\
\Leftrightarrow \quad & R \geq \tfrac{1}{2}.
\end{aligned}
$$

Hence we conclude that $\lambda(\hat{S}^{-1}S) \in [\frac{1}{2}, 1)$ which tells us that $\hat{S}$ is a good Schur complement approximation of (4.32).

The above fact motivates us to state the following theorem

Theorem 4.2. *If S and $\hat{S}$ are of the form (4.32) and (4.36) respectively with ϕ_1, ϕ_2 and $\phi_1^{-1}\phi_2$ symmetric and positive definite such that $\mathcal{M}\phi_1^{-1}\phi_2 = \phi_1^{-1}\phi_2\mathcal{M}$ then $\lambda(\hat{S}^{-1}S) \in [\frac{1}{2}, 1)$.*

Hence,

$$
\hat{S} = \frac{1}{\tau\beta_{noise}}(\mathcal{K} + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\Gamma_1)\mathcal{M}_n^{-1}(\mathcal{K} + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\Gamma_1)^T \tag{4.42}
$$

is an appropriate approximation of the Schur complement of (4.31).

It will not be feasible if we apply the inverse of $(\mathcal{K} + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\Gamma_1)$ and its transpose $(\mathcal{K} + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\Gamma_1)^T$ since that implies that we are solving the PDE directly [43]. We will use algebraic multigrid technique [44] for $(\mathcal{K} + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\Gamma_1)$ and its transpose $(\mathcal{K} + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\Gamma_1)^T$, that we require an algebraic multigrid process for the 1st diagonal block $M + \tau K + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}M \in \mathbb{R}^{n\times n}$ and algebraic multigrid process for the remaining blocks of the form $M + \tau K \in \mathbb{R}^{n\times n}$. We apply a fixed (e.g. 4) cycles of such an algebraic multigrid process N_t times to approximate the inverse of $(\mathcal{K} + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\Gamma_1)$ and N_t times to approximate the inverse of $(\mathcal{K} + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\Gamma_1)^T$.

The $(1,1)$ block A for the system (3.32), (3.33), (3.34) and (3.35) is given by

$$A = \begin{bmatrix} \tau\beta_{noise}\mathcal{M}_n & 0 \\ 0 & \beta_{prior}\mathcal{W}_p \end{bmatrix} \tag{4.43}$$

where $\mathcal{M}_n$ is given in (3.29) and $\mathcal{W}_p = \{K_D, K_N, K_D^2, K_N^2\}$. Here the $(1,1)$ block A is a block diagonal matrix. The inverse of A is given by

$$A^{-1} = \begin{bmatrix} \frac{1}{\tau\beta_{noise}}\mathcal{M}_n^{-1} & 0 \\ 0 & \frac{1}{\beta_{prior}}\mathcal{W}_p^{-1} \end{bmatrix}. \tag{4.44}$$

Since the $(1,1)$ block of (4.44) consists of mass matrix which is diagonal, so the inverse $\mathcal{M}_n^{-1}$ can be evaluated easily. On the other hand the $(2,2)$ block of (4.44) is not diagonal the inverse of $(2,2)$ block is approximated by using an algebraic multigrid preconditioner. For the case of 4th order Gaussian smoothness prior in order to approximate the inverse of K_D^2 or K_N^2 the algebraic multigrid preconditioner is applied twice.

4.5 Schur Complement Approximation for Smoothness Prior

The Schur complement approximation discussed above is not applicable if the prior covariance matrix is approximated. Because the structure of $(\mathcal{K} + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\Gamma_1)$ is changed. In the previous case Γ_1 consisted of the mass matrix M. But if the prior covariance matrix is approximated then in the Schurcomplemet approximation (4.42) the $(1,1)$ block of Γ_1 is changed. So, we need to redefine Γ_1 to approximate the Schur complement for the case of approximated priors. In the next sections we are going to discuss the approximation of the Schur complement for the case of prior matrix as 2nd order and 4th order Gaussian smoothness prior with Dirichlet and Neumann boundary conditions respectively.

4.5.1 2nd order smoothness prior with Dirichlet boundary condition

In the case of 2nd order smoothness prior with Dirichlet boundary condition the system to solve is given in (3.32) with

$$A = \begin{bmatrix} \tau\beta_{noise}\mathcal{M}_n & 0 \\ 0 & \beta_{prior}K_D \end{bmatrix} \qquad \text{and} \qquad B = \begin{bmatrix} -\mathcal{K} & \mathcal{M}_1 \end{bmatrix}.$$

The (negative) Schur complement of the system is

$$S_1 = \frac{1}{\tau\beta_{noise}}\mathcal{K}\mathcal{M}_n^{-1}\mathcal{K}^T + \frac{1}{\beta_{prior}}\mathcal{M}_1 K_D^{-1}\mathcal{M}_1^T. \tag{4.45}$$

In the 2nd term $\mathcal{M}_1 K_D^{-1}\mathcal{M}_1^T$ of (4.45) the inverse of K_D can not be found easily. So we approximate K_D by $\hat{K}_D = diag(K_D)$ [45]. Thus, the Schur complement approximation can be written as

$$\tilde{S_1} = \frac{1}{\tau\beta_{noise}}\mathcal{K}\mathcal{M}_n^{-1}\mathcal{K}^T + \frac{1}{\beta_{prior}}\mathcal{M}_1 \hat{K}_D^{-1}\mathcal{M}_1^T. \tag{4.46}$$

The 2nd term of (4.46) $\mathcal{M}_1 \hat{K}_D^{-1}\mathcal{M}_1^T$ gives us

$$\begin{bmatrix} M \\ 0 \\ \vdots \\ 0 \end{bmatrix} \hat{K}_D^{-1} \begin{bmatrix} M & 0 & \dots & 0 \end{bmatrix} = \begin{bmatrix} M\hat{K}_D^{-1}M & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}$$

$$= \Gamma_2 \mathcal{M}_n^{-1}\Gamma_2^T$$

where,

$$\Gamma_2 = \begin{bmatrix} \frac{1}{\sqrt{2}}M\hat{K}_D^{-\frac{1}{2}}M^{\frac{1}{2}} & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}.$$

Thus, the Schur complement approximation given in (4.46) can be written as

$$\tilde{S}_1 = \frac{1}{\tau\beta_{noise}}\mathcal{K}\mathcal{M}_n^{-1}\mathcal{K}^T + \frac{1}{\beta_{prior}}\Gamma_2\mathcal{M}_n^{-1}\Gamma_2. \tag{4.47}$$

A suitable approximation for the Schur complement approximation given in (4.47) is chosen as

$$\hat{S}_1 = \frac{1}{\tau\beta_{noise}}(\mathcal{K} + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\Gamma_2)\mathcal{M}_n^{-1}(\mathcal{K} + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\Gamma_2)^T. \tag{4.48}$$

4.5.2 2nd order smoothness prior with Neumann boundary condition

For 2nd order smoothness prior with Neumann boundary condition the saddle point system to solve is given in (3.33) with

$$A = \begin{bmatrix} \tau\beta_{noise}\mathcal{M}_n & 0 \\ 0 & \beta_{prior}K_N \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} -\mathcal{K} & \mathcal{M}_1 \end{bmatrix}.$$

The $(1,1)$ block of the system, A, is not invertible, since K_N is not invertible. Due to the non-invertibility of K_N as well as A, the Schur complement of the system (3.33) does not exist. So, to prescribe an approximation for a preconditioner, we recommend an approximation $\hat{K}_N$ to K_N given by $\hat{K}_N = diag(K_N)$ [45]. Which gives us the approximation to A as follows:

$$\hat{A} = \begin{bmatrix} \tau\beta_{noise}\mathcal{M}_n & 0 \\ 0 & \beta_{prior}\hat{K}_N \end{bmatrix}.$$

Hence we examine the approximated saddle point system $\begin{bmatrix} \hat{A} & B^T \\ B & 0 \end{bmatrix}$. The (negative) Schur complement of this system is

$$S_2 = \frac{1}{\tau\beta_{noise}}\mathcal{K}\mathcal{M}_n^{-1}\mathcal{K}^T + \frac{1}{\beta_{prior}}\mathcal{M}_1\hat{K}_N^{-1}\mathcal{M}_1^T. \tag{4.49}$$

The 2nd term of (4.49) $\mathcal{M}_1\hat{K}_N^{-1}\mathcal{M}_1^T$ gives us

$$\begin{bmatrix} M \\ 0 \\ \vdots \\ \\ 0 \end{bmatrix} \hat{K}_N^{-1} \begin{bmatrix} M & 0 & \dots & 0 \end{bmatrix} = \begin{bmatrix} M\hat{K}_N^{-1}M & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}$$

$$= \Gamma_3\mathcal{M}_n^{-1}\Gamma_3^T$$

where,

$$\Gamma_3 = \begin{bmatrix} \frac{1}{\sqrt{2}}M\hat{K}_N^{-\frac{1}{2}}M^{\frac{1}{2}} & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}.$$

The Schur complement given in (4.49) can be written as

$$\tilde{S}_2 = \frac{1}{\tau\beta_{noise}}\mathcal{K}\mathcal{M}_n^{-1}\mathcal{K}^T + \frac{1}{\beta_{prior}}\Gamma_3\mathcal{M}_n^{-1}\Gamma_3. \tag{4.50}$$

As in the case of 2nd order smoothness prior with Dirichlet boundary condition an approximation for the Schur complement approximation given in (4.50) is proposed as

$$\hat{S}_2 = \frac{1}{\tau\beta_{noise}}(\mathcal{K} + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\Gamma_3)\mathcal{M}_n^{-1}(\mathcal{K} + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\Gamma_3)^T. \tag{4.51}$$

4.5.3 4th order smoothness prior with Dirichlet boundary condition

In the case of 4th order smoothness prior with Dirichlet boundary condition the system to solve is given in (3.34) with

$$A = \begin{bmatrix} \tau\beta_{noise}\mathcal{M}_n & 0 \\ 0 & \beta_{prior}K_D^2 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} -\mathcal{K} & \mathcal{M}_1 \end{bmatrix}.$$

The (negative) Schur complement of the system is

$$S_3 = \frac{1}{\tau\beta_{noise}}\mathcal{K}\mathcal{M}_n^{-1}\mathcal{K}^T + \frac{1}{\beta_{prior}}\mathcal{M}_1 K_D^{-2}\mathcal{M}_1^T. \tag{4.52}$$

As in the case of 2nd order smoothness prior with Dirichlet boundary condition the 2nd term $\mathcal{M}_1 K_D^{-2}\mathcal{M}_1^T$ in (4.52) the inverse of K_D can not be found easily. So in the same way we approximate K_D by $\hat{K}_D = diag(K_D)$ [45]. Then the Schur complement approximation is given by

$$\tilde{S}_3 = \frac{1}{\tau\beta_{noise}}\mathcal{K}\mathcal{M}_n^{-1}\mathcal{K}^T + \frac{1}{\beta_{prior}}\mathcal{M}_1 \hat{K}_D^{-2}\mathcal{M}_1^T \tag{4.53}$$

The 2nd term of (4.53) $\mathcal{M}_1 \hat{K}_D^{-2}\mathcal{M}_1^T$ gives us

$$\begin{bmatrix} M \\ 0 \\ \vdots \\ \\ 0 \end{bmatrix} \hat{K}_D^{-2} \begin{bmatrix} M & 0 & \dots & 0 \end{bmatrix} = \begin{bmatrix} M\hat{K}_D^{-2}M & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}$$

$$= \Gamma_4 \mathcal{M}_n^{-1} \Gamma_4^T$$

where,

$$\Gamma_4 = \begin{bmatrix} \frac{1}{\sqrt{2}} M\hat{K}_D^{-1}M^{\frac{1}{2}} & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}.$$

Thus, the Schur complement given in (4.53) can be written as

$$\tilde{S}_3 = \frac{1}{\tau\beta_{noise}}\mathcal{K}\mathcal{M}_n^{-1}\mathcal{K}^T + \frac{1}{\beta_{prior}}\Gamma_4 \mathcal{M}_n^{-1}\Gamma_4. \tag{4.54}$$

Similarly an approximation for the Schur complement approximation given in (4.54) is proposed as

$$\hat{S}_3 = \frac{1}{\tau\beta_{noise}}(\mathcal{K} + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\Gamma_4)\mathcal{M}_n^{-1}(\mathcal{K} + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\Gamma_4)^T. \tag{4.55}$$

4.5.4 4th order smoothness prior with Neumann boundary condition

In the case of 4th order smoothness prior with Neumann boundary condition the system to solve is given in (3.35) with

$$A = \begin{bmatrix} \tau\beta_{noise}\mathcal{M}_n & 0 \\ 0 & \beta_{prior}K_N^2 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} -\mathcal{K} & \mathcal{M}_1 \end{bmatrix}.$$

The $(1,1)$ block of the system, A, is not invertible, since K_N is not invertible. Due to the non-invertibility of K_N as well as A, the Schur complement of the system (3.35) does not exist. So, to prescribe an approximation for a preconditioner, as in the case of 2nd order smoothness prior with Neumann boundary condition we recommend an approximation $\hat{K}_N$ to K_N given by $\hat{K}_N = diag(K_N)$ [45]. Which gives us the approximation to A as follows:

$$\hat{A} = \begin{bmatrix} \tau\beta_{noise}\mathcal{M}_n & 0 \\ 0 & \beta_{prior}\hat{K}_N^2 \end{bmatrix}.$$

Hence, we examine the approximated saddle point system $\begin{bmatrix} \hat{A} & B^T \\ B & 0 \end{bmatrix}$. The (negative) Schur complement of this system is

$$S_4 = \frac{1}{\tau\beta_{noise}}\mathcal{K}\mathcal{M}_n^{-1}\mathcal{K}^T + \frac{1}{\beta_{prior}}\mathcal{M}_1\hat{K}_N^{-2}\mathcal{M}_1^T. \tag{4.56}$$

The 2nd part of (4.56) $\mathcal{M}_1\hat{K}_N^{-2}\mathcal{M}_1^T$ can be written as

$$\begin{bmatrix} M \\ 0 \\ \vdots \\ \\ 0 \end{bmatrix} \hat{K}_N^{-2} \begin{bmatrix} M & 0 & \dots & 0 \end{bmatrix} = \begin{bmatrix} M\hat{K}_N^{-2}M & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}$$

$$= \Gamma_5 \mathcal{M}_n^{-1} \Gamma_5^T$$

where,

$$\Gamma_5 = \begin{bmatrix} \frac{1}{\sqrt{2}} M\hat{K}_N^{-1}M^{\frac{1}{2}} & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}.$$

The Schur complement approximation given in (4.56) can be written as

$$\tilde{S}_4 = \frac{1}{\tau\beta_{noise}} \mathcal{K}\mathcal{M}_n^{-1}\mathcal{K}^T + \frac{1}{\beta_{prior}} \Gamma_5 \mathcal{M}_n^{-1} \Gamma_5. \tag{4.57}$$

In a similar manner we propose a suitable approximation for the Schur complement approximation given in (4.57) as

$$\hat{S}_4 = \frac{1}{\tau\beta_{noise}} (\mathcal{K} + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\Gamma_5) \mathcal{M}_n^{-1} (\mathcal{K} + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\Gamma_5)^T. \tag{4.58}$$

As in the case of prior matrix is mass matrix it will not be feasible if we apply the inverse of $(\mathcal{K} + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\Gamma_i)$ and its transpose $(\mathcal{K} + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\Gamma_i)^T$, $i = 2, 3, 4, 5$. We will use algebraic multigrid technique for $(\mathcal{K} + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\Gamma_i)$ and its transpose $(\mathcal{K} + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\Gamma_i)^T$. That is we require an algebraic multigrid process for the 1st diagonal block $M + \tau K + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}} M\hat{K}_D^{-\frac{1}{2}} M^{\frac{1}{2}} \in \mathbb{R}^{n\times n}$ for the case of 2nd order smoothness prior with Dirichlet boundary condition, $M + \tau K + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}} M\hat{K}_N^{-\frac{1}{2}} M^{\frac{1}{2}} \in \mathbb{R}^{n\times n}$ for the case of 2nd order smoothness prior with Neumann boundary condition, $M + \tau K + \frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}} M\hat{K}_D^{-1} M^{\frac{1}{2}} \in$

$\mathbb{R}^{n\times n}$ for for the case of 4th order smoothness prior with Dirichlet boundary condition and $M+\tau K+\frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}M\hat{K}_N^{-1}M^{\frac{1}{2}} \in \mathbb{R}^{n\times n}$ in case of 4th order smoothness prior with Neumann boundary condition and algebraic multigrid process for the rest of the block $M+\tau K \in \mathbb{R}^{n\times n}$. We apply a fixed (i.e. 4) number of v-cycles of an algebraic multigrid process N_t times to approximate the inverse of $(\mathcal{K}+\frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\Gamma_i)$ and N_t times to approximate the inverse of $(\mathcal{K}+\frac{\sqrt{\tau\beta_{noise}}}{\sqrt{\beta_{prior}}}\Gamma_i)^T$.

Chapter 5

Numerical Results and Discussions

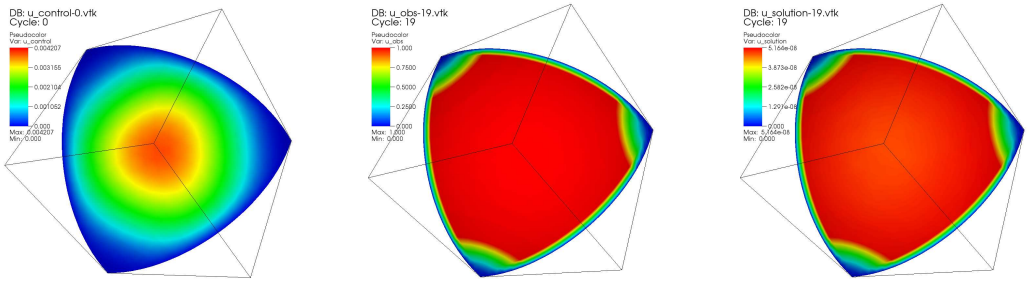
The results presented in this section are based on an implementation of our algorithms within the deal.II [46] framework. For the AMG precondtioner, we use the Trillions ML packages [47] that implements a smoothed aggregation AMG. Our implementation of MINRES was taken from [25] with a tolerance of 10^{-6} for the relative pseudo residual. For all our experiments $T = 1$ and $\tau = 0.05$ which results in 20 time steps. We discretized using $Q1$ finite elements for each of the state, control and Lagrange multiplier fields as well as the backward Euler scheme. Also note that whenever we show degrees of freedom these are only the degrees of freedom for one grid point of time (i. e. for a single time step). In fact in that case we are solving a linear system of dimension two times the number of time step times the number of degrees of freedom plus one times the degrees of freedom of the spatial discretization. For example, a spatial discretization with 274625 unknowns and 20 time steps corresponds to an overall linear system of dimension 11259625. We vary the smoothing parameters β_{noise} and β_{prior} choosing $\beta_{noise} = 10^0, 10^{-2}, 10^{-4}$ and 10^{-6} and $\beta_{prior} = 10^0, 10^{-2}, 10^{-4}$ and 10^{-6}. To exemplify our approach, in table 5.1-5.3 we show degrees of freedom (DoF), number of MINRES iterations along with CPU times (in seconds) for the system 3.30 which is with prior matrix as the mass matrix with Schur complement given in (4.42). Table 5.4-5.6 show degrees of freedom (DoF), number of MINRES iterations along with CPU times (in seconds) for the system 3.32, for the 2nd order smoothness prior with Dirichlet boundary condition with Schur complement given in 4.48 and in Table 5.7-5.9 we depict degrees of freedom (DoF), number of MINRES along with cpu times (in seconds) for the system 3.33, for the 2nd order smoothness prior with Neumann boundary condition with Schur complement given in 4.51. On the other hand, in Table 5.10-5.12 we depict degrees of freedom (DoF), number of MINRES along

with CPU times (in seconds) for the system 3.34, for the 4th order smoothness prior with Dirichlet boundary condition with Schur complement given in 4.55. And finally in Table 5.13-5.15 we depict degrees of freedom (DoF), number of MINRES along with CPU times (in seconds) for the system 3.35, for the 4th order smoothness prior with Neumann boundary condition with Schur complement given in 4.58.

Figure 5.1 shows the computed control, observed solutions and computed solutions at $t = 1$ for $\beta_{noise} = 10^0$ and $\beta_{prior} = 10^0$ with degrees of freedom 4913 for the prior matrix chosen to be the mass matrix. While the Figure 5.2 shows the computed control, the observed solutions and computed solutions at $t = 1$ for $\beta_{noise} = 10^{-2}$ and $\beta_{prior} = 10^0$ with degrees of freedom 35937 for 2nd order smoothness prior with Dirichlet boundary condition. On the other hand the Figure 5.3 shows the computed control, the observed solutions and computed solutions at $t = 1$ for $\beta_{noise} = 10^{-4}$ and $\beta_{prior} = 10^0$ with degrees of freedom 4913 for 2nd order smoothness prior with Neumann boundary condition. Figure 5.4 depicts the computed control, the observed solutions and computed solutions at $t = 1$ for $\beta_{noise} = 10^{-2}$ and $\beta_{prior} = 10^{-2}$ with degrees of freedom 35937 for the 4th order smoothness prior with Dirichlet boundary condition. Finally the Figure 5.5 depicts the computed control, the observed solutions and computed solutions at $t = 1$ for $\beta_{noise} = 10^{-4}$ and $\beta_{prior} = 10^{-4}$ with degrees of freedom 35937 for the 4th order smoothness prior with Neumann boundary condition.

It can be easily seen from the tables that the number of iterations remains almost constant with varying mesh size and the regularization parameter β_{noise} and β_{prior}. In case of 4th order smoothness prior with Dirichlet boundary condition the method converge if the ratio $\frac{\beta_{noise}}{\beta_{prior}} \leq 10^2$. If the ratio $\frac{\beta_{noise}}{\beta_{prior}} > 10^2$, then the number of iterations of MINRES method is very high and it takes more time. For the case of 4th order smoothness prior with Neumann boundary condition the method converge if the ratio $\frac{\beta_{noise}}{\beta_{prior}} \leq 1$. If the ratio $\frac{\beta_{noise}}{\beta_{prior}} > 1$, then the number of iterations of MINRES method increases substantially and it takes more time.

5.0.5 Prior matrix as mass matrix



(A) Computed control (B) Observed state (C) Computed solution

FIGURE 5.1: Plots of Control, Observed state, and Computed state for $\beta_{noise} = 10^0$ and $\beta_{prior} = 10^0$ at $t = 1$ with DOF 4913 for prior matrix as mass matrix.

	$\beta_{prior} = 10^0$	$\beta_{prior} = 10^{-2}$	$\beta_{prior} = 10^{-4}$	$\beta_{prior} = 10^{-6}$
DoF	MINRES(Time)	MINRES(Time)	MINRES(Time)	MINRES(Time)
4913	6(10)	8(14)	12(19)	12(19)
35937	6(90)	8(117)	12(162)	14(187)
274625	6(731)	8(897)	12(1259)	14(1446)

TABLE 5.1: Number of DoF, MINRES steps and CPU-time (in seconds) for different values of for regularization parameter $\beta_{noise} = 10^{-2}$ with $\beta_{prior} = 10^0, 10^{-2}, 10^{-4}$ and 10^{-6} for the prior matrix as the mass matrix.

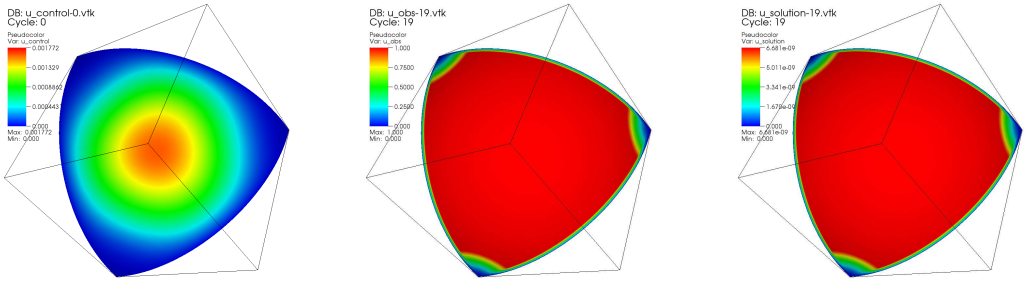
	$\beta_{prior} = 10^0$	$\beta_{prior} = 10^{-2}$	$\beta_{prior} = 10^{-4}$	$\beta_{prior} = 10^{-6}$
DoF	MINRES(Time)	MINRES(Time)	MINRES(Time)	MINRES(Time)
4913	4(7)	6(11)	8(13)	12(20)
35937	4(69)	6(93)	8(116)	12(159)
274625	4(555)	6(732)	8(927)	12(1263)

TABLE 5.2: Number of DoF, MINRES steps and CPU-time (in seconds) for different values of for regularization parameter $\beta_{noise} = 10^{-4}$ with $\beta_{prior} = 10^0, 10^{-2}, 10^{-4}$ and 10^{-6} for the prior matrix as the mass matrix.

	$\beta_{prior} = 10^0$	$\beta_{prior} = 10^{-2}$	$\beta_{prior} = 10^{-4}$	$\beta_{prior} = 10^{-6}$
DoF	MINRES(Time)	MINRES(Time)	MINRES(Time)	MINRES(Time)
4913	4(8)	4(8)	6(11)	8(14)
35937	4(71)	4(70)	6(95)	8(118)
274625	4(556)	4(554)	6(773)	8(907)

TABLE 5.3: Number of DoF, MINRES steps and CPU-time (in seconds) for different values of for regularization parameter $\beta_{noise} = 10^{-6}$ with $\beta_{prior} = 10^0, 10^{-2}, 10^{-4}$ and 10^{-6} for the prior matrix as the mass matrix.

5.0.6 2nd order smoothness prior with Dirichlet boundary condition



(A) Computed control (B) Observed state (C) Computed solution

FIGURE 5.2: Plots of Control, Observed state, and Computed state for $\beta_{noise} = 10^{-2}$ and $\beta_{prior} = 10^{0}$ at $t = 1$ with DOF 35937 for 2nd order smoothness prior with Dirichlet boundary condition.

	$\beta_{prior} = 10^{0}$	$\beta_{prior} = 10^{-2}$	$\beta_{prior} = 10^{-4}$	$\beta_{prior} = 10^{-6}$
DoF	MINRES(Time)	MINRES(Time)	MINRES(Time)	MINRES(Time)
4913	4(7)	4(9)	4(9)	8(8)
35937	4(69)	4(69)	6(93)	9(127)
274625	4(539)	4(546)	7(814)	9(1530)

TABLE 5.4: Number of DoF, MINRES steps and CPU-time (in seconds) for different values of for regularization parameter $\beta_{noise} = 10^{-2}$ with $\beta_{prior} = 10^{0}, 10^{-2}, 10^{-4}$ and 10^{-6} rrespectively for 2nd order smoothness prior with Dirichlet boundary condition.

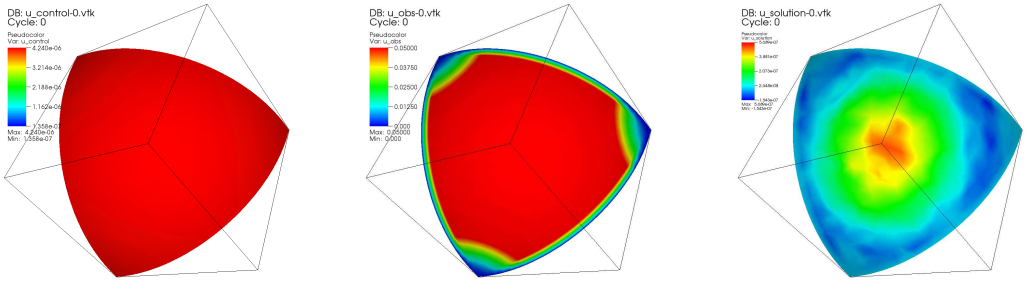
	$\beta_{prior} = 10^{0}$	$\beta_{prior} = 10^{-2}$	$\beta_{prior} = 10^{-4}$	$\beta_{prior} = 10^{-6}$
DoF	MINRES(Time)	MINRES(Time)	MINRES(Time)	MINRES(Time)
4913	4(9)	4(8)	4(8)	4(8)
35937	4(69)	4(75)	4(69)	6(93)
274625	4(534)	4(552)	4(539)	7(823)

TABLE 5.5: Number of DoF, MINRES steps and CPU-time (in seconds) for different values of for regularization parameter $\beta_{noise} = 10^{-4}$ with $\beta_{prior} = 10^{0}, 10^{-2}, 10^{-4}$ and 10^{-6} rrespectively for 2nd order smoothness prior with Dirichlet boundary condition.

	$\beta_{prior} = 10^{0}$	$\beta_{prior} = 10^{-2}$	$\beta_{prior} = 10^{-4}$	$\beta_{prior} = 10^{-6}$
DoF	MINRES(Time)	MINRES(Time)	MINRES(Time)	MINRES(Time)
4913	3(7)	4(8)	4(8)	4(9)
35937	3(57)	4(69)	4(69)	4(70)
274625	3(455)	4(548)	4(551)	4(541)

TABLE 5.6: Number of DoF, MINRES steps and CPU-time (in seconds) for different values of for regularization parameter $\beta_{noise} = 10^{-6}$ with $\beta_{prior} = 10^{0}, 10^{-2}, 10^{-4}$ and 10^{-6} respectively for 2nd order smoothness prior with Dirichlet boundary condition.

5.0.7 2nd order smoothness prior with Neumann boundary condition



(A) Computed control (B) Observed state (C) Computed solution

FIGURE 5.3: Plots of Control, Observed state, and Computed state for $\beta_{noise} = 10^{-4}$ and $\beta_{prior} = 10^{-0}$ at $t = 1$ with DOF 4913 for 2nd order smoothness prior with Neumann boundary condition.

	$\beta_{prior} = 10^{0}$	$\beta_{prior} = 10^{-2}$	$\beta_{prior} = 10^{-4}$	$\beta_{prior} = 10^{-6}$
DoF	MINRES(Time)	MINRES(Time)	MINRES(Time)	MINRES(Time)
4913	4(8)	5(10)	7(12)	12(19)
35937	5(88)	7(108)	7(109)	12(168)
274625	5(559)	7(1002)	7(955)	13(1464)

TABLE 5.7: Number of DoF, MINRES steps and CPU-time (in seconds) for different values of for regularization parameter $\beta_{noise} = 10^{-2}$ with $\beta_{prior} = 10^{0}, 10^{-2}, 10^{-4}$ and 10^{-6} respectively for 2nd order smoothness prior with Neumann boundary condition.

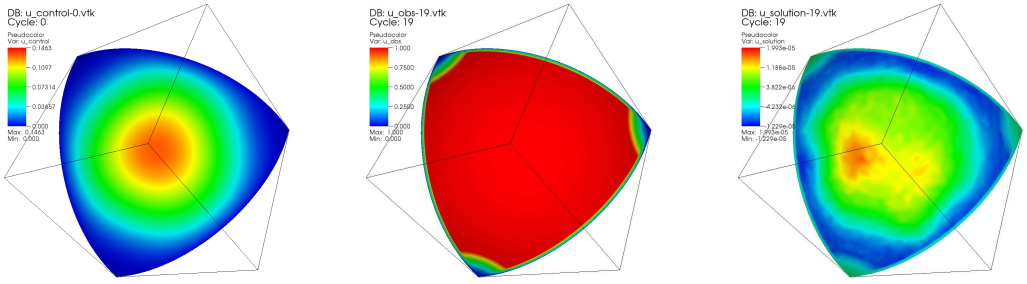
	$\beta_{prior} = 10^{0}$	$\beta_{prior} = 10^{-2}$	$\beta_{prior} = 10^{-4}$	$\beta_{prior} = 10^{-6}$
DoF	MINRES(Time)	MINRES(Time)	MINRES(Time)	MINRES(Time)
4913	4(8)	4(8)	5(9)	7(12)
35937	4(72)	5(82)	7(108)	7(108)
274625	4(557)	5(651)	7(848)	7(878)

TABLE 5.8: Number of DoF, MINRES steps and CPU-time (in seconds) for different values of for regularization parameter $\beta_{noise} = 10^{-4}$ with $\beta_{prior} = 10^{0}, 10^{-2}, 10^{-4}$ and 10^{-6} respectively for 2nd order smoothness prior with Neumann boundary condition.

	$\beta_{prior} = 10^{0}$	$\beta_{prior} = 10^{-2}$	$\beta_{prior} = 10^{-4}$	$\beta_{prior} = 10^{-6}$
DoF	MINRES(Time)	MINRES(Time)	MINRES(Time)	MINRES(Time)
4913	4(9)	4(8)	4(8)	5(9)
35937	4(72)	4(72)	5(85)	7(109)
274625	4(590)	4(568)	6(663)	7(833)

TABLE 5.9: Number of DoF, MINRES steps and CPU-time (in seconds) for different values of for regularization parameter $\beta_{noise} = 10^{-6}$ with $\beta_{prior} = 10^{0}, 10^{-2}, 10^{-4}$ and 10^{-6} respectively for 2nd order smoothness prior with Neumann boundary condition.

5.0.8 4th order smoothness prior with Dirichlet boundary condition



(A) Computed control (B) Observed state (C) Computed solution

FIGURE 5.4: Plots of Control, Observed state, and Computed state for $\beta_{noise} = 10^{-2}$ and $\beta_{prior} = 10^{-2}$ at $t = 1$ with DOF 35937 for 4th order smoothness prior with Dirichlet boundary condition.

	$\beta_{prior} = 10^{0}$	$\beta_{prior} = 10^{-2}$	$\beta_{prior} = 10^{-4}$
DoF	MINRES(Time)	MINRES(Time)	MINRES(Time)
4913	4(9)	4(9)	6(11)
35937	4(71)	7(111)	11(159)
274625	7(851)	9(1042)	18(1884)

TABLE 5.10: Number of DoF, MINRES steps and CPU-time (in seconds) for different values of for regularization parameter $\beta_{noise} = 10^{-2}$ with $\beta_{prior} = 10^{0}, 10^{-2}, 10^{-4}$ and 10^{-6} respectively for 4th order smoothness prior with Dirichlet boundary condition.

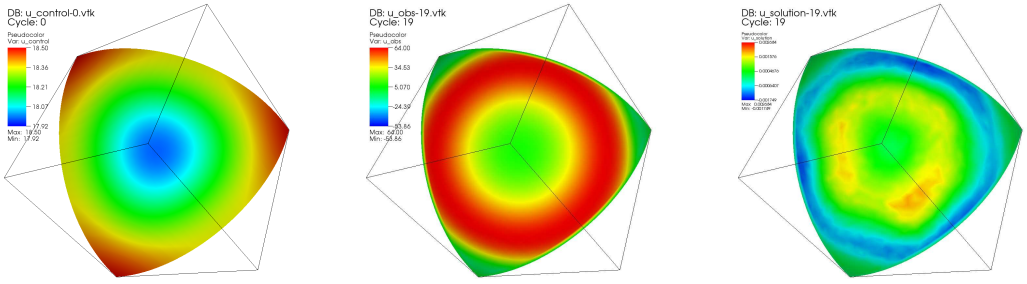
	$\beta_{prior} = 10^{0}$	$\beta_{prior} = 10^{-2}$	$\beta_{prior} = 10^{-4}$	$\beta_{prior} = 10^{-6}$
DoF	MINRES(Time)	MINRES(Time)	MINRES(Time)	MINRES(Time)
4913	4(8)	4(9)	4(8)	6(11)
35937	4(75)	4(73)	7(111)	11(157)
274625	4(569)	7(858)	9(1049)	18(1921)

TABLE 5.11: Number of DoF, MINRES steps and CPU-time (in seconds) for different values of for regularization parameter $\beta_{noise} = 10^{-4}$ with $\beta_{prior} = 10^{0}, 10^{-2}, 10^{-4}$ and 10^{-6} respectively for 4th order smoothness prior with Dirichlet boundary condition.

	$\beta_{prior} = 10^{0}$	$\beta_{prior} = 10^{-2}$	$\beta_{prior} = 10^{-4}$	$\beta_{prior} = 10^{-6}$
DoF	MINRES(Time)	MINRES(Time)	MINRES(Time)	MINRES(Time)
4913	4(9)	4(9)	4(8)	4(9)
35937	4(70)	4(75)	4(74)	7(110)
274625	4(540)	4(571)	7(857)	9(1053)

TABLE 5.12: Number of DoF, MINRES steps and CPU-time (in seconds) for different values of for regularization parameter $\beta_{noise} = 10^{-6}$ with $\beta_{prior} = 10^{0}, 10^{-2}, 10^{-4}$ and 10^{-6} respectively for 4th order smoothness prior with Dirichlet boundary condition.

5.0.9 4th order smoothness prior with Neumann boundary condition



(A) Computed control (B) Observed state (C) Computed solution

FIGURE 5.5: Plots of Control, Observed state, and Computed state for $\beta_{noise} = 10^{-4}$ and $\beta_{prior} = 10^{-4}$ at $t = 1$ with DOF 35937 for 4th order smoothness prior with Neumann boundary condition.

	$\beta_{prior} = 10^{0}$	$\beta_{prior} = 10^{-2}$
DoF	MINRES(Time)	MINRES(Time)
4913	12(20)	17(24)
35937	12(167)	17(228)
274625	10(1148)	23(2381)

TABLE 5.13: Number of DoF, MINRES steps and CPU-time (in seconds) for different values of for regularization parameter $\beta_{noise} = 10^{-2}$ with $\beta_{prior} = 10^{0}, 10^{-2}$ and 10^{-4} respectively for 4th order smoothness prior with Neumann boundary condition.

	$\beta_{prior} = 10^{0}$	$\beta_{prior} = 10^{-2}$	$\beta_{prior} = 10^{-4}$
DoF	MINRES(Time)	MINRES(Time)	MINRES(Time)
4913	10(17)	12(20)	36(54)
35937	10(145)	12(168)	17(227)
274625	11(1246)	10(1151)	23(2367)

TABLE 5.14: Number of DoF, MINRES steps and CPU-time (in seconds) for different values of for regularization parameter $\beta_{noise} = 10^{-4}$ with $\beta_{prior} = 10^{0}, 10^{-2}$ and 10^{-4} respectively for 4th order smoothness prior with Neumann boundary condition.

	$\beta_{prior} = 10^{0}$	$\beta_{prior} = 10^{-2}$	$\beta_{prior} = 10^{-4}$	$\beta_{prior} = 10^{-6}$
DoF	MINRES(Time)	MINRES(Time)	MINRES(Time)	MINRES(Time)
4913	10(18)	10(17)	12(20)	18(28)
35937	10(143)	10(145)	12(225)	27(347)
274625	11(1222)	11(1236)	10(1228)	23(2831)

TABLE 5.15: Number of DoF, MINRES steps and CPU-time (in seconds) for different values of for regularization parameter $\beta_{noise} = 10^{-6}$ with $\beta_{prior} = 10^{0}, 10^{-2}, 10^{-4}$ and 10^{-6} respectively for 4th order smoothness prior with Neumann boundary condition.

Chapter 6

Conclusion

In this thesis we have presented a setup for the inverse problem governed by a three dimensional parabolic PDE within the framework of Bayesian inference with Gaussian noise and prior probability densities. We have derived the discretized first order optimality conditions using the finite element method that leads us to a very large linear system in the form of a saddle point system. We studied Gaussian smoothness priors. Namely, 2nd order Gaussian smoothness prior with Dirichlet boundary conditions and with Neumann Boundary conditions. Additionally a 4th order Gaussian smoothness prior with Dirichlet boundary conditions and Neumann boundary conditions. We proposed an all-at-once approach to solve the linear system. The system matrix in all cases was symmetric indefinite, so the Krylov subspace method MINRES is used with the application of a block diagonal preconditioner $\mathcal{P}$. the preconditioner P was chosen to enhance the convergence behavior that allowed us to have a fast solution of the problems. We also proposed approximations of the Schur complements for the preconditioners for all cases. For the case of spectrally neutral prior the eigenvalues of $(\hat{S}^{-1}S)$ are found in $[\frac{1}{2}, 1)$, where $\hat{S}$ is the approximation of Schur complement S. This shows that the Schur complement approximation for the case of spectrally neutral prior is robust for the original system. For the cases of smoothness prior though eigenvalue bounds as in the case of spectrally neutral prior could not be guaranteed, the approximations of the Schur complement approximations perform robustly with respect to all the parameters. Our preconditioners did not need to solve the time evaluation accurately (e. g. using a direct method) but only used an AMG preconditioner. This approach not only speed up the solution of the system but also allowed for much larger problems to be solved as the AMG preconditioner can easily handle extremely large three dimensional matrices.

Bibliography

[1] A. Kirsch. *An Introduction to the Mathematical Theory of Inverse Problems.* Applied Mathematical Sciences. Springer New York, 2011. ISBN 9781441984746.

[2] A. Tarantola. *Inverse problem theory and methods for model parameter estimation.* Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 2005. ISBN 9780898717921.

[3] Jacques Hadamard. Sur les problèmes aux dérivés partielles et leur signification physique. *Princeton University Bulletin*, 13:49–52, 1902.

[4] S. L. Cotter, M. Dashti, and A. M. Stuart. Approximation of Bayesian Inverse Problems for PDEs. *ArXiv e-prints*, September 2009.

[5] H.W. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems.* Mathematics and Its Applications. Springer, 1996. ISBN 9780792341574.

[6] Jingbo Wang. *Bayesian computational techniques for inverse problems in transport processes.* PhD thesis, Cornell University, 2006.

[7] Jairi Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems.* Springer, 2004.

[8] H. Flath, L. Wilcox, V. Akçelik, J. Hill, B. van Bloemen Waanders, and O. Ghattas. Fast algorithms for Bayesian uncertainty quantification in large-scale linear inverse problems based on low-rank partial Hessian approximations. *SIAM Journal on Scientific Computing*, 33(1):407–432, 2011. doi: 10.1137/090780717.

[9] Y. Saad. *Iterative methods for sparse linear systems.* Society for Industrial and Applied Mathematics, Philadelphia, PA, 2003. ISBN 0-89871-534-2.

[10] Lloyd N Trefethen and David Bau III. *Numerical linear algebra.* Number 50. Society for Industrial Mathematics, Philadelphia, PA, 1997. ISBN 978-0-898713-61-9.

[11] Carl Meyer. *Matrix analysis and applied linear algebra book and solutions manual.* Number 71. Society for Industrial Mathematics, 2000.

[12] Daniel Kressner. *Numerical methods for general and structured eigenvalue problems.* Berlin: Springer, 2005. doi: 10.1007/3-540-28502-4.

[13] Saad Yousef. *Numerical Methods for Large Eigenvalue Problems.* Society for Industrial and Applied Mathematics, Philadelphia, PA, 2011. ISBN 978-1-61197-073-9. doi: 10.1137/1.9781611970739.

[14] Michele Benzi, Gene H. Golub, and Jörg Liesen. Numerical solution of saddle point problems. *Acta Numerica*, 14:1–137, 4 2005. ISSN 1474-0508.

[15] Fuzhen Zhang. *The Schur complement and its applications.* Springer, NY, 2005. ISBN 0-387-24271-6.

[16] Emilie V Haynsworth. On the Schur complement. Technical report, DTIC Document, 1968.

[17] Eldad Haber and Uri M Ascher. Preconditioned all-at-once methods for large, sparse parameter estimation problems. *Inverse Problems*, 17(6):1847, 2001.

[18] Roland Glowinski and Patrick Le Tallec. *Augmented Lagrangian and operator-splitting methods in nonlinear mechanics.* Society for Industrial and Applied Mathematics, Philadelphia, PA, 1989.

[19] Michel Fortin and Roland Glowinski. *Augmented Lagrangian methods: applications to the numerical solution of boundary-value problems*, volume 15. North Holland, Amsterdam-New York-Oxford, 2000. ISBN 978-0-444-86452-9.

[20] Chen Greif, Gene H. Golub, and James M. Varah. Augmented Lagrangian techniques for solving saddle point linear systems. Technical report, SIAM J. Matrix Anal. Appl, 2004.

[21] Roger A Horn and Charles R Johnson. *Matrix analysis.* Cambridge university press, New York, 1990. ISBN 9780521386326.

[22] Torgeir Rusten and Ragnar Winther. A preconditioned iterative method for saddle-point problems. *SIAM Journal on Matrix Analysis and Applications*, 13(3):887–904, 1992.

[23] Owe Axelsson. *Iterative solution methods.* Cambridge University Press, 1996. ISBN 9780521555692.

[24] Henk A. van der Vorst. *Iterative Krylov methods for large linear systems*. Cambridge University Press, Cambridge, 2009. doi: 10.1017/CBO9780511615115.

[25] Howard C. Elman, David J. Silvester, and Andrew J. Wathen. *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2005. ISBN 978-0-19-852868-5; 0-19-852868-X.

[26] Daniela Calvetti. *Introduction to Bayesian Scientific Computing Ten Lectures on Subjective Computing*. Springer Science+Business Media LLC, NY, 2007.

[27] Fredi Tröltzsch. *Optimal control of partial differential equations: theory, methods, and applications*, volume 112. American Mathematical Soc., 2010.

[28] John W. Pearson and Martin Stoll. Fast Iterative Solution of Reaction-Diffusion Control Problems Arising from Chemical Processes. *Submitted*, 2012.

[29] John T Betts and Stephen L Campbell. *Discretize then optimize*. SIAM Publications, Philadelphia, 2005.

[30] Gilbert Strang and George Fix. *An Analysis of the Finite Element Method 2nd Edition*. Wellesley-Cambridge, 2nd edition, May 2008. ISBN 0980232708.

[31] Dietrich Braess. *Finite elements*. Cambridge University Press, 2007.

[32] Martin Stoll. All-at-once solution of a time-dependent time-periodic PDE-constrained optimization problems. *Accepted, IMA Journal of Numerical Analysis*, July 2011.

[33] Jorge Nocedal and Stephen J Wright. *Numerical optimization*. Springer Science+ Business Media, 2006.

[34] Iain S. Duff. Sparse numerical linear algebra: Direct methods and preconditioning. Technical report, 1996.

[35] T. Rees, H. Dollar, and A. Wathen. Optimal solvers for pde-constrained optimization. *SIAM Journal on Scientific Computing*, 32(1):271–298, 2010. doi: 10.1137/080727154.

[36] C. Paige and M. Saunders. Solution of sparse indefinite systems of linear equations. *SIAM Journal on Numerical Analysis*, 12(4):617–629, 1975. doi: 10.1137/0712047.

[37] Walter Edwin Arnoldi. The principle of minimized iterations in the solution of the matrix eigenvalue problem. *Quart. Appl. Math*, 9(1):17–29, 1951.

[38] Magnus Rudolph Hestenes and Eduard Stiefel. Methods of conjugate gradients for solving linear systems. *Journal of Research of the National Bureau of Standards*.

[39] Youcef Saad and Martin H. Schultz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput*, 7(3): 856–869, 1986. ISSN 0196-5204.

[40] Anne Greenbaum. *Iterative methods for solving linear systems*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 1997. ISBN 0-89871-396-X.

[41] Malcolm F Murphy, Gene H Golub, and Andrew J Wathen. A note on preconditioning for indefinite linear systems. *SIAM Journal on Scientific Computing*, 21(6): 1969–1972, 2000.

[42] John W. Pearson, Martin Stoll, and Andy Wathen. Preconditioners for state constrained optimal control problems with Moreau-Yosida penalty function. *To Appear Numerical Linear Algebra with Applications*, 2012.

[43] John W Pearson, Martin Stoll, and Andrew J Wathen. Regularization-robust preconditioners for time-dependent pde-constrained optimization problems. *SIAM Journal on Matrix Analysis and Applications*, 33(4):1126–1152, 2012.

[44] JW Ruge and Klaus Stüben. Algebraic multigrid. *Multigrid methods*, 3:73–130, 1987.

[45] Tyrone Rees and Martin Stoll. A fast solver for an H1 regularized optimal control problem. *Submitted*, 2012.

[46] W. Bangerth, R. Hartmann, and G. Kanschat. deal.II-A general-purpose object-oriented finite element library. *ACM Trans. Math. Softw.*, 33(4), August 2007. ISSN 0098-3500.

[47] M.W. Gee, C.M. Siefert, J.J. Hu, R.S. Tuminaro, and M.G. Sala. ML 5.0 smoothed aggregation user's guide. Technical Report SAND2006-2649, Sandia National Laboratories, 2006.