Copyright Statement

The digital copy of this thesis is protected by the Copyright Act 1994 (New Zealand).

This thesis may be consulted by you, provided you comply with the provisions of the Act and the following conditions of use:

- Any use you make of these documents or images must be for research or private study purposes only, and you may not make them available to any other person.
- Authors control the copyright of their thesis. You will recognize the author's right to be identified as the author of this thesis, and due acknowledgement will be made to the author where appropriate.
- You will obtain the author's permission before publishing any material from their thesis.

General copyright and disclaimer

In addition to the above conditions, authors give their consent for the digital copy of their work to be used subject to the conditions specified on the Library Thesis Consent Form and Deposit Licence.
Approaches to Multiscale Inverse Problems

September, 2016

Ruanui Vincent Nicholson

Supervised by
Prof. Jari Kaipio

Co-supervised by
Dr. Steve Taylor

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy in Applied Mathematics,
The University of Auckland,
2016.
Abstract

Many scientific problems involve parameters such as conductivity, permeability or density which vary on multiple spatial and/or temporal scales. When such a parameter is to be estimated from noisy indirect measurements we face a challenging dichotomy: In practical situations the computational cost required to accurately take into account the small scale behaviour can be prohibitive. On the other hand, estimation of the conductivity is an ill-posed inverse problem meaning that any modelling errors, such as neglecting the small scale, which are not accounted for can render estimate useless.

A typical way of reducing the computational cost is to simply neglect any behaviour at the smaller scales, enabling a coarse discretisation of the problem. However, if the effects of disregarding smaller scale characteristics along with the coarse discretisation are not taken into account the estimates attained can be misleading.

In this thesis, we develop computationally feasible models to tackle the problem of estimating the multiscale distributed parameter of the Poisson equation, which is closely related to the multiscale electrical impedance tomography (EIT) problem. In the case of estimating a low dimensional representation of the conductivity (devoid of the small scale features) we show that by applying the Bayesian approximation error (BAE) approach to marginalise over small scale effects and discretisation errors we are able to substantially reduce the dimension of the problem while maintaining accurate estimates. Moreover, we show that by applying a feature extraction type modification to the BAE approach we can in some cases estimate the amplitude and correlation length of the small scale component of the conductivity.

In the case of estimating the full multiscale conductivity we show that multiscale finite element methods (MsFEM) can be implemented at both the forward modelling stage and at the inversion stage, which along with a somewhat coarsened discretisation can reduce overall computational cost. By deriving a closed form expression for the Jacobian matrix which represents the rate at which the electric potential calculated using MsFEM changes with respect to the conductivity we are able to apply gradient based optimisation techniques to estimate the conductivity. The BAE approach was also implemented in this procedure to take into account any modelling discrepancy caused by the use of MsFEM and discretisation errors. Use of such a procedure leads to estimates attained by using 7 multiscale basis functions in line with those calculated using finite elements on 85 linear basis functions.
Acknowledgement

This research would not exist had it not been for support and patience shown by many people.

To my wife Ana Ma and daughters Rae Marion Nicholson and Frances Liuaki Nicholson, the biggest thank you. Thank you for your love, your support, your patience, and your ability to pick me up time and time again. You are my world.

To my siblings, Jeanie Benson and Bill Benson, thank you for your continued support and love, it means so much.

To my Mother-in-law “Mum” Ika Ma, and Father-in-law “Dad” Iki Ma, and to all my in-laws, thank you for all that you do. I have never met such generous people, or people who smile so warmly.

I would like to thank my supervisor, Professor Jari Kaipio. You are so much more than a supervisor. Thank you for sharing your wisdom, your experience, your ideas, and your time. You may never know how much it helped me to hear you say “relax”. No student ever had a better supervisor. Thank you.

I am also greatly indebted to Dr. Steve Taylor, my co-supervisor. Your input to this research has been invaluable, and your patience with me is very much appreciated. I thank you for all your insights.

I would also like to acknowledge Prof. Tom ter Elst. Your willingness to help and ability to make things clear are amazing.

Thank you to Prof. James Sneyd, who went out of his way on numerous occasions to present me with awesome opportunities.

To my friends, the four other members of office 471, Anton Gulley, Attique Ur Rehman, Hwan Goh, and Owen Dillon, and past occupant Dr. Paul Hadwin, thanks for the memories and the endless laughter. Thank you also for all your help and discussions. Each one of you has contributed to this research in your own way, not least by pointing out my many mistakes.

I would also like to give a special thank you to Dr. Maryam Alavi, Jung Min (Sylvia) Han and Justin Clark. Thank you for your friendship and support.

Many thanks also go to the Mathematics Department as a whole including the administration staff. I have enjoyed chatting freely with all of you.

Thank you
Ruanui

This thesis is dedicated to the memories of my grandma Marion Braunstein, my grandpa Maurice Braunstein, my mother Rae Frances Nicholson, my brother Kahuakura Zane Erurea Nicholson, and my nephew (and godson) Stephen Erurea Dudley.

Let us keep close together not far apart
# CONTENTS

Abstract .......................................................................... i  
Acknowledgement ............................................................ ii  
Glossary ........................................................................... v  

1 Introduction ................................................................... 1  

2 Homogenisation Theory .................................................. 7  
  2.1 The Periodic Case .......................................................... 9  
  2.2 The Non-periodic Setting ............................................... 26  
  2.3 The Stochastic Setting .................................................. 26  
  2.4 Discussion .................................................................. 28  

3 Multiscale Finite Element Methods ................................. 31  
  3.1 Finite Element Methods .................................................. 32  
  3.2 Multiscale Finite Element Methods ................................. 36  
  3.3 Construction of Multiscale Basis Functions ..................... 37  
  3.4 Coupling the Scales ..................................................... 38  
  3.5 Basis Function Boundary Conditions .............................. 39  
  3.6 Implementation .......................................................... 41  
  3.7 Examples .................................................................... 43  
  3.8 Cost ............................................................................. 46  
  3.9 Proof of Convergence ................................................... 47  
  3.10 Discussion .................................................................. 54  

4 Inverse Problems Theory and Methodologies .................. 57  
  4.1 Deterministic Methods ................................................... 58  
  4.2 Bayesian Methods ....................................................... 62  
  4.3 Bayesian Approximation Error Approach ....................... 70
## CONTENTS

4.4 A Novel Extension of the Bayesian Approximation Error Approach to Estimate Auxiliary Unknowns ........................................ 74
4.5 Selection of a Prior Model .................................................. 80
4.6 Non-linear Least Squares ..................................................... 84
4.7 Discussion .................................................................. 89

5 Estimation of the Large Scale Component .................................. 91
5.1 The One Dimensional Case .................................................. 93
5.2 Recovery of $\alpha_S$ and $\beta_S$ ............................................. 101
5.3 The Two Dimensional Case .................................................. 107
5.4 Recovery of $\alpha_S$ and $\beta_S$ ............................................. 120
5.5 Discussion .................................................................. 121

6 Inversions using Multiscale Finite Element Methods ......................... 125
6.1 The Jacobian .................................................................. 126
6.2 Gradient Based Inversions Using MsFEM and BAE ..................... 129
6.3 In the Absence of Small Scale Behaviour ................................ 129
6.4 Presence of Small Scale Behaviour ..................................... 132
6.5 Differences in the Jacobians ............................................... 139
6.6 Discussion .................................................................. 140

7 Concluding Remarks ................................................................ 143

A Mathematical Appendix .......................................................... 145
A.1 Definitions .................................................................. 145
# GLOSSARY

## List of Mathematical Symbols

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ω</td>
<td>Bounded domain</td>
</tr>
<tr>
<td>∂Ω</td>
<td>Boundary of domain, typically piecewise continuous or Lipschitz</td>
</tr>
<tr>
<td>η</td>
<td>Outward facing unit vector normal to ∂Ω</td>
</tr>
<tr>
<td>ε</td>
<td>A small positive parameter typically much smaller than 1</td>
</tr>
<tr>
<td>σ</td>
<td>The electrical (or thermal) conductivity</td>
</tr>
<tr>
<td>σ_ε</td>
<td>Conductivity exhibiting behaviour on the ε-scale</td>
</tr>
<tr>
<td>σ_L</td>
<td>Part of conductivity varying at the (large) scale the problem is posed on</td>
</tr>
<tr>
<td>σ_S</td>
<td>Part of conductivity varying at the (small) ε-scale</td>
</tr>
<tr>
<td>σ*</td>
<td>The homogenised or effective conductivity</td>
</tr>
<tr>
<td>u</td>
<td>The electric potential</td>
</tr>
<tr>
<td>u_ε</td>
<td>Electric potential exhibiting behaviour on the ε-scale</td>
</tr>
<tr>
<td>[.]_z</td>
<td>The spatial mean of a function with respect to z</td>
</tr>
<tr>
<td>φ(x), φ</td>
<td>Finite element basis function</td>
</tr>
<tr>
<td>ϕ(x), ϕ</td>
<td>Finite element basis function on the fine mesh</td>
</tr>
<tr>
<td>ψ(x), ψ</td>
<td>Multiscale finite element basis function</td>
</tr>
<tr>
<td>π(·)</td>
<td>Probability distribution</td>
</tr>
<tr>
<td>e</td>
<td>Observation errors</td>
</tr>
<tr>
<td>ε</td>
<td>Approximation errors</td>
</tr>
<tr>
<td>Γ_x</td>
<td>Covariance of the random variable x</td>
</tr>
<tr>
<td>Γ_xy</td>
<td>Cross-covariance of random variables x and y</td>
</tr>
<tr>
<td>α</td>
<td>Variance</td>
</tr>
<tr>
<td>β</td>
<td>Correlation length</td>
</tr>
<tr>
<td>α_L</td>
<td>Variance at the large scale</td>
</tr>
<tr>
<td>β_L</td>
<td>Correlation length at the large scale</td>
</tr>
<tr>
<td>α_S</td>
<td>Variance at the small ε-scale</td>
</tr>
<tr>
<td>β_S</td>
<td>Correlation length at the small ε-scale</td>
</tr>
</tbody>
</table>
List of Mathematical Spaces

- $\mathbb{R}^d$ $d$-dimensional Euclidean space
- $T^d$ $d$-dimensional unit torus
- $C^k(\Omega), C^k$ Space of functions on $\Omega$ with continuous $k$th derivative
- $C^\infty(\Omega), C^\infty$ Space of functions on $\Omega$ with continuous $k$th derivative for all $k \in \mathbb{N}$
- $C^{k,\alpha}(\Omega), C^{k,\alpha}$ The space $C^k(\Omega)$ with $k$th derivatives $\alpha$-Hölder, and $0 \leq \alpha \leq 1$
- $L^p(\Omega), L^p$ Space of functions on $\Omega$ with finite $p$-norm, and $1 \leq p \leq \infty$
- $W^{s,p}(\Omega), W^{s,p}$ Space of functions on $\Omega$ with weak derivatives up to order $s$ in $L^p(\Omega)$
- $H^s(\Omega), H^s$ The space $W^{s,2}(\Omega)$
- $H^s_0(\Omega), H^s_0$ The space of functions in $H^s(\Omega)$ which vanish on $\partial \Omega$

List of Abbreviations

- PDE Partial differential equation
- EIT Electrical impedance tomography
- FEM Finite element method
- MsFEM Multiscale finite element method
- PG Petrov-Galerkin
- LS Least squares
- SVD Singular value decomposition
- TR Tikhonov regularisation
- MAP Maximum a posteriori
- CM Conditional mean
- ML Maximum likelihood
- MCMC Markov chain Monte Carlo
- BAE Bayesian approximation error
## LIST OF FIGURES

2.0.1 Examples of multiscale conductivity, $\sigma_\varepsilon$, in one and two dimensions. ................. 8  
2.1.1 One dimensional homogenisation example. ......................................................... 22  
2.1.2 Two dimensional homogenisation example. ......................................................... 23  
2.2.1 Non-periodic homogenisation example. .............................................................. 27  
3.3.1 Red refinement. ...................................................................................... 37  
3.3.2 A typical MsFEM mesh. ................................................................. 38  
3.5.1 Multiscale basis functions. .................................................................... 41  
3.7.1 One dimensional MsFEM example. ......................................................... 44  
3.7.2 Two dimensional MsFEM example true solution. ..................................... 45  
3.7.3 Two dimensional MsFEM example meshes. ................................................... 45  
3.7.4 Two dimensional MsFEM example comparison of solutions. .................... 46  
3.8.1 Two dimensional MsFEM example computation time. ............................... 48  
4.4.1 Feature selection spaces. ......................................................................... 80  
4.5.1 Draws from exponential squared prior. ....................................................... 82  
4.5.2 Marginal and joint hyper-priors. ................................................................. 83  
5.0.1 An example of a heterogeneous background. ............................................. 91  
5.1.1 One dimensional draws from priors. ......................................................... 94  
5.1.2 One dimensional approximation error statistics. ....................................... 95  
5.1.3 One dimensional estimates with no small scale behaviour. ..................... 97  
5.1.4 One dimensional estimates with relatively little small scale behaviour. .... 98  
5.1.5 One dimensional estimates with substantial small scale behaviour. .......... 99  
5.1.6 One dimensional physical example draws priors. ..................................... 100  
5.1.7 One dimensional physical example true conductivities. ............................ 101  
5.1.8 One dimensional physical example reconstructions. .................................... 101  
5.2.1 One dimensional recovery of auxiliary variables conditioned on $\nu$. ...... 103
5.2.2 One dimensional recovery of auxiliary variables conditioned on $F(\nu)$. . . . . . . 104
5.2.3 One dimensional variance of auxiliary variables conditioned on $F(\nu)$. . . . . . 105
5.2.4 One dimensional marginal priors and posteriors of auxiliary variables. . . . . . 105
5.2.5 One dimensional joint prior and posterior of auxiliary variables. . . . . . . . . . 106
5.2.6 One dimensional recovery of full conductivity. . . . . . . . . . . . . . . . . . . . 106
5.3.1 Two dimensional true conductivity. . . . . . . . . . . . . . . . . . . . . . . . . . 107
5.3.2 Two dimensional meshes used. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 108
5.3.3 Two dimensional data used. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 108
5.3.4 Two dimensional draws from large scale prior. . . . . . . . . . . . . . . . . . . . 109
5.3.5 Two dimensional reference case. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 110
5.3.6 Two dimensional hyper-priors. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 111
5.3.7 Two dimensional draws from small scale prior. . . . . . . . . . . . . . . . . . . . 111
5.3.8 Two dimensional approximation error meshes. . . . . . . . . . . . . . . . . . . . 112
5.3.9 Two dimensional approximation error statistics. . . . . . . . . . . . . . . . . . . 113
5.3.10 Two dimensional discretised true conductivities. . . . . . . . . . . . . . . . . . . 113
5.3.11 Two dimensional marginalisation over small scale. . . . . . . . . . . . . . . . . . 114
5.3.12 Two dimensional marginalisation over small scale with increased noise. . . . . 115
5.3.13 Two dimensional marginalisation over discretisation. . . . . . . . . . . . . . . . . 116
5.3.14 Two dimensional marginalisation over discretisation with increased noise. . . . 117
5.3.15 Two dimensional marginalisation over small scale and discretisation. . . . . . 118
5.3.16 Two dimensional marginalisation over coarsest discretisation. . . . . . . . . . . 120
5.3.17 Two dimensional marginalisation over small scale and coarsest discretisation. . 121
5.4.1 Two dimensional recovery of auxiliary unknowns. . . . . . . . . . . . . . . . . . . 122
5.4.2 Two dimensional marginal priors and posteriors of auxiliary variables. . . . . . 122
5.5.1 Two dimensional structure of the errors. . . . . . . . . . . . . . . . . . . . . . . . 123
6.3.1 MsFEM inversion draws from large scale prior. . . . . . . . . . . . . . . . . . . . 130
6.3.2 MsFEM inversion estimates with no small scale behaviour and lowest noise. . 131
6.3.3 MsFEM inversion estimates with no small scale behaviour and low noise. . . . . 132
6.3.4 MsFEM inversion estimates with no small scale behaviour and more noise. . . . 133
6.4.1 MsFEM inversion draws from priors. . . . . . . . . . . . . . . . . . . . . . . . . . 133
6.4.2 MsFEM inversion approximation error statistics. . . . . . . . . . . . . . . . . . . 135
6.4.3 MsFEM inversion estimates with lowest noise. . . . . . . . . . . . . . . . . . . . 136
6.4.4 MsFEM inversion estimates with low noise. . . . . . . . . . . . . . . . . . . . . . 137
6.4.5 MsFEM inversion estimates with more noise. . . . . . . . . . . . . . . . . . . . . 138
6.5.1 Comparison of Jacobians at initial estimate. . . . . . . . . . . . . . . . . . . . . 140
6.5.2 Comparison of Jacobians after several steps. . . . . . . . . . . . . . . . . . . . . 140
6.5.3 Comparison of posterior covariance. . . . . . . . . . . . . . . . . . . . . . . . . . 141
CHAPTER 1

INTRODUCTION

The necessity to assimilate observational data and mathematical models is a problem ubiquitous in all fields of science and leads to inverse problems. Inverse problems can loosely be defined as the problem of identifying parameters of interest from indirect noisy observations. Such problems are ill-posed in the sense of Hadamard [1]. Examples of such problems are everywhere, from the study of interacting biological species [2] to locating submarines [3], and from hydrology [4] to cosmology [5].

It is becoming increasingly apparent that in a vast array of settings, parameters of interest can vary over multiple spatial and/or temporal scales, resulting in what has become known as multiscale problems. Many scientific problems exhibit these multiscale characteristics, and examples can be found in all areas: From Wall Street stock prices [6] to analysis of defibrillators [7], and from geological prospecting [8] to searching for stars [9].

Amalgamating the study of inverse problems and multiscale behaviour is a natural step, and has lead to the relatively new field of multiscale inverse problems. Such a problem could be defined as the problem of identifying a set of multiscale parameters from indirect noisy observations. However, multiscale inverse problems are ambiguous: We may on the one hand wish to identify the set of multiscale parameters, or contrast, we may wish to find an approximation of the multiscale parameters which exhibit behaviour only at the largest scale. The definition one uses is problem dependent, and steps taken in the analysis are specific to the end goal.

We will concern ourselves with a multiscale version of the much celebrated Calderón inverse conductivity problem [10]. More precisely, we look at a multiscale adaptation of the following problem. Let $\Omega$ be a bounded domain in $\mathbb{R}^d$, $d = 1, 2, 3$ with Lipschitz boundary $\partial \Omega$. Take $\eta$ to be an outward facing unit vector normal to $\partial \Omega$, and furthermore suppose $\sigma \in L^\infty(\Omega)$ is
bounded away from zero. Then, given

\[-\nabla \cdot (\sigma \nabla u) = f, \quad \text{in} \quad \Omega\]

\[\sigma \nabla u \cdot \eta = g \quad \text{on} \quad \partial \Omega,\]

can we determine \(\sigma\) on measurements of \(u\)? A multiscale counterpart to such a problem is the case in which \(\sigma\) varies at multiple spatial scales, namely at the scale upon which the problem is posed, and on at least one much smaller scale, in this case the associated partial differential equation (PDE) is

\[-\nabla \cdot (\sigma_\varepsilon (x, x/\varepsilon) \nabla u_\varepsilon (x, x/\varepsilon)) = f(x), \quad \text{in} \quad \Omega\]

\[\sigma \nabla u_\varepsilon \cdot \eta = g \quad \text{on} \quad \partial \Omega,\]

where \(0 < \varepsilon \ll 1\). We will also consider the Dirichlet version of the problem.

The non-multiscale version of the problem has lead to several practical imaging modalities, including electrical impedance tomography (EIT) [11]. In this imaging technique the internal conductivity distribution of an object is estimated by attaching electrodes to the boundary of the object and injecting current through the electrodes. The resulting voltages on all electrodes are then measured and the conductivity distribution of the target is inferred. We will treat a version of such a problem with a simpler measurement system.

Applications of EIT include the fields of medicine [12], geophysics [13] and nondestructive testing of materials [14], for an comprehensive summary of the method, see for example [15]. We will mostly concern ourselves with the contexts of geophysical prospecting and nondestructive testing of concrete or other composite materials. In both of these settings it is natural to take conductivity as being multiscaled: The earth’s subsurface has heterogeneities varying on a huge range of scales, while the constituents making up composite materials such as concrete are of varying size.

In these applications, there are large amounts of uncertainty present. Foremost is the uncertainty around the actual size of the small scale(s), and the possible values the conductivity may take on the small scale(s). Information on the small scale dynamics is typically not available or is expensive to collect. Thus an approximation may have to be used. However, inverse problems are ill-posed in the sense of Hadamard [1], and thus, the use of an incorrect approximation can cause severe errors in the estimates.

The use of EIT in order to estimate the entire multiscale conductivity is a valid idea to pursue, and requires some attention. This type of problem could be considered a true multiscale inverse problem. Such problems inherit the ill-posedness of inverse problems while also suffering from the curse of dimensionality from attempting to capture the multiscale phenomena.

Several computational algorithms have been developed since the 1980’s in an attempt to maintain accuracy in solving multiscale forward problems while simultaneously reducing the computational overheads. One such class of methods are multiscale finite element methods.
(MsFEM) [16], which grew out of Babuška and Osborn’s generalised finite element method (GFEM) [17]. Such methods have been successfully applied to multiscale inverse problems, however, these have been carried out entirely using sampling based techniques such as Markov chain Monte Carlo (MCMC) methods [18, 19].

Implementing MsFEM can ease the computational complexity of the overall MCMC algorithm, as the cost for each forward solve is reduced by the used of a drastically smaller system matrix. However, using MsFEM to carry out forward simulations does not reduce the large number of samples required [20]. In this thesis we will show that MsFEM can in fact be used with gradient based methods for inverse problems.

Gradient based inversions, negate the need for samples, but do require both forward simulations and the Jacobian matrix which relates the information on how the calculated potential changes with respect to the conductivity. To reduce the cost of the forward simulations MsFEM can be used, which also means the Jacobian of the derivatives of the potential calculated using MsFEM with respect to the conductivity must also be computed. Moreover, the use of MsFEM can be seen as the implementation of an approximate model. Failure to account for errors induced by the approximative model can lead to misleading estimates of the conductivity [21, 22].

Another source of uncertainty, most often ignored, is the level of discretisation used for estimation [23]. Ideally a fine discretisation would be used, however, due to the small scale structure of the problem the level of discretisation required would lead to excessive (most likely prohibitive) computational time and resource requirements. This often necessitates coarser discretisation along with an approximate model, and adds another layer of errors which should be accounted for.

In the presence of these uncertainties, the Bayesian framework for inverse problems is a natural setting to attempt to solve such problems. Through repeated use of Bayes’ theorem [24] we are able to model all unknowns including the uncertainties listed as random variables. Moreover, in the Bayesian approach if any of the uncertainties are able to be parametrised by a small number of variables, it may be possible to also infer these [25]. For example we may be able to infer the amplitude and correlation length of the small scale behaviour of the conductivity.

The Bayesian approximation error approach (BAE) [22, 21] can be used to accommodate a range of uncertainties, including model uncertainty [26, 27], and uncertainty in the boundary conditions [23, 28, 29]. The BAE approach attempts to combine all errors, measurement errors, modelling errors, discretisation errors, etc. into one sole error term, called the approximation errors. The actual value of the approximation errors generally remains unknown. However, the statistics of the approximation errors can be estimated from prior information of the uncertainties involved.

With the statistics of the approximation errors in hand, the analysis required to derive the estimated conductivity is no more difficult than in the standard case of simple additive measurement error. The BAE approach was originally proposed in [22] to compensate for discretisation errors. Later it was applied to perform an approximate marginalisation over
unknown auxiliary parameters in the context of diffuse optical tomography [26]. The BAE approach has since been successfully applied to approximately marginalise over all sorts of parameters and compensate for many types of modelling discrepancies in a wide range of inverse problems.

The aim of this thesis is twofold:

- Initially we investigate the application of the BAE approach to estimation of the large scale conductivity only, and accordingly we treat the small scale part of the conductivity as an auxiliary parameter.

- Secondly we investigate the potential to use gradient based inversion methods based on a MsFEM formulation of the forward problem to recover the full multiscale conductivity.

For the first problem we show that the BAE approach can be applied to approximately pre-marginalise over the small scale behaviour of the conductivity while simultaneously accounting for errors caused by a coarse discretisation. In some cases we are also able to characterise the small scale behaviour by inferring both the amplitude and correlation length of the small scale structure of the conductivity. We also give examples showing one can then incorporate these small scale parameters into a reformulated inverse problem to recover the full multiscale conductivity.

We show that the standard BAE approach to recover auxiliary parameters may fail to give meaningful estimates of the small scale amplitude and correlation length. To this end we propose a new extension to the BAE approach which can be used to enhance estimates of auxiliary unknowns, and apply this extension to recover the small scale amplitude and correlation length.

To address the problem of recovering of the entire multiscale conductivity using gradient based inversion methods for the MsFEM formulation of the forward problem we first outline an efficient implementation algorithm for MsFEM. We use this implementation to analytically derive the associated Jacobian matrix, allowing, what we believe to be, the first use of MsFEM for parameter estimation using gradient based methods. By using the BAE approach we show one can indeed recover the full multiscale conductivity using MsFEM. We also briefly compare results and the respective Jacobians between the use of MsFEM and finite element methods (FEM).

This thesis can broadly speaking be divided into two parts. The first component focuses on giving a theoretical foundation for the rest of the thesis and an outline of the numerical methods we shall use. The second part is dedicated to providing results and applications of the methods devised in the first section. In each chapter we give a brief review of relevant literature and results.

This thesis is organised as follows: In chapter 2 we outline the theory of homogenisation which underlies most recent approaches to multiscale problems. In chapter 3 we introduce MsFEM and give an implementation algorithm. The basic theory for inverse problems is reviewed in chapter 4 and we present a new extension to the BAE approach to estimate auxiliary parameters. In chapter 5 we consider the use of the BAE approach to marginalise...
over small scale structures and discretisation errors. Finally, in chapter 6, we derive the Jacobian matrix describing how the potential calculated using MsFEM changes with respect to the conductivity, and use this to carry out gradient based inversions for a one dimensional inverse problem motivated by the EIT problem.
Loosely speaking, homogenisation can be thought of as the mathematically justified practice of taking averages. In other words, homogenisation extracts homogeneous effective parameters from disordered or heterogeneous media. Indeed, consider the multiscale Poisson equation,

$$-\nabla \cdot \left( \sigma_{\varepsilon}(x, x_{\varepsilon}) \nabla u_{\varepsilon}(x, x_{\varepsilon}) \right) = f(x), \quad \text{in } \Omega$$

$$u_{\varepsilon} = g_D \quad \text{or} \quad \sigma \nabla u_{\varepsilon} \cdot \eta = g_N \quad \text{on } \partial \Omega,$$

with $\varepsilon \ll 1$. Such a problem governs both the electric potential and temperature distribution in a body $\Omega$ with boundary $\partial \Omega$, where the electrical (resp. thermal) conductivity, $\sigma_{\varepsilon}$, varies at two scales, and $u$ is the electric potential (resp. temperature). We define the two scales as: The large, or coarse, scale, which is of order $\Omega$ which we take to be $O(1)$; and the small, or fine, scale, which is of order $\varepsilon$. Some examples of such a conductivity are shown in Figure 2.0.1. The right hand side allows for any electrical (resp. thermal) sinks or sources within $\Omega$, and $u$ represents the electric potential (resp. temperature), which, as a direct result of the conductivity, is itself multiscaled.

Poisson’s problem is relatively well understood, see for example [30], and numerically solving such a problem can be achieved by standard methods such as finite element methods (FEM), or finite difference schemes. On the other hand, the multiscale version of the problem can be difficult to solve accurately. When considering the multiscale problem several questions should be considered:

- Is there some replacement for $\sigma_{\varepsilon}(x, x_{\varepsilon})$ which varies only on the larger scale, say $\sigma^*(x)$, that when substituted into equation (2.0.1) results in a solution, say $u_0(x)$, which is a good approximation of $u_{\varepsilon}(x, x_{\varepsilon})$?

- Under what assumptions could we carry out such a substitution?

- How good would the approximation of $u_{\varepsilon}(x, x_{\varepsilon})$ by $u_0(x)$ be?
Chapter 2. Homogenisation Theory

Figure 2.0.1: Examples of multiscale conductivity, $\sigma_\epsilon$, in one and two dimensions.

The idea of substitution for simplification is an idea dating back to time immemorial, and there has been renewed interest in the field with the development of so called surrogate modelling [31]. However, in this specific setting these questions gave rise to what is now known as the theory of homogenisation.

In the early 1950’s Bishop and Hill were able to prove upper bounds for the yield surface of polycrystalline solids by approximating the strain-rate and stress fields as uniform [32]. This was not the first progress in the field, but it was one of the first results to be proven, and it gave way to many similar breakthroughs. In the 60’s Hashin and Shtrikman [33] derived a variational principle to bound the overall conductivity of a two-phase heterogeneous composite. The 1970’s saw huge developments in homogenisation theory, with the idea of asymptotic homogenisation being at the fore. The most influential work of this era is that of Bensoussan et al. [34], where a full analysis of the homogenisation procedure for periodic structures was given. The ideas of Bensoussan et al. are still the starting point for most work on homogenisation theory, and have lead to huge interest and progress in the field. Other notable efforts in homogenisation theory include the early work of Eshelby [35], the idea of compensated compactness developed by Tartar [36] and Murat [37], and the book of Jikov et al. [38].

The order of our brief analysis of homogenisation theory will loosely follow the order in which it has been developed since the 70’s. We point out that the results in this section are also available in numerous text. We start with the periodic case before dropping this assumption to look at both the non-periodic and stochastic settings. In each case there is a certain number of definitions and specific notation which are needed to be introduced before further work can be carried out. For a more in-depth overview of homogenisation see any of [34, 39, 38, 40, 41, 42].

Throughout the analysis we will concern ourselves with the Poisson problem of the form

$$-\nabla \cdot \left( \sigma_\epsilon \left( \frac{\mathbf{x}}{\varepsilon} \right) \nabla u_\epsilon \left( x, \frac{\mathbf{x}}{\varepsilon} \right) \right) = f(x) \quad \text{in } \Omega$$

$$u_\epsilon \left( x, \frac{\mathbf{x}}{\varepsilon} \right) = 0 \quad \text{on } \partial \Omega,$$

(2.0.2)
where \( \Omega \) is an open bounded subset of \( \mathbb{R}^N \) with Lipschitz boundary \( \partial \Omega \), \( \sigma_\varepsilon \in L^\infty (\Omega) \) and positive, \( f \in L^2 (\Omega) \), and \( 0 < \varepsilon < 1 \) is a small parameter. We closely follow the works of Bensoussana et al. [34] and Pavliotis and Stuart [43]. The analysis in this chapter is equally valid for \( \sigma_\varepsilon \) being a real positive definite matrix, allowing for anisotropic conductivity. Unless otherwise stated, we take \( \sigma_\varepsilon \) to be a real valued function in \( L^\infty (\Omega) \) although in some cases we keep the notation general to allow for anisotropy.

2.1 The Periodic Case

Throughout this section \( \varepsilon \) will be used to denote the size of a period. To formalise this assumption we must first introduce the following definition:

**Definition 2.1.1.** A function \( f : \mathbb{R}^N \rightarrow \mathbb{R} \) is said to be \( k \)-periodic if for \( k \in \mathbb{R} \),

\[
f(x + ke_i) = f(x) \quad \forall x \in \mathbb{R}^N \quad i = 1, 2, \ldots, N,
\]

with \( e_i \) the canonical basis of \( \mathbb{R}^N \).

It can also be helpful to think of \( k \)-periodicity, for \( k \in \mathbb{R} \), through the following definition:

**Definition 2.1.2.** The \( N \)-dimensional unit torus, denoted \( \mathbb{T}^N \), is the identification of the \( N \)-dimensional unit (hyper-)cube with the addition of periodic Dirichlet boundary conditions. Then the dilation of the unit (hyper-)cube by \( k \in \mathbb{R} \) with periodic Dirichlet boundary conditions will be denoted \( k\mathbb{T} \).

The periodic setting also comes with peculiar compatibility conditions for the Poisson problem outlined in (2.0.2) which we will draw on numerous times.

**Theorem 2.1.3.**

Let \( \sigma_\varepsilon \in L^\infty (\varepsilon \mathbb{T}) \), that is \( \sigma_\varepsilon \) is \( \varepsilon \)-periodic and in \( L^\infty (\Omega) \), then for \( f \in L^2 (\Omega) \), the periodic Poisson problem,

\[
- \nabla \cdot (\sigma_\varepsilon \nabla u) = f
\]

\[
u : \varepsilon \mathbb{T}^N \rightarrow \mathbb{R}
\]

admits a solution unique up to an additive constant if and only if

\[
\int_{\mathbb{T}^N} f \, dx = 0.
\]

The proof can be realised by application of the Lax-Milgram theorem and integration, and is presented in, for example, [34].

We will also be interested at times in taking the spatial mean of a function, to which end we introduce the following definition and notation,
Definition 2.1.4. The spatial mean, or simply mean, of a function \( f : \Omega \rightarrow \mathbb{R} \) is given by

\[
[f] = \int_{\Omega} f \, dx.
\]

The spatial mean can also be taken with respect to any set of variables, let \( f : \Omega_x \times \Omega_y \rightarrow \mathbb{R} \) with \( x \in \Omega_x \) and \( y \in \Omega_y \) then the spatial mean of \( f \) with respect to \( y \) is

\[
[f]_y(x) = \int_{\Omega_y} f(x, y) \, dy.
\]

Imposing that the solutions \( u \) in theorem 2.1.3 are mean 0, i.e. \( [u] = 0 \) ensures uniqueness of the solution. We now proceed with our exploration of homogenisation theory.

2.1.1 Two-Scale Expansions

Many methods have been developed to approach the problem of periodic homogenisation. Of these methods the most intuitive (although lacking in rigour) is the method of two scale expansions. The method is presented here, while precise mathematical justification is given in section 2.1.2.

In what follows, it is important to keep in mind the end goal: We wish to find a suitable counterpart to problem (2.0.2), the solution of which no longer possesses behaviour at the \( \varepsilon \) scale, but is simultaneously a good approximation to the solution of (2.0.2).

Initially we consider the following two-scale asymptotic expansion,

\[
u_\varepsilon(x) = \sum_{k=1}^{\infty} \varepsilon^k u_k \left( x, \frac{x}{\varepsilon} \right),
\]

where each \( u_k \) is \( \varepsilon \)-periodic in the second argument. By making the substitution \( y = \frac{x}{\varepsilon} \), we have

\[
u_\varepsilon(x) = \sum_{k=1}^{\infty} \varepsilon^k u_k (x, y),
\]

where each \( u_k \) is 1-periodic in \( y \), i.e. \( u_k : \mathbb{R}^N \times \mathbb{T}^N \rightarrow \mathbb{R} \). This substitution also implies that

\[
\nabla u_k \left( x, \frac{x}{\varepsilon} \right) = \varepsilon^{-1} \nabla_y u_k (x, y) + \nabla_x u_k (x, y),
\]

where \( \nabla = \left[ \frac{\partial}{\partial y_1}, \frac{\partial}{\partial y_2}, \ldots, \frac{\partial}{\partial y_N} \right]^T \) and \( \nabla_x = \left[ \frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \ldots, \frac{\partial}{\partial x_N} \right]^T \). From this definition of the total derivative we can expand the differential operator, \(-\nabla \cdot (\sigma_\varepsilon \nabla)\) into a power series in \( \varepsilon \) as,

\[
L_\varepsilon = \varepsilon^{-2} L_0 + \varepsilon^{-1} L_1 + L_2,
\]
with each operator defined,

\[ L_0 := -\nabla_y \cdot (\sigma \nabla_y), \quad L_1 := -\nabla_y \cdot (\sigma \nabla_x) - \nabla_x \cdot (\sigma \nabla_y), \quad \text{and} \quad L_2 := -\nabla_x \cdot (\sigma \nabla_x). \]

Upon substituting the series (2.1.2) into equation (2.0.2) and using (2.1.3) we arrive at another series in \( \varepsilon \), namely,

\[ \varepsilon^{-2}[L_0 u_0] + \varepsilon^{-1}[L_0 u_1 + L_1 u_0] + L_0 u_2 + L_1 u_1 + L_2 u_0 + \varepsilon[L_0 u_3 + L_1 u_2 + L_2 u_1] + \mathcal{O}(\varepsilon^2) = f. \tag{2.1.4} \]

By equating powers of \( \varepsilon \), we obtain the following cascading problems,

\[
\begin{align*}
\mathcal{O}(\varepsilon^{-2}) &: \quad L_0 u_0 = 0, \\
\mathcal{O}(\varepsilon^{-1}) &: \quad L_0 u_1 = -L_1 u_0, \\
\mathcal{O}(\varepsilon^0) &: \quad L_0 u_2 = -L_1 u_1 - L_2 u_0 + f, \\
\mathcal{O}(\varepsilon^1) &: \quad L_0 u_3 = -L_1 u_2 - L_2 u_1.
\end{align*}
\]

Generally, terms beyond the zeroth (constant) order term are not considered. However we retain the first order term as it is required to prove theorem 2.1.6. We also note that all higher order terms satisfy

\[ L_0 u_k = -L_1 u_{k-1} - L_2 u_{k-2}, \quad \text{for} \quad k = 4, 5, \ldots. \]

The operator \( L_0 \) is a differential operator in \( y \) only, and each \( u_k(x, y) \) is 1-periodic in \( y \). This implies that from equation (2.1.5) our solution must be of the form,

\[ u_0(x, y) = u_0(x). \tag{2.1.9} \]

Since \( u_0 \) is a function of \( x \) only and \( \sigma \) is a function of \( y \) only, we can rewrite equation (2.1.6) as

\[ L_0 u_1 = (\nabla_y \cdot \sigma^T) \cdot \nabla_x u_0, \tag{2.1.10} \]

and due to the fact that \( L_0 \) is a differential operator in \( y \) only, we seek solutions of (2.1.10) in the form

\[ u_1(x, y) = \chi(y) \cdot \nabla_x u_0(x) + \hat{u}_1(x). \tag{2.1.11} \]

By substituting solutions of this form into the original problem, equation (2.1.10), we see that
the first order corrector, $\chi(y)$, satisfies
\[
-\nabla_y \cdot \left( \sigma (\nabla_y \chi)^T \right) = \nabla_y \cdot \sigma \\
\chi : \mathbb{T}^N \to \mathbb{R}^N \quad (2.1.12)
\]
\[
\int_{\mathbb{T}^N} \chi \, dy = 0.
\]
Problem (2.1.12) is often referred to as the cell problem, and is equivalent along with the same conditions as
\[
-\nabla_y \cdot (\nabla_y \chi \sigma^T) = \nabla_y \cdot \sigma^T,
\]
which can be helpful for analysis. Moreover the first order corrector can be solved for component-wise by finding solutions of
\[
L_0 \chi_i = \nabla_y \cdot \sigma e_i \\
\chi_i : \mathbb{T}^N \to \mathbb{R}^N \quad (2.1.13)
\]
\[
\int_{\mathbb{T}^N} \chi_i \, dy = 0,
\]
where $e_i$ are the canonical basis of $\mathbb{R}^N$, and $i = 1, 2, \ldots, N$.

The compatibility condition given in theorem 2.1.3 implies that the solvability of equation (2.1.7) is ensured if
\[
\int_{\mathbb{T}^N} L_2 u_0 + L_1 u_1 \, dy = \int_{\mathbb{T}^N} f \, dy = f. \quad (2.1.14)
\]
We treat each term in (2.1.14) separately. Firstly,
\[
\int_{\mathbb{T}^N} L_2 u_0 \, dy = -\int_{\mathbb{T}^N} \nabla_x \cdot (\sigma \nabla_x u_0) \, dy = -\nabla \cdot ([\sigma]_y \nabla u_0),
\]
since $u_0 = u_0(x)$. For the second term we also substitute in the form of $u_1$,
\[
\int_{\mathbb{T}^N} L_2 u_0 \, dy = -\int_{\mathbb{T}^N} \nabla_y \cdot (\sigma \nabla_x u_1) + \nabla_x \cdot (\sigma \nabla_y u_1) = -\int_{\mathbb{T}^N} \nabla_x \cdot (\sigma \nabla_y (\chi \cdot \nabla_x u_0)),
\]
since $\sigma$ is 1-periodic. However we can further manipulate this term,
\[
-\int_{\mathbb{T}^N} \nabla_x \cdot (\sigma \nabla_y (\chi \cdot \nabla_x u_0)) = -\int_{\mathbb{T}^N} \nabla_x \cdot \left( \sigma (\nabla_y \chi)^T \nabla_x u_0 \right) \, dy = -\nabla \cdot \left( [\sigma (\nabla_y \chi)^T]_y \nabla u_0 \right),
\]
since $u_0$ is independent of $y$.

The reworking of both equation (2.1.6) and equation (2.1.7) leads us to conclude that $u_0$ is the unique solution of
\[
-\nabla \cdot \left( [\sigma + \sigma (\nabla_y \chi)^T]_y \nabla u_0 \right) = f \quad \text{in } \Omega \quad (2.1.15)
\]
\[
u_0 = 0 \quad \text{on } \partial \Omega.
\]
This new problem is usually called the homogenised, or effective equation, while \([\sigma + \sigma (\nabla_y \chi)^T]\)_y is most often denoted \(\sigma^*\), and called the homogenised or effective diffusion (tensor), and \(u_0\) is referred to as the homogenised solution. It will turn out to be useful to also define the effective diffusion elementwise (in the case of anisotropic, matrix valued conductivity),

\[
\sigma^*_{ij} = [\sigma \varepsilon (e_i + \nabla_y \chi_i) \cdot e_j]_y. \tag{2.1.16}
\]

Assuming \(\sigma = \sigma(y)\) is simplifying, but is not necessary. If instead we have \(\sigma = \sigma(x, y)\), then through similar analysis we arrive at a almost identical homogenised equation,

\[
-\nabla \cdot \left( [\sigma(x, y) + \sigma(x, y) (\nabla_y \chi(x, y))^T]_y \nabla u_0 \right) = f \quad \text{in } \Omega \\
u_0 = 0 \quad \text{on } \partial \Omega,
\]

with \(\sigma^*(x) = [\sigma(x, y) + \sigma(x, y) (\nabla_y \chi(x, y))^T]_y\).

The uniqueness of \(u_0\) is guaranteed by the Lax-Milgram theorem, which relies on \(\sigma^*\) being positive definite. This premise is satisfied so long as \(\sigma\) is positive definite, that is to say we have the following theorem.

**Theorem 2.1.5.**

Let \(\sigma(x)\) be a uniformly elliptic 1-periodic matrix, then the homogenised diffusion

\[
\sigma^* = [\sigma(x, y) + \sigma(x, y) (\nabla_y \chi(x, y))^T]_y
\]

is also uniform elliptic, moreover, symmetry of \(\sigma\) implies symmetry of \(\sigma^*\).

For a proof of this theorem see for example [43]

Even with \(\sigma = \sigma(y)\), equation (2.1.15) is generally only analytically solvable for essentially one dimensional problems, these are a class of problems such that \(\sigma(y) = \sigma_1(y_1) \cdot \sigma_2(y_2) \cdot \ldots \cdot \sigma_N(y_N)\), or in the case of so called layered or laminar materials where \(\sigma(y) = \sigma(y_k)\), for a \(k\) between 1 and \(N\). Two notable exceptions to this rule are the cases of two-phase and four-phase checkerboard composites. In the first of these cases the homogenised conductivity can analytically be solved for and turns out to be the geometric mean of \(\sigma(y_1, y_2)\), see for example [38] for consideration of this case. The four-phase checkerboard is often referred to as the Mortola-Steffe structure, as Mortola and Steffe conjectured a closed form for the homogenised coefficient [44], however the conjecture was not proven until 2000, see [45].

**Higher order terms and an interesting observation on \(\hat{u}_1(x)\) in one dimension**

In most texts higher order terms in the two-scale expansion are suppressed into an overarching corrector term [34]. The corrector term will be treated in section 2.1.4 and its importance will be illuminated. However, we now take a look at the higher order terms, and in particular the first order terms. We continue with the assumption that \(\sigma = \sigma(y)\).

The general form of the first order term in the two-scale expansion was derived previously,
Chapter 2. Homogenisation Theory

and given in equation (2.1.11). We restate it here for ease of reference,

\[ u_1(x, y) = \chi(y) \cdot \nabla_x u_0(x) + \hat{u}_1(x), \]

where \( \chi \) is the first order corrector which satisfies the cell problem, equation (2.1.12). At this point we know \( u_0 \) and \( \chi \), and hence we only seek \( \hat{u}_1(x) \) to determine \( u_1 \) completely. To gain convergence, the natural boundary conditions to enforce are

\[ \hat{u}_1(x) = -\chi(y) \cdot \nabla_x u_0(x) \quad \text{for } x \text{ and } y \text{ on } \partial\Omega., \]  

(2.1.17)

however, depending on how \( \varepsilon \) goes to 0 this may cause some problems [34], which are not of concern here.

In order to solve for \( \hat{u}_1 \) we use our knowledge of the terms studied so far and look again at equation (2.1.7),

\[ L_0 u_2 = \nabla_y \cdot (\sigma \nabla_x u_1) + \nabla_x \cdot (\sigma \nabla_y u_1) + \nabla_x \cdot (\sigma \nabla_x u_0) - \nabla_x \cdot (\sigma^* \nabla_x u) \]

(2.1.18)

\[ = \nabla_y \cdot (\sigma \nabla_x (\chi \cdot \nabla_x u + \hat{u}_1)) + \nabla_x \cdot (\sigma \nabla_y (\chi \cdot \nabla_x u_0) + (\sigma \nabla_x u_0) - \sigma^* \nabla_x u_0) \]

\[ = \nabla_y \cdot (\sigma \nabla_x \hat{u}_1) + \nabla_x \cdot \left( \left( \nabla_y \chi \sigma + \chi \left( \nabla_y \cdot \sigma^T \right)^T + \sigma (\nabla_y \chi)^T + \sigma^* \right) \nabla_x u_0 \right). \]

(2.1.19)

In equation (2.1.18) we have used the fact that \( f = \nabla_x \cdot (\sigma^* \nabla_x u_0) \). From equation (2.1.19) we deduce

\[ u_2(x, y) = \nabla_x \cdot (\Theta(y) \nabla_x u_0(x)) + \chi(y) \cdot \nabla_x \hat{u}_1(x) + \hat{u}_2(x). \]

(2.1.20)

Substituting solutions of this form back into equation (2.1.19) we see that the matrix-valued second order corrector, \( \Theta(y) \), satisfies

\[-\nabla_y \cdot (\sigma \nabla_y \Theta) = \nabla_y \chi \sigma + \chi \left( \nabla_y \cdot \sigma^T \right)^T + \sigma (\nabla_y \chi)^T + \sigma - \sigma^* \]

\[ \Theta : \mathbb{T}^N \to \mathbb{R}^{N \times N} \]

\[ \int_{\mathbb{T}^N} \Theta_{ij} \, dy = 0. \]

(2.1.21)

The fact that each element in \( \Theta \) is zero mean ensures uniqueness, as without this assumption \( \Theta \) is only known up to an additive constant. However, for our purposes we could in fact drop this restriction, as this additive constant will play no role in our analysis.

Application of the compatibility condition given in theorem 2.1.3 to the equation (2.1.8)
Chapter 2. Homogenisation Theory

\begin{align*}
0 &= -\int_{\mathbb{T}^N} L_1 u_2 + L_2 u_1 \, dy \\
&= \int_{\mathbb{T}^N} \nabla_y \cdot (\sigma \nabla_x u_2) + \nabla_x \cdot (\sigma \nabla_y u_2) + \nabla_x \cdot (\sigma \nabla_x u_1) \, dy \\
&= \int_{\mathbb{T}^N} \nabla_x \cdot (\sigma \nabla_y (\nabla_x \cdot (\Theta \nabla_x u_0) + \chi \cdot \nabla_x \hat{u}_1) + \sigma \nabla_x (\chi \cdot \nabla_x u_0 + \hat{u}_1)) \, dy \\
&= -L^* \hat{u}_1 + \int_{\mathbb{T}^N} \nabla_x \cdot (\sigma \nabla_y (\Theta \nabla_x u_0) + \sigma \nabla_x (\chi \cdot \nabla_x u_0)) \, dy. \tag{2.1.22}
\end{align*}

where \( L^* := -\nabla \cdot \left( [\sigma + \sigma (\nabla_y \chi)^T] \nabla u_0 \right) \). Hence \( \hat{u}_1 \) is the (unique) solution of

\begin{align*}
L^* \hat{u}_1 &= \int_{\mathbb{T}^N} \nabla_x \cdot (\sigma \nabla_y (\Theta \nabla_x u_0) + \sigma \nabla_x (\chi \cdot \nabla_x u_0)) \, dy, \\
\hat{u}_1(x) &= -\chi(y) \cdot \nabla_x u_0(x) \quad \text{for } x \text{ and } y \text{ on } \partial \Omega. \tag{2.1.23}
\end{align*}

Taking into account the fact that both \( \sigma \) and \( \Theta \) are independent of \( x \) and \( u_0 \) is independent of \( y \) we can reformulate equation (2.1.23) in indicial notation as

\begin{align*}
L^* \hat{u}_1 &= \left( \int_{\mathbb{T}^N} \sigma_{ij} \frac{\partial}{\partial y_k} \Theta_{jk} + \sigma_{ij} \chi_k \right) \frac{\partial^2 u_0}{\partial x_i \partial x_j \partial x_k} \\
\hat{u}_1(x) &= -\chi(y) \cdot \nabla_x u_0(x) \quad \text{on } \partial \Omega. \tag{2.1.24}
\end{align*}

This result can also be found in the literature, see for example [34]. If the problem is posed in one dimension we have the following theorem,

**Theorem 2.1.6.**

Suppose \( u_0(x) \) and \( \sigma^* \) are defined as in equations (2.1.15), while \( \chi(y) \) and \( \Theta(y) \) are defined as in equations (2.1.12), and (2.1.21) respectively, then in dimension one \( \hat{u}_1 \) is the unique solution to the problem,

\begin{align*}
-\sigma^* \frac{d^2 \hat{u}_1}{dx^2} &= 0 \quad \text{in } \Omega \\
\hat{u}_1 &= -\chi(y) \frac{du_0}{dx}(x) \quad \text{on } \partial \Omega.
\end{align*}

**Proof.** Suppose we are in a one dimensional setting. Then the homogenised conductivity is

\begin{equation}
\sigma^* = \int_0^1 \sigma(y) + \sigma(y) \frac{d\chi}{dy} \, dy, \tag{2.1.25}
\end{equation}
while the one dimensional analogue of equation (2.1.21) for $\Theta$ is
\[
- \frac{d}{dy} \left( \sigma(y) \frac{d\Theta}{dy}(y) \right) = \sigma(y) + \sigma(y) \frac{d\chi(y)}{dy} + \frac{d}{dy} (\sigma(y)\chi(y)) - \sigma^* \\
\Theta : T \to \mathbb{R} \\
\int_0^1 \Theta \, dy = 0,
\]
which solved by integration and enforcing periodicity has solutions of the form
\[
\Theta(y) = - \int_0^y \chi(z) \, dz + c
\]
where $c$ is an unimportant constant which could be solved for by enforcing the zero mean condition on $\Theta$. Thus
\[
\frac{d\Theta}{dy}(y) = -\chi(y),
\]
and therefore
\[
-\sigma^* \frac{d^2\hat{u}_1}{dx^2} = \left( \int_0^1 \frac{d\Theta}{dy} + \sigma \chi \, dy \right) \frac{d^3u_0}{dx^3} = \left( \int_0^1 -\sigma \chi + \sigma \chi \, dy \right) \frac{d^3u_0}{dx^3} = 0.
\]

2.1.2 Justification of the Homogenisation Procedure

Although the method of two-scale asymptotic expansions is insightful, it lacks rigour, assumes high degrees of regularity, and cannot tell us anything about convergence. This section provides a brief mathematically sound justification of some homogenisation results. The results are based on the so called energy method, or otherwise called (Tartar’s) oscillating test functions method, and we follow quite closely the direction of [46]. There are numerous other approaches, perhaps the other most notable is the two-scale convergence method, see for example [47].

Recall the problem at hand,
\[
-\nabla \cdot \left( \sigma \left( \frac{x}{\varepsilon} \right) \nabla u_\varepsilon \right) = f \quad \text{in } \Omega \\
u_\varepsilon = 0 \quad \text{on } \partial \Omega.
\]
To proceed with the energy method we require some classical results concerning problem (2.1.26). We must first define two types of convergence:

**Definition 2.1.7.** A sequence $\{x_n\}$ in a (real) Banach space $E$ is said to converge weakly to $x$ iff
\[
\forall f \in E', \quad f(x_n) \to f(x),
\]
where $E'$ denotes the dual of $E$. We denote weak convergence by

$$x_n \rightharpoonup x \text{ weakly in } E.$$

**Definition 2.1.8.** Let $F$ be a Banach space and $E = F'$, that is $E$ is the dual of $F$. Then a sequence $\{f_n\}$ in $E$ is said to converge weakly$^*$ to $x$ iff

$$\forall x \in F, \ f_n(x) \to f(x).$$

We denote weak$^*$ convergence by

$$f_n \rightharpoonup f \text{ weakly}^* \text{ in } E.$$

These two notions of convergence are in fact equivalent for reflexive Banach spaces, and have many interesting properties. For an in depth discussion on both weak and weak$^*$ convergence see for example [48]. With these two ideas of convergence we now present the following theorem:

**Theorem 2.1.9.**

Let $1 \leq p \leq \infty$ and let $f_\varepsilon$ be a $1$-periodic function in $L^p(\mathbb{T})$ such that

$$f_\varepsilon = f\left(\frac{x}{\varepsilon}\right) = f(y).$$

Then for $1 \leq p < \infty$ as $\varepsilon \to 0$ we have

$$f_\varepsilon \rightharpoonup [f_\varepsilon]_y \text{ weakly in } L^p(\Omega).$$

If $p = \infty$ the convergence is weak$^*$.

The proof of this result is quite long, and can be found in numerous text on functional analysis, partial differential equations or homogenisation, see for example [41].

**Theorem 2.1.10.**

Problem (2.1.26) with $f$ in $L^2(\Omega)$ admits a unique solution, $u_\varepsilon$, in $H^1_0(\Omega)$ (defined in appendix A) which satisfies

$$\|u_\varepsilon\|_{H^1_0(\Omega)} \leq C \|f\|_{L^2(\Omega)}, \quad (2.1.27)$$

where $C$ is independent of both $f$ and $\varepsilon$.

The proof of this result is based on the Lax-Milgram theorem or the Riesz representation theorem and can be found in most texts on the theory of partial differential equations or functional analysis, see for example [49].

A further useful result is

**Theorem 2.1.11.**

A bounded sequence in a Hilbert space contains a weakly convergent subsequence.
This result is based on Cantor diagonalisation and Riesz representation theorem, a full proof is available in [50] for example.

Consider now the weak form of (2.1.26),

$$\text{find } u_\varepsilon \in H^1_0(\Omega) \text{ such that }$$

$$\int_{\Omega} \sigma\left(\frac{x}{\varepsilon}\right) \nabla u_\varepsilon(x) \cdot \nabla v(x) \, dx = \int_{\Omega} f(x) v(x) \, dx \quad \text{for all } v \in H^1_0(\Omega).$$

We will denote $$\sigma\left(\frac{x}{\varepsilon}\right)$$ by $$\sigma_\varepsilon$$ and $$u_\varepsilon(x)$$ by $$u_\varepsilon$$, and we interpret $$\varepsilon$$ now as an index for the sequences of both $$\sigma_\varepsilon$$ and $$u_\varepsilon$$ which goes to 0. Now, $$\sigma_\varepsilon$$ is bounded in $$L^2(\Omega)$$ while $$u_\varepsilon$$ is bounded in $$H^1_0(\Omega)$$ by theorem 2.1.10. Then there are subsequences $$\nabla u_\varepsilon$$ and $$\sigma_\varepsilon$$ indexed both by $$\varepsilon$$, which converge weakly in $$L^2(\Omega)$$. We will denote the weak limit of $$u_\varepsilon$$ by $$u_0(x)$$ and the weak limit of $$\sigma_\varepsilon$$ by $$\sigma^*$$.

The objective of homogenisation can be recast as determining the following,

$$\lim_{\varepsilon \to 0} \int_{\Omega} \sigma\left(\frac{x}{\varepsilon}\right) \nabla u_\varepsilon(x) \cdot \nabla v(x) \, dx = \int_{\Omega} \sigma^* \nabla u_0(x) \cdot \nabla v(x) \, dx,$$  \hspace{1cm} (2.1.29)

should the limit exist.

The difficulty in evaluating the limit is the fact that the product of the limits of two weakly convergent sequences does not in general converge to the limit of the product. To remedy this situation we use (Tartar’s) oscillating test functions [51, 52]. To define these new test functions we first let $$\zeta: \mathbb{T}^N \to \mathbb{R}^N$$ be the solution of the dual cell problem

$$-\nabla_y \cdot (\sigma^T \nabla_y \zeta_i) = \nabla_y \cdot \sigma^T e_i,$$

$$\zeta_i: \mathbb{T} \to \mathbb{R}$$

$$\int_{\mathbb{T}^N} \zeta_i \, dy = 0.$$  \hspace{1cm} (2.1.30)

for $$i = 1, 2, \ldots, N$$. Now, let $$\mathcal{D}(\Omega)$$ be the set of test functions on $$\Omega$$, that is, the set containing all $$C^\infty$$ functions with compact support in $$\Omega$$. We now define the oscillating test functions,

**Definition 2.1.12.** Suppose $$\phi(x)$$ is in $$\mathcal{D}(\Omega)$$ and $$\zeta\left(\frac{x}{\varepsilon}\right)$$ is the solution to problem (2.1.30), then the oscillating test functions are defined,

$$\phi_\varepsilon(x) = \phi(x) + \varepsilon \zeta\left(\frac{x}{\varepsilon}\right) \cdot \nabla \phi(x).$$  \hspace{1cm} (2.1.31)

It should be noted that $$\zeta \in C^\infty(\Omega)$$ (i.e. $$\zeta$$ is smooth with respect to the large scale variable, $$x$$)\(^2\) and since $$\zeta$$ is bounded and periodic on the small scale we have $$\zeta_i$$ bounded in $$H^1(\Omega)$$ for $$i = 1, 2, \ldots, N$$.

\(^1\)Recall, $$\sigma_\varepsilon = \sigma\left(\frac{x}{\varepsilon}\right)$$ and thus is constant with respect to the large scale variable, $$x$$, hence since $$\Omega$$ is bounded we have $$\sigma_\varepsilon \in L^2(\Omega)$$.

\(^2\)On the large scale $$\zeta$$ inherits the same regularity as that of the large scale of $$\sigma_\varepsilon$$. 

---

Chapter 2. Homogenisation Theory 18
As \( \phi_\varepsilon(x) = \phi_\varepsilon \left( x, \frac{y}{\varepsilon} \right) = \phi_\varepsilon(x, y) \) we have

\[
\nabla \phi_\varepsilon = \nabla_x \phi_\varepsilon + \varepsilon^{-1} \nabla_y \phi_\varepsilon = \left( I + (\nabla_y \zeta)^T \right) \nabla_x \phi + \varepsilon \nabla^2 \phi \zeta, \quad (2.1.32)
\]

where \( \nabla^2 \) denotes the Hessian matrix.

Now substituting the oscillating test function into the left hand side of the weak form of problem (2.1.28) yields

\[
\int_\Omega \sigma_\varepsilon \nabla u_\varepsilon \cdot \nabla \phi_\varepsilon \, dx = \int_\Omega \sigma_\varepsilon \nabla u_\varepsilon \cdot \left( I + (\nabla_y \zeta)^T \right) \nabla_x \phi \, dx + \varepsilon \int_\Omega \sigma_\varepsilon \nabla u_\varepsilon \cdot \nabla^2 \phi \zeta \, dx. \quad (2.1.33)
\]

The second term in equation (2.1.33) is bounded by a constant multiplied by \( \varepsilon \) so in the limit will be zero. The first term however can be manipulated using integration by parts,

\[
\int_\Omega \sigma_\varepsilon \nabla u_\varepsilon \cdot \left( I + (\nabla_y \zeta)^T \right) \nabla_x \phi \, dx = -\int_\Omega u_\varepsilon \nabla \cdot \left( \sigma_\varepsilon \left( I + (\nabla_y \zeta)^T \right) \nabla_x \phi \right) \, dx, \quad (2.1.34)
\]

since \( u_\varepsilon = 0 \) on \( \partial \Omega \). Before we proceed with the calculations, we remark that the term being considered is dependent on both \( x \) and \( y \). Thus when calculating the divergence we must consider the total derivative representation of the divergence operator, that is

\[
\nabla \cdot \to \nabla_x \cdot + \varepsilon^{-1} \nabla_y \cdot .
\]

With this in mind we proceed,

\[
\nabla \cdot \left( \sigma_\varepsilon^T \left( I + (\nabla_y \zeta)^T \right) \nabla_x \phi \right) = (\nabla_x + \varepsilon^{-1} \nabla_y) \cdot \left( \sigma_\varepsilon^T \left( I + (\nabla_y \zeta)^T \right) \nabla_x \phi \right). \quad (2.1.35)
\]

The potentially troublesome looking term in equation (2.1.35) vanishes, that is,

\[
\varepsilon^{-1} \nabla_y \cdot \left( \sigma_\varepsilon^T \left( I + (\nabla_y \zeta)^T \right) \nabla_x \phi \right) = 0,
\]

since \( \zeta \) satisfies the dual cell problem. The other term in equation (2.1.35) weakly converges to its mean, by theorem 2.1.9, that is

\[
\nabla_x \cdot \left( \sigma_\varepsilon^T \left( I + (\nabla_y \zeta)^T \right) \nabla_x \phi \right) \rightharpoonup \nabla_x \cdot \left( \left[ \sigma_\varepsilon^T \left( I + (\nabla_y \zeta)^T \right) \right] y \nabla_x \phi \right) \quad \text{weakly as } \varepsilon \to 0.
\]

Thus with a final application of integration by parts we have

\[
\lim_{\varepsilon \to 0} \int_\Omega \sigma \left( \frac{x}{\varepsilon} \right) \nabla u_\varepsilon(x) \cdot \nabla v(x) \, dx = -\int_\Omega u_0 \nabla \cdot \left( \left[ \sigma_\varepsilon^T \left( I + (\nabla_y \zeta)^T \right) \right] y \nabla_x \phi \right) \, dx. \quad (2.1.36)
\]

By reviewing the relationship between the first order corrector, \( \chi \), and the solution of the dual cell problem, \( \zeta \), we aim to characterise the effective conductivity. Consider the
componentwise cell and dual cell problems,
\[
\nabla_y \cdot (\sigma_\varepsilon (\nabla_y \chi_j + e_i)) = 0 \\
\nabla_y \cdot (\sigma_\varepsilon^T (\nabla_y \zeta_i + e_i)) = 0.
\]

Multiplying the dual cell problem by $\chi_j$, and the cell problem by $\zeta_i$ and then integrating we have
\[
\int_{\mathcal{T}_N} \nabla_y \cdot (\sigma_\varepsilon^T (\nabla_y \zeta_i + e_i)) \chi_j dy = \int_{\mathcal{T}_N} \nabla_y \cdot ((\sigma_\varepsilon \nabla_y \chi_j + e_j)) \zeta_i dy,
\]
which by using integration by parts becomes
\[
\int_{\mathcal{T}_N} \sigma_\varepsilon^T (\nabla_y \zeta_i + e_i) \cdot \nabla_y \chi_j dy = \int_{\mathcal{T}_N} \sigma_\varepsilon (\nabla_y \chi_j + e_j) \cdot \nabla_y \zeta_i dy.
\]

Then cancelling like terms, and with some slight rearranging we have
\[
\int_{\mathcal{T}_N} \sigma_\varepsilon \nabla_y \chi_j \cdot e_i dy = \int_{\mathcal{T}_N} \sigma_\varepsilon^T \nabla_y \zeta_i \cdot e_j dy. \tag{2.1.37}
\]

By setting
\[
\Lambda = \left[ \sigma_\varepsilon^T \left( I + (\nabla_y \zeta_i)^T \right) \right]_{y},
\]
we see that elementwise we have
\[
\Lambda_{ij} = \left[ \sigma_\varepsilon^T (e_i + \nabla_y \zeta_i) \cdot e_j \right]_{y} = \left[ \sigma_\varepsilon^T e_i \cdot e_j + \sigma_\varepsilon^T \nabla_y \zeta_i \cdot e_j \right]_{y} = \sigma_\varepsilon (e_j + \nabla_y \chi_j) \cdot e_i = \sigma_{ji}^*, \tag{2.1.38}
\]
where we used the result of equation (2.1.37) and the element-wise definition of the effective conductivity given in equation (2.1.16). That is,
\[
\Lambda = (\sigma^*)^T,
\]
and following on from equation (2.1.36) we see that
\[
-\int_{\Omega} u_0 \nabla x \cdot \left( \left[ \sigma_\varepsilon^T \left( I + (\nabla_y \zeta_i)^T \right) \right]_{y} \nabla_x \phi \right) dx = -\int_{\Omega} u_0 \nabla x \cdot \left( (\sigma^*)^T \nabla_x \phi \right) dx
\]
\[
= \int_{\Omega} \sigma^* \nabla_x u_0 \cdot \nabla_x \phi dx. \tag{2.1.39}
\]

Hence we have shown that
\[
\lim_{\varepsilon \to 0} \int_{\Omega} \sigma \left( \frac{x}{\varepsilon} \right) \nabla u_\varepsilon (x) \cdot \nabla v (x) \ dx = \int_{\Omega} \sigma^* \nabla_x u_0 \cdot \nabla_x \phi dx,
\]
justifying the following theorem,

**Theorem 2.1.13.**
Suppose $f \in L^2(\Omega)$ and let $u_\varepsilon$ be the weak solution of
\[-\nabla \cdot (\sigma_\varepsilon \nabla u_\varepsilon) = f \quad \text{in } \Omega \]
\[u_\varepsilon = 0 \quad \text{on } \partial \Omega.\]

Furthermore let $u_0$ be the weak solution of
\[-\nabla \cdot (\sigma^* \nabla u_0) = f \quad \text{in } \Omega \]
\[u_0 = 0 \quad \text{on } \partial \Omega,\]

where $\sigma^* = [\sigma + \sigma (\nabla_y \chi)^T]_y$ and $\chi(y)$ is the weak solution of the cell problem, then as $\varepsilon \to 0$,
\[u_\varepsilon \rightharpoonup u_0 \quad \text{weakly in } H^1_0(\Omega).\]

We remark that strong convergence in $L^2(\Omega)$ can also be proven, see for example [43].

### 2.1.3 Examples

In this section we present a one dimensional and two dimensional example. We emphasise again that in dimensions greater than 1, no closed form generally exists for the homogenised conductivity. One of the few cases in which a closed form does exist is looked at as our two dimensional example.

**A One Dimensional Example**

For the one dimensional case consider the problem
\[-\frac{d}{dx} \left( \sigma_\varepsilon(x) \frac{d}{dx} u_\varepsilon \right) = f \quad \text{in } (0,1) \]
\[u_\varepsilon(0) = u_\varepsilon(1) = 0,\]

where $f = 1$ and
\[\sigma_\varepsilon = \left( 2 + \sin \left( \frac{2\pi x}{\varepsilon} \right) \right)^{-1}.\]

In the one dimensional setting the cell problem is
\[-\frac{d}{dy} \left( \sigma(y) \frac{d\chi(y)}{dy} \right) = \frac{d\sigma(y)}{dy} \quad \text{for } y \in (0,1)\]
\[\chi(0) = \chi(1) \]
\[\int_0^1 \chi(y) dy = 0,\]

and the homogenised coefficient is
\[\sigma^* = \int_0^1 \left( \sigma(y) + \sigma(y) \frac{d\chi(y)}{dy} \right) dy.\]
By examining equation (2.1.41) we see that

$$\sigma(y) \frac{d\chi(y)}{dy} = -\sigma(y) + c_1,$$

and thus

$$\sigma(y) + \sigma(y) \frac{d\chi(y)}{dy} = c_1,$$

which is precisely of the form of the homogenised conductivity, $\sigma^*$. By rearranging and integrating equation (2.1.42) we have

$$\chi(y) = -y + c_1 \int_0^y \frac{1}{\sigma(z)} \, dy + c_2.$$

Then by using the periodicity of $\chi(y)$ and the fact that $\chi$ is zero-meaned we find

$$c_1 = \left( \int_0^1 \frac{1}{\sigma(y)} \, dy \right)^{-1} = \sigma^*.$$

Hence for our example we have

$$\sigma^* = \left( \int_0^1 2 + \cos(2\pi y) \, dy \right)^{-1} = \frac{1}{2}.$$

The conductivity and resulting solutions are compared to the homogenised conductivity and homogenised solution in Figure 2.1.1 for $\varepsilon = 0.5$, $\varepsilon = 0.25$, and $\varepsilon = 0.125$.
We see that the true potential, \( u_\varepsilon \), appears to converge quickly to \( u_0 \) as \( \varepsilon \) goes to 0., showing the effectiveness of the homogenisation procedure.

### A Two Dimensional Example

As in the one dimensional case we first formulate our problem,

\[
- \nabla \cdot (\sigma_\varepsilon(x) \nabla u_\varepsilon) = f \quad \text{in } \Omega \\
u_\varepsilon = 0 \quad \text{on } \partial \Omega,
\]

(2.1.43)

with \( \bar{\Omega} = [0, 1]^2 \). We take \( \sigma_\varepsilon \) to be a two-phase checkerboard, see Figure 2.1.2, defined by

\[
\sigma_\varepsilon(x) = \begin{cases} 
\sigma_1, & \{x, \{x_1, x_2 \leq \frac{\varepsilon}{2}\} \cup \{x_1, x_2 > \frac{\varepsilon}{2}\}\} \\
\sigma_2, & \{x, \{x_1 \leq \frac{\varepsilon}{2}\} \cap \{x_2 > \frac{\varepsilon}{2}\}\} \cup \{\{x_2 \leq \frac{\varepsilon}{2}\} \cap \{x_1 > \frac{\varepsilon}{2}\}\}\}
\end{cases}
\]

with \( \sigma_\varepsilon \) being \( \varepsilon \)-periodic. In this special case, \( \sigma_\varepsilon \) is invariant under rotations by \( \frac{\pi}{2} \) and moreover \( \sigma_\varepsilon(x)\sigma_\varepsilon(R_{\frac{\pi}{2}}x) = cI \), where \( R_{\frac{\pi}{2}} \) denotes rotation by \( \frac{\pi}{2} \). Then using the Helmholtz orthogonal decomposition of \( L^2 \) into spaces of solenoidal and potential functions [53], and an alternative approach to finding \( \sigma^* \) [38], one finds

\[
\sigma^* = \sqrt{\sigma_1 \sigma_2},
\]

that is the homogenized conductivity is the geometric mean of the periodic conductivity. For an in-depth analysis of this case see for example [38].

We take \( \sigma_1 = 1 \) and \( \sigma_2 = 0.3 \), implying that \( \sigma^* = \sqrt{0.3} \). As in the one dimensional case the conductivity and resulting solutions, shown in Figure 2.1.2, are compared to the homogenised solution in Figure 2.1.2 for \( \varepsilon = 1, \varepsilon = 0.2 \), and \( \varepsilon = 0.04 \).

### 2.1.4 Corrector Theory

We now briefly investigate the rate of convergence for our original two-scale asymptotic expansion. There is a fair amount of literature concerned with this problem, from the late 60’s through to the present, and the ideas therein often fall under the umbrella title of corrector theory (for homogenisation) [54, 55, 56, 57, 58]. The necessity of these results will become apparent in the chapter 3.

We begin by stating that the two-scale asymptotic expansion does not converge at the naturally assumed rate by way of the following theorem,

**Theorem 2.1.14.**

Let \( u_\varepsilon \) and \( u_0 \) be solutions of (2.1.26) and (2.1.15) respectively, and suppose \( u_0 \in W^{2,\infty}(\Omega) \). Then

\[
\|u_\varepsilon - (u_0 + \varepsilon \chi \cdot \nabla u_0)\|_{H^1(\Omega)} \leq C\sqrt{\varepsilon},
\]

where \( \chi \) is the solution of the cell problem, and \( C \) is independent of \( \varepsilon \).
Figure 2.1.2: The two dimensional homogenisation **Top Row:** Conductivities $\sigma_\varepsilon(x)$ for decreasing values of $\varepsilon$ (left to right) **Centre Row:** The corresponding solutions $u_\varepsilon(x)$ **Bottom:** homogenised solution.

The proof of this theorem is fairly standard, and among various other places can be found in [34]. The above estimate is independent of $\hat{u}_1(x)$, since $\epsilon \hat{u}_1$ is $O(\sqrt{\varepsilon})$. The reason for the slower than expected convergence is due to the appearance of boundary layers, see for example [55].

To improve our two-scale asymptotic expansion we should introduce a *boundary layer corrector*, $\theta$, such that $\theta$ satisfies

$$-
abla \cdot (\sigma_\varepsilon \nabla \theta) = 0 \quad \text{in } \Omega$$
$$\theta = -u_1(x,y) \quad \text{on } \partial \Omega.$$  \hspace{1cm} (2.1.44)

Such a corrector accounts for the errors up to order $\varepsilon$. With the addition of the boundary layer corrector we have improved convergence.

**Theorem 2.1.15.**
Let \(u_\varepsilon\) and \(u_0\) be as in theorem 2.1.14, and \(\theta\) the solution of problem (2.1.44), then
\[
\|u_\varepsilon - (u_0 + \varepsilon \chi \cdot \nabla u_0 + \varepsilon \theta)\|_{H^1_0(\Omega)} \leq C\varepsilon,
\]
where \(C\) is independent of \(\varepsilon\).

**Proof.** Let \(u_\varepsilon\) and \(u_0\) be the unique solutions of (2.0.2) and (2.1.15) respectively. Furthermore, assume \(u \in W^{2,\infty}(\Omega)\), \(u_1\) and \(\theta\) are defined by (2.1.11) and (2.1.44) respectively, and let \(\hat{u}_1 = 0\) justified by the previous argument based on the order of \(\hat{u}_1\).

Define now the scaled remainder,
\[
r_\varepsilon(x) = \frac{1}{\varepsilon} \left( u_\varepsilon(x) - u_0(x) - \varepsilon u_1(x) - \varepsilon \theta(x) \right),
\]
which satisfies
\[
-\nabla \cdot (\sigma_\varepsilon \nabla r_\varepsilon) = \frac{1}{\varepsilon} \left( f + \nabla \cdot (\sigma_\varepsilon \nabla u_0) \right) + \nabla \cdot (\sigma_\varepsilon \nabla u_1) \quad \text{in } \Omega
\]
\[
r_\varepsilon = 0 \quad \text{on } \partial \Omega.
\]

Taking into account the set of cascading equations in section 2.1.1 and using the fact that
\[
\nabla = \nabla_y + \frac{1}{\varepsilon} \nabla_x,
\]
then for any \(\phi \in H^1_0(\Omega)\), we have
\[
\left| -\int_\Omega \nabla \cdot (\sigma_\varepsilon \nabla r_\varepsilon) \, dx \right| = \left| \int_\Omega \left( \frac{1}{\varepsilon} \left( f + \nabla \cdot (\sigma_\varepsilon \nabla u_0) \right) + \nabla \cdot (\sigma_\varepsilon \nabla u_1) \right) \phi \, dx \right|
\]
\[
= \left| \int_\Omega \left( \frac{1}{\varepsilon} \nabla_y \cdot (\sigma_\varepsilon \nabla u_2) + \nabla_x \cdot (\sigma_\varepsilon \nabla u_1) \right) \phi \, dx \right|
\]
\[
\leq \left| \int_\Omega - \left( \nabla_x \cdot (\sigma_\varepsilon \nabla y u_2) + \frac{1}{\varepsilon} \nabla_y \cdot (\sigma_\varepsilon \nabla y u_2) \right) \phi \, dx \right|
\]
\[
+ \left| \int_\Omega \left( \nabla_x \cdot (\sigma_\varepsilon \nabla y u_2) + \nabla_x \cdot (\sigma_\varepsilon \nabla x u_1) \right) \phi \, dx \right|
\]
\[
\leq \int_\Omega \sigma_\varepsilon \nabla_y u_2 \cdot \nabla \phi \, dx + C \|\phi\|_{H^1_0(\Omega)} \leq C \|\phi\|_{H^1_0(\Omega)}.
\]

The fact that (2.1.47) holds for all \(\phi \in H^1_0(\Omega)\) implies that, in particular, it holds for the supremum over all \(\|\nabla \phi\|_{H^1_0(\Omega)} = 1\), which gives us
\[
\|r_\varepsilon(x)\|_{H^1_0(\Omega)} \leq \frac{1}{C_m} \left\| \frac{1}{\varepsilon} \left( f + \nabla \cdot (\sigma_\varepsilon \nabla u) \right) \right\|_{H^{-1}(\Omega)} \leq C,
\]
which concludes the proof.

**Remark 2.1.16.** For \(f \in L^2(\Omega)\) and \(\Omega\) a convex polygon we can only be sure that \(u_0 \in H^2(\Omega) \supset W^{2,\infty}(\Omega)\). The extra regularity requirements are due to the terms involving \(\nabla_x u_2\). However there are two ways of deferring this problem, both of which enforce stronger conditions:

- We could recast the problem imposing stronger regularity assumptions on \(f\),
• It has been shown in [57] that if each component of $\chi(y)$ is in $W^{1,\infty}(\mathbb{T})$ then for $u_0 \in H^2(\Omega)$ we still have

$$\|u_\varepsilon - (u_0 + \varepsilon \chi \cdot \nabla u_0 + \varepsilon \theta)\|_{H^0(\Omega)} \leq C\varepsilon,$$

where $C$ is independent of $\varepsilon$. However, the elements of $\chi(y)$ are generally in $H^1(\Omega) \supset W^{1,\infty}(\mathbb{T})$ and since the right hand side of the cell problem is dependent only on $\sigma$, to achieve $\chi(y)$ elements in $W^{1,\infty}(\mathbb{T})$ we would require stricter regularity on $\sigma$, such as being bounded and Hölder continuous [59].

Typically $\theta$ cannot be solved for analytically. However, in chapter 3 we will only be concerned with the behaviour of $\theta$ rather than a full characterisation. In particular, a fundamental trait of interest is the fact that $\nabla \theta$ tends towards 0 exponentially fast as we move away from the boundary, hence the term boundary corrector. More formally, there exists $\gamma > 0$ such that $\nabla \theta$ decays like $e^{-\gamma z_k}$ where $z_k$ denotes the coordinate in the inward normal direction from the $k$th boundary [60].

This result depends on the geometry of $\Omega$ and the regularity of $u_0$ and is still a topic of research. The analysis is advanced, and no attempt to reproduce a proof is given here. We note only that the result holds for convex polygons\(^3\), see for example [61] or [62]. From this result it is deduced that the boundary layer has width of order $\varepsilon$ [60].

### 2.2 The Non-periodic Setting

There is a large amount of literature devoted to the study of deterministic non-periodic homogenisation, leading to several key concepts. In particular several ideas of convergence grew out of the study of non-periodic deterministic homogenisation which have become applicable elsewhere: De Giorgi’s theory of $\Gamma$-convergence [63] [64]; Spagnolo’s theory of $G$-convergence [65] [66]; and Tartar and Murat’s theory of $H$-convergence [52]. Moreover, Tartar’s method of oscillating test functions is adaptable to the non-periodic case.

Although the theory developed to cope with the deterministic non-periodic setting is rich, it is somewhat detached from reality, and is of little use for our purposes. Without attaching any statistics to the medium, and without the assumption of periodicity it is hard to say what is the physical meaning of $\varepsilon$ tending to zero.

Consider the following example: Suppose we have an isotropic composite material consisting of just two ingredients, $A$ and $B$, in equal amounts, that is $|\Omega_A| = |\Omega_B|$ and $|\Omega_A| + |\Omega_B| = |\Omega|$. Furthermore, suppose $A$ and $B$ have conductivities $\sigma_A$ and $\sigma_B$ respectively. Due to the absence of any statistics associated with the medium, there is an infinite number of effective conductivities which are possible. Figure 2.2.1 shows several possibilities for the effective conductivity along with the respective solutions which are clearly quite different.

It can be shown using the aforementioned types of convergence that there does exist an effective conductivity which only varies on the large scale. However, we have shown that

\(^3\)In chapter 3 we will be interested in triangles in particular due to our discretisation of our domain.
without any statistics or the assumption of periodicity, it is hard to give a physically motivated argument as to why the conductivity would limit to a homogenised counterpart. Our aim is not to undermine any of the work carried out in the non-periodic setting, it is only meant to illustrate that we will be unable to use the results in our application of what follows in this thesis.

2.3 The Stochastic Setting

We end the chapter on homogenisation with a brief discussion of stochastic homogenisation. Throughout this section we take $(\Sigma, \mathcal{F}, \mathbb{P})$ to be a probability space, where $\Sigma$ is the sample space, $\mathcal{F}$ denotes the $\sigma$-algebra, and $\mathbb{P}$ denotes the probability measure. Furthermore, let realisations be denoted by $\omega \in \Sigma$. We also require the following definition,

Definition 2.3.1. For $f \in L^1(\Sigma)$, we denote by

$$\mathbb{E}(f) = \int_{\Sigma} f \, d\mathbb{P}$$

the expected value of $f$.

In the stochastic setting, the conductivity is taken to be random, so that $\sigma = \sigma(x, \omega)$, with

$$\sigma : \Omega \times \Sigma \to \mathbb{R}_+$$

We also assume the random field $\sigma$ to be statistically homogeneous, or stationary, which can be interpreted as a periodic field with arbitrary period [67]. Furthermore we assume that

$$\sigma \left( \frac{x}{\varepsilon}, \omega \right) = \sigma \left( T_{x/\varepsilon} \omega \right) = \sigma \left( T_y \omega \right)$$

where $T$ is an ergodic dynamical system $T_y : \Sigma \to \Sigma$. There is an abundance of literature on
ergodic theory and ergodic dynamical systems, see for example, [68]. However, for a succinct
set of notes on the required ideas for stochastic homogenisation see [69] or [38].

To establish the main result we first present the stochastic version of the cell problem.

For any \( p \in \mathbb{R}^N \)
\[
-\nabla_y \cdot \left( \sigma \left( \nabla_y \chi_p \right)^T \right) = \nabla_y \cdot (\sigma p), \quad \text{a.s. in } \mathbb{R}^N,
\]

where \( \nabla \chi_p \) is stationary and \( \mathbb{E} \left( \int_{\mathbb{T}^N} \nabla \chi \, dy \right) = 0 \).

It has been shown that the abstract form of this problem in general cannot be solved for \( \chi \), however it can be solved for \( \nabla \chi \) [70], which is in fact all we will require. The other main
difference in the form of the cell problem between the periodic and stochastic setting is that
the cell problem in the stochastic setting is in fact posed over all of \( \mathbb{R}^N \), and hence solving it
numerically would be more expensive than solving the original problem. To this end various
approximation techniques have been proposed, such as periodisation, and implementation of
fictitious domain techniques and associated boundary conditions. For an in-depth look at the
stochastic cell problem see [70], [71] or [72] for example.

The main result in this setting is the following:

**Theorem 2.3.2.**

Let \((\Sigma, \mathcal{F}, \mathbb{P})\) be a probability space, with \( \mathcal{F}_x : \Sigma \rightarrow \Sigma \) an ergodic dynamical system. Suppose
\( f \in L^2(\Omega) \) and \( \sigma \) is bounded away from 0, with
\[
\sigma \left( \frac{x}{\epsilon}, \omega \right) = \sigma \left( \mathcal{F}_{x/\epsilon \omega} \right) = \sigma \omega.
\]

Then if we let \( u_\epsilon \) be the weak solution of
\[
-\nabla \cdot (\sigma \omega \nabla u_\epsilon) = f \quad \text{a.s. in } \Omega, \\
u_\epsilon = 0 \quad \text{on } \partial \Omega,
\]

and \( u_0 \) be the weak solution of
\[
-\nabla \cdot (\sigma^* \nabla u_0) = f \quad \text{in } \Omega, \\
u_0 = 0 \quad \text{on } \partial \Omega,
\]

where
\[
\sigma^*_{ij} = \mathbb{E} \left( \int_{\mathbb{T}^d} (e_i + \nabla \chi_{\epsilon \omega})^T \sigma \omega e_j \, dy \right)
\]
and \( \nabla \chi \) comes from the stochastic cell problem, then as \( \epsilon \rightarrow 0 \),
\[
u_\epsilon \rightharpoonup u_0 \quad \text{weakly in } H^1(\Omega).
\]

The proof of this theorem is based on three key ideas, namely a (the) div-curl lemma,
Weyl’s decomposition of \( L^2 \) into solenoidal and potential vector fields and Birkhoff’s ergodic
Chapter 2. Homogenisation Theory

Theorem. The proof can be found in multiple texts on stochastic/random homogenisation, for example see [69].

The stochastic version of the homogenisation procedure may be helpful for our end goal. However, it is unclear how the effective conductivity in this case would behave if the exact probability distribution of $\sigma_\varepsilon$ is unknown.

2.4 Discussion

We have given a short overview of homogenisation, showing that in some cases we can replace $\sigma_\varepsilon(x, \frac{x}{\varepsilon})$ by a specifically chosen slowly varying counterpart. For the periodic and stochastic cases we have given characterisation of the effective conductivity. We have also briefly discussed some issues around how well this approximation converges to the true solution.

When working under the assumption of periodicity there are a great deal of insightful results, however, periodicity is not always a warranted assumption. The deterministic non-periodic setting offers few results which are applicable to real world problems. In the stochastic case the resulting stochastic cell problem is computationally more expensive to solve than the original problem, and for $\varepsilon > 0$ it is hard to say much about the convergence.

A more physically motivated example is the case in which $\varepsilon$ is strictly greater than 0 but its exact value is in fact unknown. Such a problem is hard to deal with in the homogenisation framework: in the periodic case, we would have to solve for the boundary corrector, which can be computationally as expense as the original problem; while in the stochastic setting, it is unclear whether or not such a problem can be addressed. These difficulties lead us to consider the computational counterpart to homogenisation theory, so called multiscale numerical methods.
CHAPTER 3

MULTISCALE FINITE ELEMENT METHODS

The main shortcoming of homogenisation theory is the difficulty in applying the theory when $\varepsilon$ does not tend towards 0. To this end we will generally be required to use a numeric solver. Solving multiscale partial differential equations (PDE) numerically by traditional methods, such as finite element methods (FEM), can be extremely costly, even with the advent of supercomputers. Smaller scales must be resolved over the range of the larger scales, leading to the construction of a prohibitively large linear system which must be solved. Developing a numerical method which can capture the effects of the small scale behaviour while only requiring resolution of the large scale would go a long way to remedying this problem. Numerical schemes with such an idea at their heart have become known as multiscale numerical methods.

The first major advances in multiscale methods were in the early 80’s by Babuška and Osborn [17] who developed the idea of generalised finite element method (GFEM). Since the 90’s there has been continued growth in the field, resulting in numerous methods. Some of the most widely used algorithms are:

- Residual free bubble (RFB) methods of Brezzi and Russo [73];
- Multiscale finite element methods (MsFEM) developed by Hou and Wu [16];
- Heterogeneous multiscale methods (HMM) of E and Engquist [74].

This is far from an exhaustive list, there is a large number of methods which have been developed and are used in practice. A noteworthy and theoretically valuable tool which has also arisen out of the study of such problems is the variational multiscale methods (VMM) which are generally attributed to Hughes [75], although some similar ideas were present in an earlier paper by Hughes and Franca [76].
We will limit our discussion to multiscale finite element methods (MsFEM) which in some sense is a further development of GFEM. The first work on MsFEM was that of Hou and Wu in the late 90’s. Since then many others have made considerable improvements to the method, while the work of Efendiev and Hou [77] now provides a thorough introduction and analysis of the method.

In this chapter we first introduce some ideas behind FEM, and give some notes on implementation. The remainder of the chapter is devoted to MsFEM, where we outline the concept, show how we can implement the method, give some illustrative examples and briefly discuss the cost of the method. We end the chapter with a proof on the convergence of the method.

### 3.1 Finite Element Methods

The finite element method (FEM), or finite element methods, are often attributed to the works of Hrennikoff [78], Courant [79], and Bubușka [80, 81, 82]. These methods were originally conceived to deal with complex elasticity and structural and aeronautical engineering problems. Nowadays however, FEM is considered one of the go to method for solving many classes of problems involving partial differential equations [83].

We present a brief outline of FEM, a more thorough exploration can be found in a large number of texts, see any of [84, 85, 86] for example. We will restrict ourselves to the study of elliptic boundary value problems only, but note that some of the advantages of FEM are that it is not limited by time dependence or nonlinearities. To initiate our summary we first make clear our objective: we wish to solve the following problem

\[
-\nabla \cdot (\sigma(x) \nabla u) = f \quad \text{in } \Omega
\]

\[
u = g_D \quad \text{on } \Gamma_D
\]

\[
\sigma \nabla u \cdot \eta = g_N \quad \text{on } \Gamma_N,
\]

where \( \Omega \) is an open bounded domain with Lipschitz boundary \( \partial \Omega, \Gamma_D \cup \Gamma_N = \partial \Omega \) and \( \Gamma_D \cap \Gamma_N = \emptyset \). We take \( \eta \) as an outward facing unit vector normal to \( \partial \Omega \). Furthermore, \( f, g_N \in L^2(\Omega) \), and \( g_D \in H^{\frac{1}{2}}(\Gamma_D) \), see appendix A for an explanation of this space. We will in fact only be interested in the distinct cases

1. \( \Gamma_D = \partial \Omega \), and \( g_D = 0 \), that is the homogeneous Dirichlet problem, and

2. \( \Gamma_N = \partial \Omega \), that is the Neumann problem.

We will first discuss the Dirichlet problem, before moving on to the Neumann problem.

#### 3.1.1 The Dirichlet Problem

In both the Dirichlet and Neumann cases, the next step is to formulate the variational equivalent of problem (3.1.1), which is done using integration by parts. To this end, suppose
\( u \in C^2(\Omega) \cap C^1(\bar{\Omega}) \) and take \( v \in C_0^\infty(\Omega) \), then
\[
- \int_{\Omega} \nabla \cdot (\sigma \nabla u) \, v \, dx = \int_{\Omega} f \, v \, dx, \tag{3.1.2}
\]
then integrating by parts we have
\[
\int_{\Omega} (\sigma \nabla u) \cdot \nabla v \, dx = \int_{\Omega} f \, v \, dx. \tag{3.1.3}
\]
Here we make the observation that we in fact only require \( u, v \in H^1(\Omega) \) for equation (3.1.3) to make sense, moreover \( C_0^\infty(\Omega) \) is dense in \( H_0^1(\Omega) \), and thus we only need (3.1.3) to hold for all \( v \in H_0^1(\Omega) \). We also enforce the boundary conditions, which is done by imposing that \( u \) must lie in \( H_0^1(\Omega) \). Thus the weak formulation of the Dirichlet problem is
\[
\text{find } u \in \mathcal{V} \text{ such that } \int_{\Omega} (\sigma \nabla u) \cdot \nabla v \, dx = \int_{\Omega} f \, v \, dx \quad \forall v \in \mathcal{V}, \tag{3.1.4}
\]
where \( \mathcal{V} = H_0^1(\Omega) \).

### 3.1.2 The Neumann Problem

We now consider constructing the variational form for the Neumann problem. Again, suppose \( u \in C^2(\Omega) \cap C^1(\bar{\Omega}) \) but this time we take \( v \in C^\infty(\Omega) \), then using integration by parts and the Neumann boundary condition we have
\[
\int_{\Omega} (\sigma \nabla u) \cdot \nabla v \, dx - \int_{\partial \Omega} g_N v \, ds = \int_{\Omega} f \, v \, dx. \tag{3.1.5}
\]
Taking \( v \equiv 1 \) we arrive at the compatibility condition,
\[
\int_{\partial \Omega} g_N v \, ds + \int_{\Omega} f \, v \, dx = 0,
\]
which must be satisfied for (3.1.5) to have a solution. Again, we can relax our assumptions on \( u \) and \( v \), and demand only that \( u, v \in H^1(\Omega) \). However, as it stands, if \( u \) is a solution of (3.1.5), then so is \( u + c \), where \( c \) is any constant. To remedy this problem we restrict your space of possible solutions to those in \( H^1(\Omega) \) with
\[
\int_{\Omega} u \, dx = 0.
\]
Thus the weak formulation of the Neumann problem is
\[
\text{find } u \in \mathcal{V} \text{ such that } \int_{\Omega} (\sigma \nabla u) \cdot \nabla v \, dx = \int_{\Omega} f \, v \, dx + \int_{\partial \Omega} g_N v \, ds \quad \forall v \in \mathcal{V}, \tag{3.1.6}
\]
where \( \mathcal{V} = \{ v \in H^1(\Omega) : \int_{\Omega} v \, dx = 0 \} \).

Notice that in both the Dirichlet problem given by (3.1.4), and the Neumann problem of
(3.1.6) can be written in the form

\[
\text{find } u \in \mathcal{V} \text{ such that } B(u, v) = F(v) \quad \forall v \in \mathcal{V},
\]  

(3.1.7)

where \( B(u, v) \) is the left hand side of either (3.1.4) or (3.1.6), and is known as the bilinear form, while \( F(v) \) is the linear functional on the right hand side of (3.1.4) or (3.1.6) respectively.

### 3.1.3 The Ritz-Galerkin Approximation

The next step in finite element methods, whether it be the Dirichlet problem or Neumann problem, is to take a finite dimensional subspace, of \( \mathcal{V} \), say \( \mathcal{V}_h \) with a basis of functions, say \( \{\phi_1(x), \phi_2(x), \cdots, \phi_N(x)\} \). We mention that the space in which our solutions lie is termed the solution space and the space in which the functions \( v \) which we test against lie is known as the test or trial space, and they need not be the same in general.

Given \( \{\phi_1, \phi_2, \cdots, \phi_N\} \) forms a basis for \( \mathcal{V}_h \) we write the approximation

\[
u(x) \approx u^h(x) = \sum_{j=1}^{N} \lambda_j \phi_j(x)
\]  

(3.1.8)

One should (especially when dealing with inverse problems) make a finite dimensional approximation of \( \sigma \). In the current finite element setting we will limit ourselves to the case of letting

\[
sigma(x) \approx \sigma^h(x) = \sum_{\ell=1}^{N} \xi_\ell \phi_\ell(x).
\]  

(3.1.9)

However, examples exist where \( \sigma \) is represented on a distinct basis, see for example section 7.6 of [21].

### 3.1.4 The Linear System

By substituting the approximation back into the variational form, and testing for each function in the test space, we have

\[
\sum_{j=1}^{N} \lambda_j B(\phi_j, \phi_i) = F(\phi_i) \quad \text{for } i = 1, 2, \ldots N.
\]

In both the Dirichlet problem and Neumann problem we have

\[
B(\phi_j, \phi_i) = \int_{\Omega} \left( \sum_{\ell=1}^{N} \xi_\ell \phi_\ell(x) \right) \nabla \phi_j \cdot \nabla \phi_i \, dx = K_{ij},
\]  

(3.1.10)
with the matrix $K$ generally called the \textit{stiffness matrix}. As for the right hand side, in the Dirichlet problem we have

\[ F(\phi_i) = \int_{\Omega} f \phi_i \, dx = F_i, \quad (3.1.11) \]

and in the Neumann problem

\[ f(\phi_i) = \int_{\Omega} f \phi_i + \int_{\partial\Omega} g_N \phi_i \, dx = F_i \]

where the vector $F$ is known as the \textit{load} or \textit{forcing} vector. It is the coefficients of $u$, i.e. the $\lambda_i$’s, that we wish to solve for. In this manner, we now see that our finite dimensional approximation of (3.1.7) is simply linear system of the form

\[ K\lambda = F. \quad (3.1.12) \]

We have yet to declare what our basis functions $\phi_1, \phi_2, \cdots, \phi_N$ are. In this section we shall focus our discussion on continuous piece wise linear polynomials. Specifically, suppose we have a triangulation $T_h$ of our domain, then the basis functions we will typically be using throughout this thesis for finite element methods will be the set of basis functions defined by

\[ \phi_i(x_j) = \begin{cases} 
1, & i = j \\
0, & \text{otherwise}, 
\end{cases} \quad i, j = 1, 2, \ldots, N \]

where $x_j$ is the $j$th node of our triangulation, with $\phi_i$ piecewise linear. In this way we approximate $\xi_i = \sigma(x_i) = \sigma_i$ and $\lambda_i = u(x_i) = u_i$. Additionally, basis functions of this form have small support leading to an extremely sparse stiffness matrix and our solution is then

\[ u(x_i) = \lambda_i. \quad (3.1.13) \]

Higher order bases functions such as quadratic, cubic, etc. are also used in practice and are defined in a similar manner with the addition of defining edge (and potentially internal) nodes. For a general discussion on the choice of basis functions see for example [84]. We shall see, in section 3.2, we are not restricted to the use of only polynomial basis functions.

3.1.5 \textbf{Notes on Implementation}

There are two issues to do with implementation that require attention, and they both stem from the same obstacle: The stiffness matrix is singular.

- For the Dirichlet problem enforcing the associated boundary conditions solves this problem. To apply the boundary conditions we define $D$ as the set of indices for which $x_j$ are on $\partial\Omega$ for all $j$ in $D$, i.e. the boundary nodes. Then we can either:

1. Set $K_{ii} = 1$, $K_{ij} = 0$, and $F_i = 0$ for all $i \in D$ and for $j = 1, 2, \ldots, N$ with $i \neq j$,.
this resulting system in then uniquely solvable.

2. Annihilate $K_{ij}$, $K_{ji}$, and $F_i$ for all $i \in D$ and $j = 1, 2, \ldots, N$, forming the uniquely solvable system $\tilde{K}\tilde{\lambda} = \tilde{F}$, then reinstate $\lambda_i = 0$ for all $i \in D$.

The first method is more intuitive, while the second method can be implemented more efficiently since $\tilde{K} = \tilde{K}^T$. Moreover the first method can be seen as using the fact that the test functions are by definition 0 at the boundary.

• For the Neumann problem we must also enforce the constraint

$$\int_{\Omega} u \, dx = 0,$$

which in the finite dimensional setting becomes

$$\int_{\Omega} \sum_{j=1}^{N} \lambda_j \phi_j \, dx = \lambda \cdot C = 0, \quad \text{where} \quad C_j = \int_{\Omega} \phi_j \, dx.$$

We then interpret the problem as a constrained minimisation problem, where we wish to minimise the Lagrangian,

$$L(\lambda, \mu) = \frac{1}{2} \lambda^T K \lambda - \lambda^T F - \mu^T C \cdot \lambda,$$

where $\mu$ is taken as a Lagrange multiplier. By taking

$$\frac{\partial L}{\partial \lambda} = K \lambda - F + \mu C = 0 \quad \text{and} \quad \frac{\partial L}{\partial \mu} = C \cdot \lambda = 0$$

we arrive at the uniquely solvable linear system

$$\begin{bmatrix} K & C \\ C^T & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \mu \end{bmatrix} = \begin{bmatrix} F \\ 0 \end{bmatrix}. \quad (3.1.14)$$

It should be pointed out that $\mu$ is of no interest to us at all, and is discarded. For more on enforcing Neumann boundary conditions see [86] for example. For more on the link between variational forms and Lagrangians see for example [30].

We have thus been able to reduce both the Dirichlet and Neumann problems to a uniquely solvable (and sparse) linear system. This concludes our brief outline of finite element methods.

### 3.2 Multiscale Finite Element Methods

We will start our exploration of multiscale finite element methods by considering the Dirichlet problem,

$$- \nabla \cdot (\sigma(x) \nabla u_\varepsilon) = f \quad \text{in} \Omega$$

$$u_\varepsilon = 0 \quad \text{on} \partial \Omega. \quad (3.2.1)$$
where $\varepsilon$ is indicative of some small scale (not necessarily periodic) structure of the problem. The finite element methods developed previously were aimed to tackle such a problem. However, the number of basis functions needed to refine the small scale behaviour can quickly become prohibitive. This drawback lead to the advent of several methods aimed at solving multiscaled PDE with reduced cost, as mentioned in at the beginning of the chapter.

Multiscale finite element methods consist of two tasks:

1. Localised construction of so-called multiscale basis functions which inherit the multiscale structure of the problem.

2. a global coarse scale solver, which acts to couple the scales.

Firstly we discuss how one chooses a set of basis functions and some associated concepts, and then we examine how to go about coupling the scales.

### 3.3 Construction of Multiscale Basis Functions

Firstly, denote by $\mathcal{T}_H$ an admissible partition of $\Omega$ into a finite set of triangle\(^1\). We will call this partition the coarse grid or coarse mesh, which we will assume can be refined to give a finer grid which we will call the fine grid or fine mesh, and we denote by $\mathcal{T}_h$. There appears to be no set method in the literature for carrying out refinements for triangular elements when using MsFEM, however for this thesis we shall restrict ourselves the use of so-called red refinement [86].

Red refinement is perhaps the most natural way of refining triangular elements, it ensures that the quality of the grid is maintained since all smaller elements are congruent to their respective parent element, see Figure 3.3.1. For more on mesh refinement the reader is directed to [86] and the references therein.

\[\text{Figure 3.3.1: Red refinement.}\]

Let $x_i$ be the nodes of the coarse mesh, and let the finite element subspace $\mathcal{V}_h$ be spanned by the typical piecewise linear functions denoted by $\phi_i(x)$. Analysis for higher order elements is much the same, for an in-depth account of higher order MsFEM, see for example [87], or [88] among others. Furthermore let $S_i$ denote the support of $\phi_i(x)$. Then we define the

\(^1\)Quadrilaterals would also be fine
corresponding set of multiscale basis functions $\psi_i(x)$ as the solution of the following,

$$-\nabla \cdot (\sigma \nabla \psi_i) = 0 \quad \text{in } K,$$

$$\psi_i = \phi_i \quad \text{on } \partial K, \quad \forall K \in T_H, K \subset S_i,$$

where $K$ is taken to denote an element of our mesh. We take the finite element space spanned by $\psi_i$ as $W_H$,

$$W_H = \text{span}\{\psi_i\} \subset H_0^1(\Omega).$$

We make several remarks at this point. Firstly by construction we have $\text{supp}(\psi_i) = \text{supp}(\phi_i)$, that is the multiscale basis functions have small support. Secondly the boundary conditions imposed on the multiscale basis functions are important, and is a topic of ongoing research. Thirdly, we could take $K$ to be smaller than a coarse element, this technique is usually only applied when one has a priori knowledge of periodicity of the medium, or the location of heterogeneities, for more on this approach see [77].

In general, equation (3.3.1) is solved using standard FEM, hence the multiscale basis functions can be written,

$$\psi(x) = \sum_j \alpha_j \varphi_j(x).$$

In our case we take the $\varphi_j$’s to be the piecewise linear basis functions over $T_h$. Keeping this representation in mind will useful at the implementation stage, see section 3.6.

### 3.4 Coupling the Scales

By constructing a set of multiscale basis functions we can represent the fine-scale solution using a reduced number of degrees of freedom. This is done by making the standard finite
element projection onto our updated basis,

\[ u \approx u_{MS} = \sum_{i} u_i \psi_i(x) \] (3.4.1)

To attain a global solution we look to establish a suitable variational formulation which will reduce computational costs. To this end, there are two choices of test functions by which we could multiply equation (3.2.1) by, before making substitution (3.4.1):

1. One approach would be to implement a Galerkin finite element method, where the test space and the solution space are both \( \mathcal{W}_H \). This scheme results in the following variational form,

\[ \sum_{K \in T_H} \int_K \sigma \nabla u_{ms} \cdot \nabla v_H \, dx = \int_{\Omega} f v_H \, dx, \quad \forall v_H \in \mathcal{W}_H. \] (3.4.2)

2. On the other hand, we could implement a Petrov-Galerkin (P-G) finite element method. In this scheme we take \( \mathcal{V}_H \) as our test space and maintain \( \mathcal{W}_H \) as our solution space. This approach leads to the following variational form,

\[ \sum_{K \in T_H} \int_K \sigma \nabla u_{ms} \cdot \nabla v_H \, dx = \int_{\Omega} f v_H \, dx, \quad \forall v_H \in \mathcal{V}_H. \] (3.4.3)

There are benefits to employing the P-G finite element form of the problem:

- The implementation procedure is less complex and less expensive [89].
- With the addition of adapted boundary conditions, discussed in the following subsection, the P-G method in some situations has improved accuracy.

For these reasons we will limit ourselves to the P-G version of MsFEM.

In the above we have shown that the global formulation and coupling of the scales is problem dependent. In the present case simply modifying the stiffness matrix to a multiscale stiffness matrix is sufficient to couple the scales.

### 3.5 Basis Function Boundary Conditions

The choice of boundary conditions to be imposed upon multiscale basis functions has been, and still is, the topic of a lot of research. In general there are three classes of boundary conditions we could choose to impose along \( \partial K \). The need to investigate boundary conditions is mainly due to the appearance of so-called resonance error. This type of error is characterised by the appearance of large errors in \( u_{MS} \) when \( h \approx \varepsilon \), this will be made formal at the end of this chapter.

The cause of the resonance term is in fact related to the first order boundary layer corrector, \( \theta \), discussed in the previous chapter. This boundary layer corrector is necessary to
correct for the artificial boundary condition we impose on $u_{MS}$ by enforcing boundary conditions upon the multiscale basis functions. In an attempt to reduce such effects at least two concepts have been investigated.

### 3.5.1 Reduced problem methods

In this approach one solves a dimensionally reduced elliptic problem on either side of $\partial K$ to find values of $\psi|_{\partial K}$. Reduced problem methods are in general suited to cases where $\sigma_\epsilon$ is separable, i.e. $\sigma_\epsilon(x_1,x_2) = \sigma_1(x_1)\sigma_2(x_2)$, however for the general case an oversampling strategy may be more suited [16].

### 3.5.2 Oversampling methods

In the simple form of this regime one extends the computational domain of the basis function by $O(\epsilon)$ beyond $\partial K$ to say $\partial K^{OS}$ and then enforces linear boundary conditions. That is, we solve

\[
-\nabla \cdot (\sigma_\epsilon \nabla \psi_{OSi}) = 0 \quad \text{in} \quad K^{OS},
\]

\[
\psi_{OSi} = \phi_i \quad \text{on} \quad \partial K^{OS}, \quad \forall K^{OS}, K \in T_H, K \subset S_i.
\]

In the case that the extension of $K$ goes to beyond $\partial \Omega$ we enforce boundary conditions also along $\partial \Omega \cap \hat{K}$. We then form the actual multiscaled basis functions by taking a linear combination of the oversampled basis function,

\[
\psi_i = \sum_{j=1}^{d} c_{ij} \psi_{OSj}^{i},
\]

with the $c_{ij}$ chosen so that $\psi_j(x_i) = \delta_{ij}$, where the $x_i$ are the corner nodes of the coarse mesh [16]. More compactly this can be written as

\[
\Psi = C\Psi^{OS}.
\]

Thus if one takes care to order the basis functions and nodes accordingly we have $C = \hat{\Psi}^{-1}$, where $\hat{\Psi}_{ij} = \psi_{OSj}^{i}(x_i)$.

The oversampling method is motivated by the idea that first order element boundary layer corrector only decays exponentially fast, and thus only influences regions within $O(\epsilon)$ of $\partial K$, as discussed in section 2.1.4. Results on convergence of these methods is given in section 3.9. It should be noted that there are several variants of this method, and research is ongoing, see [90] for an in-depth look at several different oversampling strategies.

Examples of multiscale basis functions are shown in Figure 3.5.1 for the standard and oversampled framework. It should be made clear that oversampling in general leads to basis functions which can be discontinuous across $\partial K$ resulting in what is known as a nonconforming method [84].

Table 3.5.1 summarises the different finite element basis functions used.
Figure 3.5.1: Multiscale basis functions Left: Conforming Right: Non-conforming.

Table 3.5.1: The different basis functions used in MsFEM.

<table>
<thead>
<tr>
<th>Basis functions</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi(x)$</td>
<td>Linear finite element (on coarse mesh)</td>
</tr>
<tr>
<td>$\varphi(x)$</td>
<td>Linear finite element (on fine mesh)</td>
</tr>
<tr>
<td>$\psi(x)$</td>
<td>Multiscale finite element solved for over $K$</td>
</tr>
<tr>
<td>$\psi^{OS}(x)$</td>
<td>Multiscale finite element solved for over $K^{OS}$</td>
</tr>
</tbody>
</table>

3.6 Implementation

As stated in section 3.2, MsFEM comprises of two main steps, construction of the basis functions and coupling of the scales. Constructing the basis functions is generally carried out using regular finite element methods as discussed in section 3.3. The task is not difficult, solving the Laplace’s equation using standard linear FEM is in fact a classroom problem, for efficient computations for construction of the basis functions see for example [86].

On the other hand, the task of coupling the scales is problem peculiar to MsFEM. We present here a straightforward method to compute the PG form of the multiscale stiffness matrix.

Initially we shall recall some notation. The linear finite element basis on the fine and coarse mesh are denoted by $\varphi$ and $\phi$ respectively, while the multiscale basis functions (on the coarse scale) are denoted $\psi$. We will use the notation $\mathcal{K}$ to represent the multiscale stiffness matrix, that is

\[
\mathcal{K}_{ij} = \int_{\Omega} \sigma \nabla \phi_i \nabla \psi_j \, dx = \sum_{K \in \mathcal{T}_h} \left( \int_{K} \sigma \nabla \phi_i \cdot \nabla \psi_j \, dx \right) \quad i, j = 1, 2, \ldots, N.
\]

We will demonstrate the construction of the matrix over a single element of the coarse mesh, as it is then trivial to construct the entire matrix.
In a given \( K \in \mathcal{T}_H \), \( \nabla \phi_i \) is constant, and thus we have
\[
\int_K \sigma \nabla \phi_i \cdot \nabla \psi_j \, dx = \nabla \phi_i \cdot \sum_{k \in \mathcal{T}_h \cap K} \left[ \int_k \sigma \nabla \psi_j \, dx \right].
\]

We will now simply write \( \sum_k \) in place of \( \sum_{k \in \mathcal{T}_h \cap K} \). By projecting \( \sigma \) onto the fine basis functions, and recalling that the multiscale basis functions are in fact linear combinations of the fine basis functions (see (3.3.2)) we have
\[
\int_K \sigma \nabla \phi_i \cdot \nabla \psi_j \, dx = \nabla \phi_i \cdot \sum_k \left[ \int_k \sigma (\sum_l \varphi_l) \left( \sum_m \psi_j^m \varphi_m \right) \, dx \right] = \nabla \phi_i \cdot \sum_k \left[ \int_k (\sum_l \varphi_l) \left( \sum_m \psi_j^m \nabla \varphi_m \right) \, dx \right] = \nabla \phi_i \cdot \sum_k \left[ \left( \sum_m \psi_j^m \nabla \varphi_m \right) \int_k (\sum_l \varphi_l) \, dx \right],
\]
where \( \psi_j^m = \psi_j(x_m) \) and we have utilised the fact that for each \( k \) we have \( \nabla \varphi \) constant.

The steps outlined here are for an arbitrary dimension, \( d \), and any coarse mesh, \( \mathcal{T}_H \). We note that in the one dimensional problem or problems posed on a regular coarse mesh some simplifications may be possible.

To continue, let
\[
\mathcal{N} = [\nabla \varphi_1, \nabla \varphi_2, \ldots, \nabla \varphi_{d+1}] \in \mathbb{R}^{d \times (d+1)} \quad \text{and} \quad \mathcal{M} = [\nabla \varphi_{m_1}, \nabla \varphi_{m_2}, \ldots, \nabla \varphi_{m_{d+1}}] \in \mathbb{R}^{d \times (d+1)},
\]
where the \( \varphi_n \)'s for \( n = 1, 2, \ldots, d+1 \) are the linear finite element basis functions sharing support in \( K \), and the \( \varphi_{m_s} \)'s for \( s = 1, 2, \ldots, d+1 \) are the \( d+1 \) linear finite element basis functions sharing support in a small element, \( k \).

Then for all \( i \) and \( j \) such that \( \phi_i \) and \( \psi_j \) share support in \( K \), we have
\[
K_{ij} = \int_K \sigma \nabla \phi_i \cdot \nabla \psi_j \, dx = \mathcal{N}^T \sum_k \left[ \mathcal{M} \Psi \sum_{r=1}^{d+1} \sigma_{m_r} \int_k \varphi_{m_r} \, dx \right] = \mathcal{N}^T \sum_k \left[ \mathcal{M} \Psi \frac{1}{6} J_k \sum_{r=1}^{d+1} \sigma_{m_r} \right],
\]
where \( \Psi \in \mathbb{R}^{(d+1) \times (d+1)} \) with \( \Psi_{ps} = \psi_{i_s}^m \psi_{i_p} \) \( (x_{m_p}) \), \( J_k \) denotes the Jacobian of the mapping of \( k \) to the local triangle, while the volume of \( \psi \) after changing to local variables is \( \frac{1}{6} \).

Finally, with a little linear algebra (3.6.1) can be further simplified to give
\[
\int_K \sigma \nabla \phi_i \cdot \nabla \psi_j \, dx = \mathcal{N}^T \mathcal{M} D \mathcal{Q},
\]
where \( \mathcal{M} \) is a block column matrix, \( D \) is a block diagonal matrix, and \( \mathcal{Q} \) is a block row matrix,
Chapter 3. Multiscale Finite Element Methods

with

\[ M_j = M_j, \quad D_{ii} = \left( \frac{|J_k|}{6} \sum_{r=1}^{d+1} \sigma_{i_r} \right) I_d, \quad \text{and} \quad Q_i = \Psi_i, \]

where \( i \) and \( j \) run from 1 through to \( N_k \), where \( N_k \) is the number of fine elements in a coarse element. In this way we have \( M \in \mathbb{R}^{d \times (d+1)N_k}, \ D \in \mathbb{R}^{(d+1)N_k \times 3(d+1)}, \) and \( Q \in \mathbb{R}^{(d+1)N_k \times (d+1)}, \) and hence the product of (3.6.2) lies in \( \mathbb{R}^{(d+1) \times (d+1)} \) as required. We note two important but perhaps not obvious facts: Firstly, The matrices \( N, M, D, \) and \( Q \) are in general different for each coarse element, \( K; \) and secondly, for efficiency the product \( DQ \) is computed using \( D \) as a block column matrix and taking the Hadamard product.

The final step in the implementation to solve the system

\[ \mathcal{K} \lambda = F \]

where is \( F \) combines boundary conditions and any forcing terms. Our solution is then

\[ u_H = B \lambda = B \mathcal{K}^{-1} F \]

where \( B = [\psi_1, \psi_2, \ldots, \psi_N] \).

3.7 Examples

Now that we have given an outline of the implementation of MsFEM we will carry out examples in both one and two dimensions.

3.7.1 A One Dimensional Example

Consider the problem

\[ -\frac{d}{dx} \left( \sigma_\varepsilon(x) \frac{du_\varepsilon}{dx} \right) = f \quad \text{in} \ (0, 1) \]

\[ u_\varepsilon(0) = u_\varepsilon(1) = 0, \]

where \( f = 1 \) and

\[ \sigma_\varepsilon(x) = \left( 1.01 \cos \left( \frac{2\pi x}{\varepsilon} \right) \right)^{-1} + 20 \exp \left( -\left( \frac{x - 0.7}{\varepsilon} \right)^2 \right), \]

with \( \varepsilon = 0.1 \), as shown in Figure 3.7.1.

In this example we used two different regular discretisations, in the first case \( H_1 = \frac{1}{7} \) with \( h_1 = \frac{1}{273} \), meaning that each coarse interval (element) contains 39 fine intervals, while in the second case we have \( H_2 = \frac{1}{15} \) with \( h_2 = \frac{1}{585} \), again meaning that each coarse interval contains 39 fine intervals.

The multiscale basis functions were solved using linear finite element methods, and are shown Figure 3.7.1 for each case.
Figure 3.7.1: One Dimensional MsFEM Left Multiscale conductivity, $\sigma_\varepsilon$ Centre: (top to bottom) Multiscale scale basis functions on $H_1$; multiscale scale basis functions on $H_2$ Right Column: (top to bottom) Comparison of MsFEM solution ($u_{MS}$), coarse FEM solution ($u_C$) and fine FEM solution ($u_F$) on $T_{H_1}$ and $T_{h_1}$; Same comparisons on $T_{H_2}$ and $T_{h_2}$.

For comparison we also show the approximate solution attained by using linear FEM on the coarse discretisation, denoted by $u_C$, and the approximate solution attained by using linear FEM on the fine discretisation, denoted by $u_F$. These solutions along with shown along with MsFEM solutions are given in Figure 3.7.1.

It is clear that MsFEM is far more accurate than the coarse scale FEM, and even when carried out with just 8 basis functions seems to do a reasonable job. By increasing the number of multiscale basis functions to 16 we see that MsFEM and the fine scale FEM are almost identical.

### 3.7.2 A Two Dimensional Example

We now present a two dimensional example. Consider the problem

$$
- \nabla \cdot (\sigma_\varepsilon(x) \nabla u_\varepsilon) = f \quad \text{in } \Omega \\
\quad u_\varepsilon = 0 \quad \text{on } \partial \Omega,
$$

where $f = 1$ and

$$\sigma_\varepsilon(x) = \exp \left( \cos \left( 2\pi \frac{r}{\varepsilon} \right) + \tan 2 \left( \frac{x}{y} \right) \right),$$

with $\varepsilon = 0.05$, as shown in Figure 3.7.2.

For this example, we used two separate coarse meshes, $T_{H_1}$ and $T_{H_2}$, comprised of 81 nodes and 128 elements, and 1,089 nodes and 2,048 elements respectively. The fine mesh, $T_h$, can be computed as either 4 (red) refinements of $T_{H_1}$ or 2 (red) refinements of $T_{H_2}$. In either case $T_h$ has 16,641 fine nodes and 32,768 fine elements.
Refining $\mathcal{T}_{H_1}$ results in each coarse element containing 153 fine nodes and 256 fine elements, on the other hand, refining $\mathcal{T}_{H_2}$, leads to each coarse element containing 15 fine nodes and 16 fine elements. The meshes $\mathcal{T}_{H_1}$ and $\mathcal{T}_{H_2}$ can be seen in Figure 3.7.3. Once again the multiscale basis functions were solved for using standard linear FEM on the fine mesh over the coarse elements making up the respective supports.

To compare results we solved the problem using FEM with six-node quadratic basis functions on the 16,641 (4,225 corner and 12,416 midpoint) nodes of $\mathcal{T}_h$, the solution of which we take to be accurate, and is shown in Figure 3.7.2. For comparisons we also show the results computed using MsFEM on both $\mathcal{T}_{H_1}$ and $\mathcal{T}_{H_2}$, FEM on $\mathcal{T}_h$, and FEM on both $\mathcal{T}_{H_1}$ and $\mathcal{T}_{H_2}$, all of which can be seen in Figure 3.7.4.

We end this example by pointing out two observations, which are apparent when looking at the results:

- Neither the MsFEM solutions nor the FEM solutions computed on the coarse meshes
reach the maximal values of both the quadratic FEM solution and the FEM solution computed on the fine mesh. We do see that by using $T_{H_1}$ rather than $T_{H_2}$ (i.e. increasing the number of basis functions (both coarse and multiscaled)), which is inline with standard finite element analysis results [84].

- The MsFEM solutions capture the multiscale ripples in the solution.

3.8 Cost

We will now look at analysing the cost of the conforming multiscale finite element method for the two dimensional case. A similar analysis is carried out in [91] using rectangular elements, and in [16]. We first formulate our problem,

$$
- \nabla \cdot (\sigma_\varepsilon(x) \nabla u_\varepsilon) = f \quad \text{in } \Omega \\
\quad u_\varepsilon = 0 \quad \text{on } \partial \Omega,
$$

with $\bar{\Omega} = [0,1]^2$. Let $T_H$ denote the coarse triangulation of $\Omega$.

We will assume for ease of the analysis that the mesh is regular enough so that the number of intervals in the $x$- and $y$-directions are approximately equal, and we denote this number by $L$. Then the total number of coarse elements, which we will denote by $K$, is of order $L^2$, that is $K = O(L^2)$. Moreover, we assume each coarse element is red refined so as to have $\ell$ fine nodes on each coarse edge (by red refining, $p$ is related to the number of refinements, $r$, by $\ell = 2^r + 1$, $r = 0, 1, \ldots$), then letting $m$ denote the number of fine elements in a coarse element we have $m = O(\ell^2)$. With this notation the total number of fine elements is of order
\[(L \ell)^2 = Mm.\]

The theoretical cost of solving the problem on a serial machine and on a parallelized machine using MsFEM, using linear finite elements on the fine scale (denoted fFEM), and using linear finite elements on the coarse scale (denoted cFEM) are presented in Table 3.8.1.

Table 3.8.1: Costs of the solving (3.8.1) using MsFEM, FEM on the fine scale (fFEM) and FEM on the coarse scale (cFEM).

<table>
<thead>
<tr>
<th>Cost</th>
<th>MsFEM</th>
<th>fFEM</th>
<th>cFEM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sequential</td>
<td>(O(M + m))</td>
<td>(O(Mm))</td>
<td>(O(M))</td>
</tr>
<tr>
<td>Parallel</td>
<td>(O\left(\frac{M}{P} + m\right))</td>
<td>(O\left(\frac{Mm}{P}\right))</td>
<td>(O\left(\frac{M}{P}\right))</td>
</tr>
</tbody>
</table>

It should be noted that construction of the stiffness matrix for linear FEM is very well understood and thus much easier to optimise compared to the multiscale stiffness matrix associated with MsFEM. The CPU time required for solving using FEM on the fine scale denoted fFEM and the time required using MsFEM are presented in Figure 3.8.1 for two separate cases:

- In the first case we vary the number of coarse nodes and red refine the mesh 3 times, meaning the number of fine nodes per element is held at 45, we denote this case by MsFEM1.
- In the second case we vary the number of coarse nodes and red refine the mesh 4 times, meaning the number of fine nodes per element is held at 153, we denote this case by MsFEM2.

Both examples were carried out on a 2013 Imac with a 3.2GHz Intel Core i5 processor with four independent cores, and 8Gb of memory. Parallelisation was implemented simply by using the parfor command in MATLAB.

We see that the time required for linear FEM appears linear in the total number of nodes. On the other hand the MsFEM algorithm appears to be faster than the fine FEM, but the time taken appears to be increasing at an order greater than linear. This is likely due to increased complexity of computing the multiscale stiffness matrix in comparison to the standard stiffness matrix. We could reduce the time by simply using a larger number of processors (in the order of hundreds), which is often the case in practice [16].

### 3.9 Proof of Convergence

To the end this chapter we will cover some results on convergence analysis of MsFEM for linear elliptic problems. The steps here provide a mathematical justification for use of MsFEM and offer insight into some of the associated problems of multiscale numerical methods in general. As with most multiscale methods analysis is carried out in the periodic setting so we can draw on several results from homogenisation theory, most of which are covered in chapter 2. Our analysis follows the approach given in [60].
Consider the problem
\[
-\nabla \cdot \left( \sigma_{\varepsilon} \left( \frac{x}{\varepsilon} \right) \nabla u_{\varepsilon} \right) = f \quad \text{in } \Omega \\
u_{\varepsilon} = 0 \quad \text{on } \partial \Omega,
\]
(3.9.1)
where \( \Omega \) is the unit square in \( \mathbb{R}^2 \) (i.e. \( \Omega \) is bounded with a convex Lipschitz boundary), and \( \varepsilon \ll 1 \). For our purposes here we will assume \( \sigma_{\varepsilon} \in W^{1,\infty}(T_{\varepsilon}) \), which is the set of \( \varepsilon \)-periodic functions in \( W^{1,\infty}(\Omega) \). We will also assume \( f \in L^2(\Omega) \), implying that the solution \( u_{\varepsilon} \) is in \( H^2(\Omega) \) [92].

We let \( T_H \) denote a triangulation of \( \Omega \) and note by \( K \) a single triangular element (bounded with a convex Lipschitz boundary), such that \( H^2 \approx |K|, H \approx |\partial K| \), and \( 0 < H \leq 1 \), that is \( H \) is the representative size of a coarse element. We will also at times make the substitution \( y = \frac{x}{\varepsilon} \), so that any \( \varepsilon \)-periodic function of \( x \) will become 1-periodic in \( y \).

Recall that \( u_{\varepsilon} \) can be expanded as
\[
u_{\varepsilon} = u_0(x) + \varepsilon u_1(x, y) + \ldots,
\]
with \( u_0 \) the solution of the homogenised problem,
\[
-\nabla \cdot (\sigma^* \nabla u_0) = f \quad \text{in } \Omega \\
u_0 = 0 \quad \text{on } \partial \Omega.
\]
(3.9.2)
The next term in the expansion satisfies
\[
u_1 = \chi(y) \cdot \nabla u_0,
\]
with
\[-\nabla_y \cdot \left( \sigma (\nabla_y \chi)^T \right) = \nabla_y \cdot \sigma \]
\[\chi : \mathbb{T}^2 \rightarrow \mathbb{R}^2 \]
\[\int_{\mathbb{T}^2} \chi \, dy = 0. \] (3.9.3)

We will suppose that the homogenised solution $u_0$ is in $H^2(\Omega) \cap W^{1,\infty}(\Omega)$. This assumption can be relaxed to $u_0 \in H^2(\Omega)$ but having $u_0$ also in $W^{1,\infty}(\Omega)$ enables us to use results which ease the analysis. We also point out that $\sigma_\varepsilon \in W^{1,\infty}$ implies that each component of $\chi(y)$ is in $W^{2,\infty}(\mathbb{T}) \cap C^{1,\alpha}(\bar{\mathbb{T}})$ where $\alpha = 1 - \frac{n}{p}$ and $C^{1,\alpha}(\Omega)$ is a Hölder space [93], similar results also appear in [50].

The analysis here follows a similar vein to that provided in [94] and [60]. However, a slightly more general setting is presented in [60], namely $u \in H^2(\Omega)$ and $\chi(y)$ is in $W^{2,p}(\mathbb{T}) \cap C^{1,\alpha}(\bar{\mathbb{T}})$ with $p > 2$ and $\alpha = 1 - \frac{2}{p}$.

To carry out our analysis we first define some necessary spaces. Initially we will define the standard finite element space of piecewise affine functions,
\[V_H = \text{span}\{\phi_i\} \subset H^1_0(\Omega) = V\]
where the $\phi_i$'s are piecewise linear polynomials. Now we define the set of multiscale basis functions,
\[W_H = \text{span}\{\psi_i\} \subset H^1_0(\Omega) = W,\]
where the multiscaled basis functions satisfy
\[-\nabla \cdot (\sigma_\varepsilon \nabla \psi_i) = 0 \quad \text{in } K, \]
\[\psi_i = \phi_i \quad \text{on } \partial K, \quad \forall K \in \mathcal{T}_H, K \subset \text{supp}(\phi_i). \] (3.9.4)

We will take $u_h$ to be the solution of either Galerkin form of the problem,
\[\sum_{K \in \mathcal{T}_H} \int_K \sigma \nabla u_h \cdot \nabla v_h \, dx = \int_\Omega fv_h \, dx, \quad \forall v_h \in W_H, \] (3.9.5)
or the solution of the PG form of the problem,
\[\sum_{K \in \mathcal{T}_H} \int_K \sigma \nabla u_h \cdot \nabla v_h \, dx = \int_\Omega fv_h \, dx, \quad \forall v_h \in V_H. \] (3.9.6)

The main result on convergence of the conforming multiscale finite element methods will be the following theorem:

**Theorem 3.9.1.** Let $u$ and $u^H$ be the solutions of (3.9.1) and (3.9.5) (or (3.9.6)) respectively. Then there exist constants $C_1$ and $C_2$, independent of $\varepsilon$ and $h$, such that
\[\|u - u^H\|_{H^1(\Omega)} \leq C_1 H \|f\|_{L^2(\Omega)} + C_2 \sqrt{\frac{\varepsilon}{H}} \] (3.9.7)
To prove this we will focus on the Galerkin form of the problem, but note the result is valid for the PG implementation. Several intermediate result will be needed, one of which was obtained by Cea in 1964 [95].

**Theorem 3.9.2 (Cea’s Lemma).** Let $u$ and $u_H$ be the solutions of (3.9.1) and (3.9.5) respectively, and let

$$ B(u, v) = \int_{\Omega} \sigma \nabla u \cdot \nabla v \, dx $$

then there exists $C$ independent of $H$ (but dependent on $\sigma$) such that

$$ \|u - u^H\|_{H^1(\Omega)} \leq C \|v - u\|_{H^1(\Omega)}, \quad \forall v \in V. \quad (3.9.8) $$

**Proof.** Let $u$ and $u^H$ be the solutions of (3.9.1) and (3.9.5), then subtracting (3.9.5) from (3.9.1) we arrive at the (Galerkin) orthogonality condition,

$$ a(u - u_h, v) = 0, \quad \forall v \in V. $$

Then due to the coercivity (see appendix A) of $B$, we have

$$ c_1 \|u - u^H\|_{H^1(\Omega)}^2 \leq B(u - u^H, u - u^H) = B(u - u^H, u - v) + B(u - u^H, v - u^H), $$

then, since $v - u^H \in V$, using the orthogonality property we have

$$ c_1 \|u - u^H\|_{H^1(\Omega)}^2 \leq c_2 \|u - u^H\|_{H^1(\Omega)} \|u - v\|_{H^1(\Omega)} $$

and thus (3.9.8) follows. \qed

To give a priori error estimates for the multiscale finite element method we will also require several other key results.

**Theorem 3.9.3.** Let $u_I \in V$ be the interpolant of the homogenised solution $u_0$ using the multiscale basis functions $\psi_i$, i.e. $u_I = \sum_i u_0(x_i) \psi_i$, which implies $Lu_I = 0$ in $K$. As in chapter 2, $u_I$ can be expanded in $K$ as

$$ u_I = u_{I0} + \varepsilon u_{I1} + \varepsilon \theta_I + \ldots \quad (3.9.9) $$

with $L_0 u_{I0} = 0$ in $K$, $u_{I1} = \chi \cdot \nabla u_{I0}$ and $L \theta_I = 0$ in $K$ with $\theta = -u_{I1}$ on $\partial K$. Then there exists a constant $C$, independent of both $\varepsilon$ and $h$, such that

$$ \|u_I - (u_{I0} + \varepsilon u_{I1} + \varepsilon \theta_I)\|_{H^1(\Omega)} \leq C \varepsilon \|f\|_{L^2(\Omega)} \quad (3.9.10) $$

**Proof.** Since $u_{I0}$ satisfies the homogenised problem, and $\Omega$ is convex, then we immediately have

$$ \|\nabla u_{I0}\|_{L^2(\Omega)} \leq \|u\|_{H^2(\Omega)} \leq C \|f\|_{L^2(\Omega)}, $$

$$ \|\Delta u_{I0}\|_{L^2(\Omega)} \leq \|f\|_{L^2(\Omega)} \leq C \|f\|_{L^2(\Omega)}, \quad (3.9.11) $$
for some $C$ independent of $u$. Due to the fact that $u_0$ satisfies homogeneous Dirichlet boundary conditions we also have

$$\| \Delta u_{0b} \|_{L^2(K)} \leq \| u_0 - u_{0b} \|_{L^2(K)} + \| u_0 \|_{L^2(K)} \leq C \| u_0 \|_{L^2(K)},$$

where $u_{0b}$ is the bilinear interpolant of $u_0$ on $K$ and a different constant $C$ independent of $u$ [84].

The fact that $L_0 u_{I0} = 0$ implies that $L_0 (u_{I0} - u_{0b}) = L_0 u_{0b}$, and furthermore implies $\| u_{I0} \|_{L^2(K)} \leq C \| u_0 \|_{L^2(K)}$ [60]. Using theorem 2.1.15, we then have

$$\| u_I - (u_{I0} + \varepsilon u_{I1} + \varepsilon \theta_I) \|_{H^1(K)} \leq C \varepsilon \| \Delta u_{I0} \|_{L^2(K)} \leq C_1 \varepsilon \| \Delta u_0 \|_{L^2(K)}.$$ 

Lastly, taking the sum over all elements $K$, and using the regularity condition, (3.9.11), we arrive at the desired result.

The next theorem is key to the result.

**Theorem 3.9.4.** Let $u$ be the solution of (3.9.1) and let $u_I \in V$ be the interpolant of of the homogenised solution $u_0$ using the multiscale basis functions $\psi_i$. Then there exist constants $C_1$ and $C_2$ independent of both $\varepsilon$ and $h$, such that

$$\| u - u_I \|_{H^1(\Omega)} \leq C_1 H \| f \|_{L^2(\Omega)} + C_2 \sqrt{\frac{\varepsilon}{H}}. \tag{3.9.12}$$

**Proof.** To prove this result we will compare the expansions of $u_\varepsilon$ and $u_I$.

Thanks to theorem 2.1.15 and (3.9.11) we have

$$\| u - (u_0 + \varepsilon u_{I1} + \varepsilon \theta_I) \|_{H^1(\Omega)} \leq C \varepsilon \| \Delta u_{I0} \|_{L^2(K)} \leq C_1 \varepsilon \| \Delta u_0 \|_{L^2(K)}. \tag{3.9.13}$$

Then if we take the asymptotic expansion of both $u$ and $u_I$ and use the triangle inequality repeatedly, we have

$$\| u - u_I \|_{H^1(\Omega)} \leq \| u_0 - u_{I0} \|_{H^1(\Omega)} + \| u_{I1} - u_{I1} \|_{H^1(\Omega)} + \| \theta - \theta_I \|_{H^1(\Omega)} + C \| f \|_{L^2(\Omega)}. \tag{3.9.14}$$

To simplify notation, we will make the following substitutions:

$$r_0 = u_0 - u_{I0}, \quad \text{in } K$$

$$r_1 = u_{I1} - u_{I1}, \quad \text{in } K \quad \text{and}$$

$$r_\theta = \theta - \theta_I \quad \text{in } K.$$

The given estimate for the first is

$$\| r_0 \|_{H^1(\Omega)} \leq C h \| f \|_{L^2(\Omega)},$$

found in [60].
Due to our regularity assumptions we have \( \| \chi(y) \|_{L^\infty(\Omega)} \leq C \), for each component, and that furthermore for the second term we have

\[
\| \varepsilon r_1 \|_{L^2(\Omega)} \leq C \varepsilon h \| f \|_{L^2(\Omega)}.
\]

We now return to an element-wise view, where we have \( \| \nabla \chi_i \|_{L^\infty(K)} \leq C \) for \( j = 1, 2 \).

This implies there exists two constants, \( C_1 \) and \( C_2 \), such that

\[
\| \varepsilon \nabla r_1 \|_{L^2(K)} \leq C_1 \varepsilon \| \nabla r_0 \|_{L^2(K)} + C_2 \varepsilon \| \Delta u_0 \|_{L^2(K)}.
\]

Then again taking a sum over all elements we arrive at

\[
\| \varepsilon r_1 \|_{H^1(\Omega)} \leq (C_1 h + C_2) \varepsilon \| f \|_{L^2(\Omega)}.
\] (3.9.15)

For the third term, \( r_\theta \), a different view is required, as it is hard to relate \( \theta \) and \( \theta_I \) as they correspond to boundary layer correctors for the domain \( \Omega \) and an element \( K \) respectively. Hence we first treat them independently.

Due to the Sobolev embedding theorem [96], we are able to use the fact

\[
\| \theta \|_{H^1(\Omega)} \leq \frac{C}{\sqrt{\varepsilon}}
\]

which immediately implies \( \| \varepsilon \theta \|_{H^1(\Omega)} \leq C \sqrt{\varepsilon} \),

from [55].

The element boundary layer corrector \( \theta_I \) is now considered. By the so-called maximum principle we have

\[
\| \theta \|_{L^\infty(K)} \leq \| \chi(y)^T \nabla u_I(0) \|_{L^\infty(K)} \leq C,
\]

which implies

\[
\frac{1}{h^2} \| \theta \|_{L^1(K)} \leq C,
\]

and then summing over all \( K \) we have

\[
\| \theta \|_{L^1(\Omega)} \leq C \quad \text{implying} \quad \| \varepsilon \theta_I \|_{L^1(\Omega)} \leq C \varepsilon.
\] (3.9.17)

Moreover, from [77] we also have the following result,

\[
\| \varepsilon \nabla \theta_I \|_{L^2(K)} \leq \sqrt{h \varepsilon} \| u_0 \|_{W^{1,\infty}(K)},
\]

which is proven using specially chosen cutoff functions, an equivalent estimate is found in [60] using the trace theorem and an interpolation inequality. Furthermore, since by assumption
For $u_0 \in W^{1,\infty}(\Omega)$ we have
\[
\|\varepsilon \nabla \theta_I\|_{L^2(K)} \leq C\sqrt{h\varepsilon}.
\]

Now, summing over all $K$ gives
\[
\|\varepsilon \nabla \theta_I\|_{L^2(\Omega)} \leq C \sqrt{\frac{\varepsilon}{h}}.
\]  

Hence from (3.9.17) and (3.9.18), we have
\[
\|\varepsilon \theta\|_{H^1(\Omega)} \leq C_1 \sqrt{\frac{\varepsilon}{h}} + C_2 \varepsilon \leq C \sqrt{\frac{\varepsilon}{h}},
\]
and thus the result follows. \hfill \Box

Remark 3.9.5. It is shown in [60] that the following estimate also holds,
\[
\|u - u_h\|_{L^2(\Omega)} \leq C \left( h + \frac{\varepsilon}{h} \right).
\]

3.9.1 Convergence for Non-conforming MsFEM

Here we simply state the main convergence theorems for the Galerkin and PG versions of the non-conforming (oversampled) MsFEM. We do not show proofs, we note only that one cannot use Cea’s lemma as the methods are non-conforming.

Theorem 3.9.6 (Galerkin Version). Let $u_\varepsilon$ be the solution of (3.9.1) and $u_H$ be the solution of (3.9.5) where the basis functions have been computed using oversampling. Assuming that $u_0$, the homogenised counterpart of $u_\varepsilon$ is in $W^{1,\infty}(\Omega)$ and the distance between $\partial K$ and $\partial K^{OS}$ is of order $H$, that is we overstep a coarse element by a distance of approximately $H$ in all directions, then we have
\[
\|u_\varepsilon - u_H\|_{H^1(\Omega)} \leq C_1 \frac{\varepsilon}{H} + C_2 \sqrt{\varepsilon} + C_3 H,
\]
where $\|\cdot\|_{H^1(\Omega)}$ is defined as
\[
\|f\|_{H^1(\Omega)} = \left( \sum_{K \in T_H} \int_K |\nabla f|^2 \, dx \right)^{\frac{1}{2}}.
\]

For a full proof see [97].

Theorem 3.9.7 (Petrov-Galerkin Version). Let $u_\varepsilon$ be the solution of (3.9.1) and $u_h$ be the solution of (3.9.6) where the basis functions have been computed using oversampling. Assuming that $u_0$, the homogenised counterpart of $u_\varepsilon$ is in $H^2(\Omega)$ and the distance between $\partial K$ and
\( \partial K^{OS} \) is sufficiently large, the we have
\[
\| u_\varepsilon - u_H \|_{H^1(\Omega)} \leq C_1 \varepsilon + C_2 \sqrt{\varepsilon} + C_3 H,
\]
where \( \| \cdot \|_{H^1(\Omega)} \) is defined as above.

It would appear as though using the Petrov-Galerkin method nullifies any resonance error. However, the result is in fact based on an underlying assumption that the ratio \( \frac{\varepsilon}{h} \) is less than some constant. For details of this and a full proof of the result see [98].

### 3.10 Discussion

To end this chapter we make mention of several points. Firstly, we state clearly that if the conductivity is changed then the multiscale basis functions will change. This will be alluded to in chapter 6 where we consider the convergence of the multiscale basis functions. This is a major difference between MsFEM and FEM.

To study the electrical impedance tomography (EIT) problem and many other problems, we should enforce Neumann boundary conditions. However, our numerical experiments indicate the MsFEM outlined here do not cope well when Neumann boundary conditions are implemented. This is most likely due to the fact that that the multiscale basis function do not satisfy Neumann boundary conditions on \( \partial \Omega \). By construction the multiscale basis functions in fact satisfy linear Dirichlet boundary conditions on \( \partial \Omega \). The idea of imposing Neumann boundary conditions on basis functions at the boundary is an idea that may warrant further investigation.

Throughout our numerical experiments we have encountered several instances in which the oversampling procedure outlined above in fact underperforms when compared to the conforming Petrov-Galerkin case. Such situations are known in the field [99], and result from the fact that the benefits of oversampling were proven under the assumption of periodicity. We wish to be free from the assumption of periodicity in later chapters and thus from here on out only implement the conforming Petrov-Galerkin formulation of the problem.

If instead of estimating the conductivity we were perhaps interested in the problem of source identification within a multiscaled media then MsFEM would be be a very good choice. In this problem the multiscale basis functions would only need to be computed once, leading to a large saving in overheads.

We would like to note that MsFEM continues to be be enhanced, the latest major breakthrough being advent of the so-called mixed generalised multiscale finite element method which employs principle component analysis to further reduce costs, and while mass conservation is also achieved by considering flux between multiscale basis functions [100].

In this chapter we have introduced the multiscale finite element methods (MsFEM) and discussed some associated ideas, and showed that it can be an attractive alternative in an effort to lower computational costs. We have also shown how in some settings we can improve the accuracy of the method by using the so-called oversampling strategy. Unfortunately,
oversampling is not guaranteed to improve convergence in general and it leads to an increase in computational cost.

We have also shown that implementing MsFEM can be carried out by way of simple linear algebra, which will be drawn on heavily in future chapters. Specifically it will be useful to keep in mind the form of the solution,

$$u = B \mathcal{K}^{-1} F,$$

and the components which make up the matrices $B$ and $\mathcal{K}$.

In this chapter we have considered solving multiscale problems with MsFEM. However the analysis up to this point was based on the assumption that we knew $\sigma$. Suppose then that we wish to infer $\sigma$ from scans of the potential, or boundary measurements, such a topic is the focus of the next chapter.
Inverse problems arise in countless areas of science, from astronomy to underwater acoustics, and from medical imaging to land mine detection. They are characterised by the wish to infer some parameters of interest from noisy indirect measurements, they are also typically ill-posed in the sense of Hadamard [1]. Such problems combine many theories of mathematics with many branches of physics, and date back to at least 1846 when Urbain Jean Le Vierrier inferred the existence of the planet Neptune [101].

Rigorous mathematical foundations of inverse problems are generally attributed to the work of Tikhonov and Arsenin [102]. Since then there has been a huge amount of interest and progress in the field. One of the foremost developments in the field was the ability to frame inverse problems using the Bayesian paradigm. The Bayesian framework allows for a full characterisation of all possible solutions, and their relative probabilities, whilst addressing modelling issues in a clear and precise fashion [103]. There are several key texts on inverse problems in the Bayesian setting: The ideas presented in [104] give a good holistic introduction to the methodology; likewise, [22] provides a good foundation for the Bayesian approach, especially for problems involving differential equations; moreover, in [105] computational aspects of using the Bayesian framework are explored; while [103] gives an analysis of the method in the function space setting.

The task of inferring parameters which exhibit multiscale behaviour is a natural one. However, such a task has only recently been achievable due to increases in computing resources. It is for this reason that the intersection of multiscale modelling and inverse problems seems to be relatively small.

When dealing with multiscale inverse problems the following fundamental question regarding the parameters of interest must first be answered:

- Are we looking for a low dimension homogenised solution, or
• Are we looking for high dimensional *multiscale* solution?

The first question was addressed by in [106] for one dimension by applying results from homogenisation, while [107] and [108] considered both the related problem of high contrast inverse problems and also the source identification problem of interferometry in a *cluttered* medium. More recently [109] considered the problem of finding a perfectly conducting *anomaly* in a random heterogeneous background using electrical impedance tomography (EIT). We should also note [110], where the related question of thermal based damage detection in porous materials was considered.

The second question is closely related to the subfield of so-called large scale inverse problems [111], the main difference being the underlying multiscale structure in the parameters of interest. Large scale inverse problems are characterised by having large dimensional parameters, and have received large amounts of interest recently, see for example any of [111, 112, 113] for associated concepts and applications. We also point out the book [114], which gives a thorough introduction to multiscale modelling within the Bayesian framework.

Most attempts at carrying out inversions for multiscale parameters have revolved around Markov chain Monte Carlo (MCMC) methods [77, 115, 116]. However MCMC type methods generally require tens of thousands of samples for even small dimensional problems, for more see, for example, [117]. In this thesis we will consider so-called gradient based methods, which do away with the need for samples.

Throughout this chapter we will be concerned with inverse problems of the following type: Given some data $D \in \mathcal{H}_2$ find $x \in \mathcal{H}_1$ such that

$$D = A(x) + e,$$

(4.0.1)

where $A : \mathcal{H}_1 \rightarrow \mathcal{H}_2$ is the *forward operator* and $e$ denotes error in the data which comes form some known distribution. Generally $\mathcal{H}_1$ and $\mathcal{H}_2$ are Hilbert or Banach spaces, in practice they will be replaced by $\mathbb{R}^N$ and $\mathbb{R}^M$ respectively. Such a model is called the additive error model for obvious reasons, for more general models such as $D = A(x, e)$, see [22].

In this chapter we review deterministic methods for inverse problems and then move on to the Bayesian approach to inverse problems. We also give an in-depth discussion of the the Bayesian approximation error (BAE) approach.

In section 4.4 we propose a novel extension is to the BAE approach which can be used to aid in the recovery of auxiliary parameters. We attempt to give both an intuition and theoretical grounding for the idea.

### 4.1 Deterministic Methods

In this section we briefly outline some of the typical deterministic methods used for inverse problems.
4.1.1 Least Squares

The methods of least squares (LS) and nonlinear least squares (NLLS) are widely used, and are based on minimisation of the functional

$$\mathcal{F} = \min_x \|D - A(x)\|_2^2.$$ 

Let $A^*$ denote the dual operator of $A$ and $\mathcal{N}(A)$ denote the nullspace of $A$, and suppose for a moment that we have $\mathcal{N}(A) = \{0\}$, implying $\mathcal{N}(A^*A) = \{0\}$. Then the minimiser, $x_{LS}$, of $\mathcal{F}$ is the solution to the normal equations

$$x_{LS} = (A^*A)^{-1}A^*D.$$ 

If the operator $A$ is non-linear, which it is in general, optimisation techniques such as Gauss-Newton or Levenberg-Marquardt iterative methods are used. The last section of this chapter is dedicated to such methods.

In the brief discussion which follows we assume $A : \mathbb{R}^N \to \mathbb{R}^M$ and is linear. This exploration extends naturally to the non-linear case since the search direction used in the aforementioned iterative methods can be viewed as a solution to a linear least squares problem, [118]. We will give a simplistic overview of the least squares method. The method involves two main steps:

1. Project $D$ onto the range of $A$.
2. Solve the resulting system.

These steps illuminate two of the potential hazards of using the least squares method:

1. It is possible that $D$ is nearly orthogonal to the range of $A$, i.e. $D$ is near $\mathcal{N}(A^*)$. However, this would generally be attributed to modelling error(s) [119].

2. If $A$ has condition number $\kappa(A)$, then the condition number of the normal equations is $(\kappa(A))^2$ [120].

The second of these drawbacks is often can pose a major difficulty and is a common occurrence in practice [22]. There are still further drawbacks to the method which often go unstated:

- The method is based on an underlying assumption that $\mathcal{N}(A) = \{0\}$, otherwise $(A^*A)^{-1}$ does not exist.
- The method affords no way of incorporating any prior knowledge we may have about $x$.

The assumption that $A$ is of full rank is cumbersome. In practice the null space of the forward operator $A$ is non-trivial and non-intuitive. Moreover, one may also find that for $\|x\| = 1$ and $x \notin \mathcal{N}(A)$ we still have $\|Ax\| \approx 0$ [121].
Furthermore, in practice the data is corrupted by noise, meaning we are in fact attempting to solve a problem more along the lines of

\[ Ax \approx D, \]

which, in light of the ill-posedness of these types of problems, exasperates our difficulties.

Incorporating prior knowledge would clearly be an advantage, for example in many applications parameters such as conductivity must be non-negative. This idea will be investigated thoroughly throughout the rest of the chapter.

To overcome the general ill-posed nature of inverse problems several regularisation schemes have been developed that in general involve finding and solving well-posed counterparts to the original problem \[22\]. We will now consider two of the most common deterministic regularisation methods along with a discussion on inverse problems in the Bayesian framework.

### 4.1.2 Regularisation by Truncated Singular Value Decomposition

Throughout this section we assume \( \mathcal{H}_1 \) and \( \mathcal{H}_2 \) are Hilbert spaces with inner products \( \langle u, v \rangle_{\mathcal{H}_1} \) for \( u, v \in \mathcal{H}_1 \) and \( \langle u, v \rangle_{\mathcal{H}_2} \) for \( u, v \in \mathcal{H}_2 \) respectively. Furthermore we take \( A : \mathcal{H}_1 \to \mathcal{H}_2 \) to be a compact operator. With these assumptions we have both \( A^*A : \mathcal{H}_1 \to \mathcal{H}_1 \) and \( AA^* : \mathcal{H}_2 \to \mathcal{H}_2 \) are symmetric positive semi-definite and compact operators \[122\]. It follows then by the Hilbert-Schmidt theorem that both the operators \( A^*A \) and \( AA^* \) admit orthogonal eigenvalue decompositions, namely \( \{ \lambda_i, v_i \}_{i \in \mathbb{N}} \) and \( \{ \lambda_i, u_i \}_{i \in \mathbb{N}} \) respectively, with \( \lambda_1 \geq \lambda_2 \geq \cdots \geq 0 \).

Therefore we can write,

\[ A^*Av_i = \lambda_i v_i \quad \text{and} \quad AA^*u_i = \lambda_i u_i. \]

Furthermore, letting \( \vartheta_i = \sqrt{\lambda_i} \), called the singular values of \( A \), we have

\[ Av_i = \vartheta_i u_i \quad \text{and} \quad A^*u_i = \vartheta_i v_i, \]

as shown in \[123\] for example.

Letting \( \mathcal{R}(A^*A) \) denote the range of \( A^*A \), the orthonormal set \( \{ v_i \} \) forms a basis for \( \mathcal{R}(A^*A) = \mathcal{N}(A^*A)^\perp = \mathcal{N}(A)^\perp \subset \mathcal{H}_1 \), and the orthonormal set \( \{ u_i \} \) forms a basis for \( \mathcal{R}(AA^*) = \mathcal{N}(AA^*)^\perp = \mathcal{N}(A^*)^\perp \subset \mathcal{H}_2 \), and we call \( \{ \vartheta_i, v_i, u_i \} \) the singular system for \( A \).

Since \( x \in \mathcal{H}_1 \) and we assume \( D \in \mathcal{H}_2 \) then we have

\[ x = P_{\mathcal{N}(A)}x + \sum_{i=1}^{\infty} \langle x, v_i \rangle_{\mathcal{H}_1} v_i \quad \text{and} \quad D = P_{\mathcal{N}(A^*)}D + \sum_{i=1}^{\infty} \langle D, u_i \rangle_{\mathcal{H}_2} u_i, \]

where \( P_{\mathcal{N}(A)} \) and \( P_{\mathcal{N}(A^*)} \) denote the projection operators onto \( \mathcal{N}(A) \) and \( \mathcal{N}(A^*) \) respectively. Additionally,

\[ Ax = \sum_{i=1}^{\infty} \langle x, v_i \rangle_{\mathcal{H}_1} A v_i \quad \text{and} \quad A^*D = \sum_{i=1}^{\infty} \langle D, u_i \rangle_{\mathcal{H}_2} A^*u_i. \]
We can now express $x$ in terms of $\mathcal{D}$,

$$\langle \mathcal{D}, u_i \rangle_{\mathcal{H}_2} = \langle Ax, \vartheta_i^{-1} A v_i \rangle_{\mathcal{H}_2} = \langle x, \vartheta_i^{-1} A A^* v_i \rangle_{\mathcal{H}_1} = \vartheta_i \langle x, v_i \rangle_{\mathcal{H}_1},$$

and we define the singular value decomposition (SVD) solution to (4.0.1) as

$$x_{\text{SVD}} = \sum_{i=1}^{\infty} \vartheta_i^{-1} \langle \mathcal{D}, u_i \rangle_{\mathcal{H}_2} v_i. \quad (4.1.1)$$

The solution is unique if and only if $N(A) = \emptyset$, however, we also have the following result with regard to existence of the solution.

**Theorem 4.1.1.** (Picard) The equation $Ax = \mathcal{D}$ is solvable in $\mathcal{H}_1$ if and only if

$$\sum_{i=1}^{\infty} \lambda_i^{-1} |\langle \mathcal{D}, u_i \rangle_{\mathcal{H}_2}|^2 < \infty,$$

in which case the solution is given by equation $(4.1.1)$.

This result is based mostly on projections, for a complete proof see [124] among others.

The singular values can (and do) accumulate at zero, and hence for noisy data $\mathcal{D}$ the Picard condition does not hold leading to the ill-posedness of inverse problems. We also note that the degree to which a problem is ill-posed depends on the rate at which the singular values decay to zero. A natural remedy to this problem is to truncate the series, i.e. take only the first $p \in \mathbb{N}$ singular values in account. This process leads to a new solution to (4.0.1),

$$x_{\text{TSVD}} = \sum_{i=1}^{p} \vartheta_i^{-1} \langle \mathcal{D}, u_i \rangle_{\mathcal{H}_2} v_i.$$

This technique is referred to as truncated singular value decomposition (TSVD) and has been used widely since the 1960’s. There are various methods for finding a suitable truncation value, see for example [22]

### 4.1.3 Tikhonov(-Phillips) Regularisation

The fact that the singular values of $A$ tend to zero, results in the norm of the solution growing arbitrarily large. Although TSVD looks to be an attractive remedy to this problem another of the drawbacks to SVD and TSVD is that in most cases it is not analytically available, and extremely expensive to calculate the singular system for $A$ [125, 22].

The goal regularisation in the deterministic setting is to simultaneously minimise the residual $r = \|Ax - \mathcal{D}\|$ while also enforcing any assumptions we may have about $x$, such as boundedness. To accommodate this, a popular choice is the Tikhonov regularised solution,

$$x_{\text{TR}} = \min_x \{ \|Ax - \mathcal{D}\|_{\mathcal{H}_2}^2 + \delta \|x\|_{\mathcal{H}_1}^2 \},$$
where $\delta > 0$ is referred to as the regularisation parameter. There are several methods used to determine a suitable $\delta$, with one of the most widely employed being the so called Morozov discrepancy principle [22].

The method can also be generalised to enforce a range of assumptions on $x$, e.g. smoothness, band limitedness or nonnegativity to name a few. This introduces the generalised1 Tikhonov regularisation scheme, where we attempt to solve the following problem,

$$x_{TR} = \min_x \{ \|Ax - D\|^2_2 + \delta G(x) \},$$

where $G : \mathcal{H}_1 \to \mathbb{R}$ is a nonnegative functional. $G$ is generally termed the penalty function and is used to enforce conditions such as those mentioned above. The most common form is

$$G(x) = \|L(x - x_0)\|^2_{\mathcal{H}},$$

with $\mathcal{H}$ a Hilbert space and $L$ being a linear operator such that $L : \mathcal{D}(L) \to \mathcal{H}$, with $\mathcal{D}(L)$ a subset of $\mathcal{H}_1$, and $x_0 \in \mathcal{H}_1$ given [22]. Typically one enforces some degree of smoothness on $x$, and thus $L$ is some type of differential operator (assuming $\mathcal{H}_1$ is a function space).

In the finite dimensional setting we typically have $L$ a matrix and the functional to be minimised can be written as

$$\|Ax - D\|^2_2 + \delta \|L(x - x_0)\|^2_2 = \left\| \begin{bmatrix} A \\ \sqrt{\delta L} \end{bmatrix} x - \begin{bmatrix} \mathcal{D} \\ \sqrt{\delta L}x_0 \end{bmatrix} \right\|^2_2.$$

Hence the minimiser is

$$x_{TR} = \hat{A}^\dagger \begin{bmatrix} \mathcal{D} \\ \sqrt{\delta L}x_0 \end{bmatrix}, \quad \text{with} \quad \hat{A} = \begin{bmatrix} A \\ \sqrt{\delta L} \end{bmatrix},$$

provided $\mathcal{N}(A) \cap \mathcal{N}(L) = \emptyset$ [22], and where $\hat{A}^\dagger$ is the Moore-Penrose pseudoinverse.

### 4.2 Bayesian Methods

There are several key philosophical differences between deterministic regularisation schemes and Bayesian methods. Two of the most significant differences are as follows:

1. In the Bayesian setting we admit we have a lack of information and address this matter by modelling unknown parameters as random variables with probability densities which encode our level of information concerning the parameters [105].

2. In the Bayesian setting the solution of the inverse problem is in fact the posterior probability distribution [22].

The first difference should not be understated. In the broadest sense inverse problems are concerned with inferring some parameters based on collected data. To connect the two,

---

1 Although this form more closely resembles the one put forward by Tikhonov [102].
data and parameters, we must have, or construct, a model which exhibits the dependence of
the data on the parameter. Unfortunately, it is quite likely the model is inaccurate and in
many cases will contain auxiliary sometimes referred to as nuisance parameters. Moreover, as
stated earlier, the measurements will contain noise [22]. In the Bayesian setting we account
for these underlying uncertainties by modelling all unknowns as random variables, having
joint and marginal probability densities, allowing us to encode our level of uncertainty into
the relevant distributions. We make explicit that in this thesis we assume all random variables
are absolutely continuous, and thus have associated probability density functions [22]. We
will use the terms density and distribution when referring to the probability density function.

The second difference is fundamental to the interpretation of a solution. Solutions in the
classical setting come in the form of a single point estimate, while in the Bayesian setting
the solution is an entire probability distribution, meaning we have far more insight about the
parameters of interest. If computational resources and time are extremely limited (as is often
the case) a single representative estimate, known as a point estimate is generally found as
the solution along with interval estimates, known as spread estimates [105]. Different types
of point estimates will be discussed in the next section. The rest of this section is devoted to
definitions concerning probability theory, where we assume we are in the finite dimensional
setting.

Firstly we explicitly state that our parameter of interest, \( x \), and the noise, \( e \), are random
variables. Since, by assumption,

\[
D = A(x, e) = A(x) + e,
\]

our data \( D \) is also a random variable.

Any information we have about \( x \) prior to taking measurements in the Bayesian setting is
encoded into a probability density, \( \pi_{\text{prior}}(x) \), called the prior density, or simply the prior, for
obvious reasons.

If we assume our parameter of interest was some specific value, say \( x = x^* \), and
\( \pi(x, D) \) is the joint probability density, then the conditional distribution of \( D \) is

\[
\pi(D|x^*) = \frac{\pi(x^*, D)}{\pi(x^*)}, \quad \pi(x^*) > 0,
\]

which is known as the likelihood function.

If we assume instead that we have measurement data \( D = D^* \), which has come from a
probability density \( \pi(D) \), then the conditional distribution of \( x \) is

\[
\pi(x|D^*) = \frac{\pi(x, D^*)}{\pi(D^*)}, \quad \pi(D^*) > 0,
\]

and is known as the posterior distribution or simply the posterior. This distribution relates
back to the interpretation of a solution, since this distribution answers the question: Given
data \( D = D^* \), find the conditional distribution of \( x \), i.e. find \( \pi(x|D^*) \), the posterior.
We end this section with the central result concerning inverse problems in the Bayesian framework.

**Theorem 4.2.1.** Suppose that the random variable $x \in \mathbb{R}^N$ has a prior probability density $\pi_{\text{prior}}(x)$ and that the data $\mathcal{D}$ is an observed value of another random variable $y \in \mathbb{R}^M$. The posterior probability distribution of $x$ given the data $\mathcal{D}$ is then

$$
\pi_{\text{post}}(x) = \pi(x \mid \mathcal{D}) = \frac{\pi_{\text{prior}}(x) \pi(\mathcal{D} \mid x)}{\pi(\mathcal{D})}, \quad \pi(\mathcal{D}) > 0.
$$

(4.2.1)

**Proof.** This result is simply Baye’s theorem [24] in the inverse problems setting. \qed

Throughout this thesis we will often use the following form for the posterior probability distribution,

$$
\pi_{\text{post}}(x) \propto \pi(\mathcal{D} \mid x) \pi_{\text{prior}}(x).
$$

(4.2.2)

### 4.2.1 Estimators

As stated previously, in the Bayesian setting the solution of an inverse problem is a posterior probability distribution. If computational resources and time were not an issue, visualisation of the posterior distribution could be a possibility. However, in most real world applications such a task is computationally infeasible [22]. This generally results in the computation of point estimates and spread estimates to establish the most likely values for the parameters of interest and how confident we are that the parameters are within certain bounds. For a good discussion on taking the posterior distribution as the solution to an inverse problem and on the solutions of inverse problems in general see chapters 3 and 4 of [105].

We will discuss here three commonly used point estimates: the **maximum likelihood** (ML) estimate, $x_{\text{ML}}$; the **conditional mean** (CM) estimate, $x_{\text{CM}}$; and the **maximum a posteriori** (MAP) estimate, $x_{\text{MAP}}$. Of these estimates the latter are the two most commonly used point estimates in inverse problems [126].

The maximum likelihood estimate is defined as

$$
x_{\text{ML}} = \arg \max_{x \in \mathbb{R}^N} \pi(\mathcal{D} \mid x),
$$

if the maximiser exists. This is a common estimate used in frequentist statistics. However, maximum likelihood estimation can be troublesome or infeasible, and can be interpreted as non-regularised least squares [105]. We thus give little attention to the maximum likelihood estimate.

The conditional mean estimate is defined as

$$
x_{\text{CM}} = \mathbb{E}(x \mid \mathcal{D}) = \int_{\mathbb{R}^N} x \pi(x \mid y) \, dx
$$

if the integral exists. The drawback of the conditional mean is related to the dimension of the problem. As can be seen from the definition, finding the conditional mean usually boils
down to an integration problem. Quadrature in high dimensions can be infeasible [22], which may lead people to use Markov chain Monte Carlo (MCMC) methods. However, we consider the case of limited time and computational resources, and thus do not use MCMC type methods.

The maximum a posteriori estimate is defined as

$$x_{\text{MAP}} = \arg \max_{x \in \mathbb{R}^N} \pi(x|\mathcal{D}),$$

if the maximiser exists. The maximum a posteriori estimate can naturally be interpreted as the mode.

We have just listed several particular estimates that are used in the Bayesian framework as representative values for the posterior distribution, and as we will see throughout this thesis these estimates, particularly the MAP estimate, may coincide with classical regularised solutions. However, the Bayesian setting can offer further insights, estimates being only a small part of the information we gather about the unknown parameters [105, 22].

We should mention at this point that despite the aforementioned difficulties in calculating $x_{\text{CM}}$, there is still ongoing debate as to whether $x_{\text{CM}}$ or $x_{\text{MAP}}$ is theoretically optimal. These arguments often revolve around ideas from Bayesian decision theory and the concept of the Bayesian cost [22, 126, 127]. It is beyond the scope of this thesis to establish the optimality of any of the point estimates. However, the matter is given a thorough analysis in [126] where a convincing argument is presented for the use of the MAP estimate. It should be noted that, in the case of a normal distribution, we have $x_{\text{CM}} = x_{\text{MAP}}$ [22].

### 4.2.2 Construction of the Likelihood Function and Posterior Density

We shall now consider construction of the likelihood function and the posterior density. Consider again the problem at hand,

$$\mathcal{D} = A(x) + e,$$  \hspace{1cm} (4.2.3)

where $x \in \mathbb{R}^N$, and $\mathcal{D}, e \in \mathbb{R}^M$. Note that the following is also valid for the general Hilbert space seeing also. We assume $x$ and $e$ have respective probability distributions $\pi_{\text{prior}}(x)$ and $\pi(e)$ which are known. Then by definition of the likelihood and Bayes’ formula, (4.2.2), we have

$$\pi(\mathcal{D} | x) = \int_{\mathbb{R}^M} \pi(\mathcal{D}, e|x) \, de = \int_{\mathbb{R}^M} \pi(\mathcal{D}|x, e)\pi(e|x) \, de.$$  

For any single realisation of $x$ and $e$ the data, $\mathcal{D}$, is fixed and therefore $\pi(\mathcal{D}|x)$ is a delta distribution, and by assumption $x$ and $e$ are independent. We continue, with

$$\pi(\mathcal{D} | x) = \int_{\mathbb{R}^M} \delta(\mathcal{D} - A(x) - e)\pi(e|x) \, de = \pi_{e|x}(\mathcal{D} - A(x)|x),$$  \hspace{1cm} (4.2.4)
where $\pi_{e|x}$ denotes the probability density of $e$ conditioned on $x$. Thus

$$\pi(x|\mathcal{D}) = \pi_{\text{post}}(x) \propto \pi_{e|x}(\mathcal{D} - A(x))\pi_{\text{prior}}(x). \quad (4.2.5)$$

Equation (4.2.5) is the posterior density, and will be used extensively throughout this thesis.

### 4.2.3 Gaussian Densities

Gaussian probability densities are fundamental to the study of inverse problems, particularly in statistical inversion theory. Not only are they common when modelling parameters such as pressure, temperature, and electric current [105], but due to the central limit theorem Gaussian densities often offer good approximations to non-Gaussian distributions when the observation is physically based on a large number of mutually independent random events [22]. Furthermore, they offer closed forms for many quantities, and are one of the easiest type of distributions to work with.

**Definition 4.2.2.** Let $x_0 \in \mathbb{R}^N$, and $\Gamma \in \mathbb{R}^{N \times N}$ be a symmetric positive definite matrix. An $n$-variate Gaussian, or $n$-variate normal, random variable $x$ with mean $x_0$ and covariance $\Gamma_x$ is a random variable with probability density

$$\pi(x) = \left(\frac{1}{2\pi}\right)^{\frac{N}{2}} \frac{1}{\sqrt{|\Gamma_x|}} \exp\left(-\frac{1}{2} (x - x_0)^T \Gamma_x^{-1} (x - x_0)\right),$$

where $|\Gamma|$ denotes the determinant of $\Gamma$. In such a case we use the notation

$$x \sim \mathcal{N}(x_0, \Gamma_x).$$

Assume $x$ and $e$ are independent and are both Gaussian with means $x_0$ and $e_0$ respectively. We assume for the moment that $A$ is linear, which implies that $\mathcal{D}$ is also Gaussian, and all marginal and conditional distributions are Gaussian [22]. Using this fact, and equation (4.2.4), we have

$$\pi(\mathcal{D} | x) \propto \exp\left(-\frac{1}{2} (\mathcal{D} - Ax - e_0)^T \Gamma_e^{-1} (\mathcal{D} - Ax - e_0)\right) = \exp\left(-\frac{1}{2} \|L_e (\mathcal{D} - Ax - e_0)\|^2\right),$$

where $L_e^T L_e = \Gamma_e^{-1}$. Also, the prior density for $x$ can be written as

$$\pi_{\text{prior}}(x) \propto \exp\left(-\frac{1}{2} \|L_x (x - x_0)\|^2\right),$$

where $L_x^T L_x = \Gamma_x^{-1}$. By substituting these into equation (4.2.5) we can represent the posterior density as

$$\pi_{\text{post}}(x) \propto \exp\left(-\frac{1}{2} \left(\|L_e (\mathcal{D} - Ax - e_0)\|^2 + \|L_x (x - x_0)\|^2\right)\right). \quad (4.2.6)$$
The following theorem provides insight and in some cases, provides an opportunity to derive closed formulas for the MAP (or CM) estimate of $x$ and associated covariances.

**Theorem 4.2.3.** Let $\mathbb{R}_+$ denote the positive reals, and $x \in \mathbb{R}^N$ and $y \in \mathbb{R}^M$ be $N$-variate and $M$-variate Gaussian random variables respectively, with joint probability density $\pi: \mathbb{R}^N \times \mathbb{R}^M \rightarrow \mathbb{R}_+$ of the form

$$
\pi(x, y) \propto \exp \left( -\frac{1}{2} \begin{bmatrix} x^T - x_0^T \\ y^T - y_0^T \end{bmatrix}^T \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} x - x_0 \\ y - y_0 \end{bmatrix} \right),
$$

where the positive definite symmetric block matrix $P$ is called the precision matrix with

$$
\begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} = \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix}^{-1} = \Gamma^{-1}, \quad P_{11}, \Gamma_{11} \in \mathbb{R}^{N \times N} \quad \text{and} \quad P_{22}, \Gamma_{22} \in \mathbb{R}^{M \times M}.
$$

Then the probability density of $x$ conditioned on $y$, $\pi(x|y) : \mathbb{R}^n \rightarrow \mathbb{R}_+$ is of the form

$$
\pi(x|y) \propto \exp \left( -\frac{1}{2} (x - \hat{x})^T P_{11}^{-1} (x - \hat{x}) \right),
$$

where

$$
\hat{x} = x_0 - P_{11}^{-1} P_{12} (y - y_0).
$$

**Proof.** To begin with we shift the coordinate origin to $(x_0, y_0)$ simply for ease of computation and to save space. Secondly, recall that from Bayes' theorem $\pi(x|y) \propto \pi(x, y)$ and thus we treat the joint density as a function of $x$ only. Consider now the quadratic form associated with the joint density presented in (4.2.3) with the adjusted coordinate system

$$
-\frac{1}{2} \begin{bmatrix} x^T \\ y^T \end{bmatrix}^T \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = -\frac{1}{2} (x^T P_{11} x + 2x^T P_{12} y + y^T P_{22} y),
$$

since $P_{12} = P_{21}^T$. Then by simply completing the square we see that

$$
-\frac{1}{2} (x^T P_{11} x + 2x^T P_{12} y + y^T P_{22} y) = -\frac{1}{2} (x + P_{11}^{-1} P_{12} y)^T P_{11} (x + P_{11}^{-1} P_{12} y) - y^T C y,
$$

where

$$
C = -\frac{1}{2} (P_{22} - P_{22}^T P_{11}^{-1} P_{12}),
$$

and $P_{11}^{-1}$ exist by the positive definiteness of $P$. The term $y^T C y$ is mutually independent of $x$ and hence can be factored out of the density and taken only as a proportionality constant. 

**Remark 4.2.4.** It is also common to present the results of the above theorem in terms of the blocks of $\Gamma$. Taking this path results in

$$
\pi(x|y) \propto \exp \left( -\frac{1}{2} (x - \hat{x})^T (\Gamma_{11} - \Gamma_{12} \Gamma_{22}^{-1} \Gamma_{21})^{-1} (x - \hat{x}) \right),
$$

where

$$
\hat{x} = x_0 - P_{11}^{-1} P_{12} (y - y_0).
$$
These results can be useful and are based on the Schur complements of the blocks of $\Gamma$, details of which are given in [22].

We can now apply these results to our problem at hand. Recall,

$$
\pi(x|\mathcal{D}) = \pi_{\text{post}}(x) \propto \pi_e(\mathcal{D} - A(x))\pi_{\text{prior}}(x),
$$

(4.2.8)

and $x$ and $e$ are mutually independent, are both Gaussian, have means $x_0$ and $e_0$ respectively, and $A$ is linear. Then

$$
E\{\mathcal{D}\} = E\{Ax + e\} = Ax_0 + e_0.
$$

By definition

$$
E \left\{ (x - x_0)(x - x_0)^T \right\} = \Gamma_x \quad \text{and} \quad E \left\{ (e - e_0)(e - e_0)^T \right\} = \Gamma_e,
$$

and we deduce

$$
E \left\{ (\mathcal{D} - \mathcal{D}_0)(\mathcal{D} - \mathcal{D}_0)^T \right\} = E \left\{ (A(x - x_0) + e - e_0)(A(x - x_0) + e - e_0)^T \right\} = A\Gamma_x A^T + \Gamma_e
$$

and

$$
E \left\{ (x - x_0)(\mathcal{D} - \mathcal{D}_0)^T \right\} = E \left\{ (x - x_0)(A(x - x_0) + e - e_0)^T \right\} = \Gamma_x A^T.
$$

From these two results we can define the covariance matrix,

$$
cov \left( \begin{bmatrix} x \\ \mathcal{D} \end{bmatrix} \right) = E \left\{ \begin{bmatrix} x - x_0 \\ \mathcal{D} - \mathcal{D}_0 \end{bmatrix} \begin{bmatrix} x^T - x_0^T \\ \mathcal{D}^T - \mathcal{D}_0^T \end{bmatrix}^T \right\} = \begin{bmatrix} \Gamma_x & \Gamma_x A^T \\ A\Gamma_x & A\Gamma_x A^T + \Gamma_e \end{bmatrix}.
$$

In light of theorem 4.2.3 and remark 4.2.4 we can immediately conclude the following.

**Theorem 4.2.5.** Suppose $x \in \mathbb{R}^N$ and $e \in \mathbb{R}^M$ are mutually independent Gaussian random variables, with

$$
x \sim \mathcal{N}(x_0, \Gamma_x), \quad e \sim \mathcal{N}(e_0, \Gamma_e),
$$

where $\Gamma_x \in \mathbb{R}^{N \times N}$ and $\Gamma_e \in \mathbb{R}^{M \times M}$ are symmetric positive definite. Suppose the model is linear so that $A \in \mathbb{R}^{M \times N}$ and we have noisy measurements $\mathcal{D} \in \mathbb{R}^M$, so that

$$
\mathcal{D} = Ax + e.
$$

Then the posterior probability of $x$ given a particular value for $\mathcal{D}$ is

$$
\pi(x|\mathcal{D}) \propto \exp \left( -\frac{1}{2} (x - \hat{x})^T \Gamma_{\text{post}}^{-1} (x - \hat{x}) \right),
$$
\[
\hat{x} = x_0 + \Gamma_x A^T (A \Gamma_x A^T + \Gamma_e)^{-1} (\mathcal{D} - Ax_0 - e_0)
\]

and

\[
\Gamma_{\text{post}} = \Gamma_x - \Gamma_x A^T (A \Gamma_x A^T + \Gamma_e)^{-1} A \Gamma_x.
\]

**Remark 4.2.6.** We note that along with the above, equivalent forms for both posterior mean and covariance are given in texts such as [104] and [22].

**Remark 4.2.7.** It is also worth taking note that by the above theorem we have

\[
\Gamma_x - \Gamma_{\text{post}} = \Gamma_x A^T (A \Gamma_x A^T + \Gamma_e)^{-1} A \Gamma_x.
\]

Moreover, the matrix \((A \Gamma_x A^T + \Gamma_e)^{-1}\) is symmetric and positive definite and thus admits a factorisation of the form \(C^T C = (A \Gamma_x A^T + \Gamma_e)^{-1}\). Therefore

\[
\Gamma_x - \Gamma_{\text{post}} = \Gamma_x A^T (A \Gamma_x A^T + \Gamma_e)^{-1} A \Gamma_x = \Gamma_x A^T C^T C A \Gamma_x = (C A \Gamma_x)^T (C A \Gamma_x).
\]

That is to say, that in the sense of quadratic forms

\[
\Gamma_x \geq \Gamma_{\text{post}}.
\]

Theorem 4.2.5 characterises the CM estimate. However, the MAP estimate satisfies

\[
x_{\text{MAP}} = \max_x \left\{ \exp \left( -\frac{1}{2} (x - \hat{x})^T \Gamma_{\text{post}}^{-1} (x - \hat{x}) \right) \right\},
\]

and thus

\[
x_{\text{MAP}} = x_{\text{CM}} = \hat{x},
\]

as expected since \(x\) and \(e\) were Gaussian random variables and the model was linear. Unfortunately, in general the model is nonlinear, and hence we cannot conclude that the posterior \(\pi_{\text{post}}(x) = \pi(x|\mathcal{D})\) is Gaussian and hence the characterisations of \(x_{\text{CM}}\) and \(x_{\text{MAP}}\) may no longer be valid. However maintaining the assumption that \(x\) and \(e\) are mutually independent Gaussian random variables, for a general model \(A(x)\) we can still derive the MAP estimate as follows,

\[
x_{\text{MAP}} = \max_x \{ \pi(x|\mathcal{D}) \} = \max_x \{ \pi_{\text{post}}(x) \} \propto \max_x \{ \pi_e(\mathcal{D} - A(x) - e_0) \pi_{\text{prior}}(x) \}
\]

\[
= \max_x \left\{ \exp \left( -\frac{1}{2} \left( \|L_e (\mathcal{D} - A(x) - e_0)\|^2_2 + \|L_x (x - x_0)\|^2_2 \right) \right) \right\}
\]

\[
= \min_x \left\{ \frac{1}{2} \left( \|L_e (\mathcal{D} - A(x) - e_0)\|^2_2 + \|L_x (x - x_0)\|^2_2 \right) \right\}.
\]

(4.2.9)

The resulting problem is a nonlinear least-squares problem and thus there are well developed methods we can employ to tackle this problem, which will be outlined at the end of the current chapter.
For general nonlinear models we do not have a closed form for the posterior covariance. To this end once the MAP estimate is calculated we take the localised approximate covariance

\[ \Gamma_{\text{post}} = (J^T \Gamma_e^{-1} J + \Gamma_x^{-1})^{-1} \]  

(4.2.10)

as an approximation to posterior spread, where \( J \) is the Jacobian of \( A \) with respect to \( x \) evaluated at \( x = x_{\text{MAP}} \). This is analogous to estimating the posterior density as a Gaussian centered at \( x_{\text{MAP}} \) [104].

### 4.3 Bayesian Approximation Error Approach

This thesis is concerned with estimating high dimensional parameters which vary on multiple scales. The above theory is adequate to reconstruct these parameters, however, the computational cost associated with reconstructing the full high dimensional parameters can be prohibitive. Thus some type of model reduction may be warranted. However, reducing the model generally leads to errors in our model and may lead to large errors in our parameter estimates. To this end, the so called *Bayesian approximation error* (BAE) approach was introduced in [22] and [21]. The approach can also be applied to handle a broad variety of modelling and approximation errors, for an in depth discussion on the approach and some interesting examples see [128].

We will develop the approach for the case in which there are auxilliary parameters and, simultaneously, a reduced numerical approximation for the forward problem is employed similar to the derivation in [128]. Suppose all unknowns are \((x, z, e)\), where \( e \) is the additive error, and \((x, z)\) are two, not necessarily mutually independent, distributed parameters where (initially) only \( x \) is of interest and we will call \( z \) the *auxiliary unknowns*.

Let

\[ \mathcal{D} = \bar{A}(\bar{x}, z) + e \in \mathbb{R}^M \]

denote the relationship between the data and an accurate forward model \( \bar{A}(\bar{x}, z) \), and let \( e \) be mutually independent of both \( x \) and \( z \). We approximate the accurate primary parameter of interest, \( \bar{x} \), by a lower dimensional projection, \( x = \mathcal{P} \bar{x} \), where \( \mathcal{P} \) denotes a projection operator.

The main idea of the approach is instead of using both the accurate model \( \bar{A} \) and accurate representation of the unknowns \((\bar{x}, z)\), we fix the random variable \( z = z_0 \) and use computationally approximative model,

\[ A_{z_0}(x) = A(x, z_0). \]

In general as one might expect, the predictions of the models can be very different. To this
Chapter 4. Inverse Problems Theory and Methodologies

restate our observational model,

\[ \mathcal{D} = \bar{A}(\bar{x}, z) + e \]

\[ = A_{z_0}(x) + (\bar{A}(\bar{x}, z) - A_{z_0}(x)) + e \]

\[ = A_{z_0}(x) + \epsilon + e, \]

where \( \epsilon = \epsilon(\bar{x}, z) = \bar{A}(\bar{x}, z) - A_{z_0}(x) \) is defined as the approximation error. In this way the approximation error can viewed as the discrepancy between the predictions of the accurate and the predictions of the approximative models [128]. In the absence of noise it could also be viewed as the residual between data and predictions of the approximative model, a viewpoint that will be useful in section 4.4.

4.3.1 Approximation Error Marginalisation

As we are maintaining the Bayesian viewpoint, all parameters are treated as random variables, and thus there exists a joint probability density, \( \pi(\bar{x}, z, \epsilon) \). Specifically this means that once we fix both the accurate and approximative models, \( \bar{A} \) and \( A \) respectively, we have

\[ \pi(\epsilon|\bar{x}, z) = \delta(\epsilon - (\bar{A}(\bar{x}, z) - A_{z_0}(x))) , \]  

(4.3.1)

since \( x = \mathcal{P}\bar{x} \) and \( z_0 \) is known. We now make the approximation that \( \epsilon(\bar{x}, z) \approx \epsilon(x, z) \). This approximation holds for inverse problems in most cases [128], and as a consequence we have \( \pi(\epsilon|\bar{x}, z) \approx \pi(\epsilon|x, z) \).

It will be useful to make the following substitution,

\[ \nu(\bar{x}, z) = \epsilon(\bar{x}, z) + e, \text{ implying } \pi(\nu|\bar{x}, z, e) = \delta(\nu - \epsilon - e). \]  

(4.3.2)

We can then restate our problem as

\[ \mathcal{D} = A_{z_0}(x) + \nu, \]  

(4.3.3)

and we will at times refer to \( \nu \) as the noisy residual since

\[ \nu(\bar{x}, z) = \epsilon(\bar{x}, z) + e = \mathcal{D} - A_{z_0}(x). \]

Furthermore we update the posterior density,

\[ \pi_{post}(x) \propto \pi_{\nu|x}(\mathcal{D} - A_{z_0}(x)|x)\pi_{prior}(x), \]  

(4.3.4)

with \( \pi_{\nu|x} \) denoting the probability density of \( \nu \) given \( x \). In general we lack a closed form and sometimes even intuition as to exactly what the density \( \pi_{\nu|x} \) is or how it behaves, we do know however that it is a function of \( x, z \) and \( e \). To this end we can apply Bayes’ theorem to the
associated joint distribution,

\[
\pi(\nu|x) = \int \int \int \pi(\nu, \epsilon, z, e|x) \, dz \, d\epsilon \\
= \int \int \int \delta(\nu - \epsilon - e) \pi(\epsilon, z|x) \pi(\epsilon) \, dz \, d\epsilon \\
= \int \pi(\epsilon|z) \pi(\nu - \epsilon) \, d\epsilon \\
= \int \pi(\nu|x) \pi_{\epsilon}(\nu - \epsilon) \, d\epsilon,
\]

and thus

\[
\pi_{\nu|x}(D - A_{z_0}(x)|x) = \int \pi(\epsilon|x) \pi_{\epsilon}(D - A_{z_0}(x) - \epsilon) \, d\epsilon.
\]

Substituting this into equation (4.3.4), we attain the \textit{approximation error posterior distribution},

\[
\pi_{\text{post}}(x) \propto \left( \int \pi_{\epsilon}(D - A_{z_0}(x) - \epsilon) \pi(\epsilon|x) \, d\epsilon \right) \pi_{\text{prior}}(x).
\]

Everything up to this point is analytically correct. However, here we make the approximation that \(\pi_{\epsilon}\) is Gaussian, and thus all relevant random variables are jointly Gaussian, that is

\[
\begin{bmatrix}
x \\
z \\
\epsilon \\
e
\end{bmatrix} \sim \mathcal{N}
\begin{pmatrix}
x_0 \\
z_0 \\
\epsilon_0 \\
e_0
\end{pmatrix},
\begin{pmatrix}
\Gamma_x & \Gamma_{xz} & \Gamma_{xe} & 0 \\
\Gamma_{zx} & \Gamma_z & \Gamma_{ze} & 0 \\
\Gamma_{ex} & \Gamma_{ez} & \Gamma_e & 0 \\
0 & 0 & 0 & \Gamma_e
\end{pmatrix},
\] \hspace{1cm} (4.3.5)

This approximation has been used to give feasible approximations for the posterior distributions in a large variety of inverse problems, see for example \cite{129, 130, 131} among many.

Use of the normal approximation yields an updated version of equation (4.2.8),

\[
\pi_{\text{post}} \propto \exp \left( -\frac{1}{2} \left( \| L_{\nu|x} (D - A_{z_0}(x) - \nu_0|x) \|_2^2 + \| L_x (x - x_0) \|_2^2 \right) \right),
\] \hspace{1cm} (4.3.6)

along with a revised MAP estimate,

\[
x_{\text{MAP}} = \max_x \left\{ \exp \left( -\frac{1}{2} \left( \| L_{\nu|x} (D - A_{z_0}(x) - \nu_0|x) \|_2^2 + \| L_x (x - x_0) \|_2^2 \right) \right) \right\},
\] \hspace{1cm} (4.3.7)

where \(\Gamma_{\nu|x}^{-1} = L_{\nu|x}^T L_{\nu|x}\). In a similar fashion to theorem 4.2.5 the approximative conditional
expectation and covariance can be computed as
\[ \nu_{0|x} = e_0 + \epsilon_{0|x} + \Gamma_{\epsilon x} \Gamma_x (x - x_0) \]  
and \[ \Gamma_{\nu|x} = \Gamma_e + \Gamma_{\epsilon|x} = \Gamma_e + \Gamma_{\epsilon} - \Gamma_{\epsilon x} \Gamma_x^{-1} \Gamma_x \epsilon, \]
respectively. However, in several applications \( x \) and \( \epsilon \) have been approximated as mutually independent resulting in
\[ \nu_{0|x} = e_0 + \epsilon_0 \]  
and \[ \Gamma_{\nu|x} \approx \Gamma_{\nu} = \Gamma_{\epsilon} + \Gamma_{\epsilon}, \]
and the estimate
\[ x_{MAP} = \max_x \left\{ \exp \left( -\frac{1}{2} \left( \| L_{\nu} (D - A z_0 (x) - \nu_0) \|^2_2 + \| L_x (x - x_0) \|^2_2 \right) \right) \right\}. \]
This approach is often referred to as the \textit{enhanced error model}, see for example [21, 132]. In what follows we will use this enhanced error model approximation.

We end this section with a couple of definitions which are necessary.

\textbf{Definition 4.3.1.} \textit{We say that the approximation error, \( \epsilon \), dominates the measurement errors, \( e \), if}
\[ \|e_0\|^2 + \text{trace} \Gamma_e < \|\epsilon_0\|^2 + \text{trace} \Gamma_\epsilon. \]  
If (4.3.13) holds then neglecting the approximation errors from the likelihood model will likely result in meaningless posterior estimates. Moreover, if
\[ \epsilon_0(k)^2 + \Gamma_\epsilon(k,k) < \epsilon^2_0 + \Gamma_\epsilon(k,k) \]
holds for any \( k \) then this could also be the case [128].

\textbf{Definition 4.3.2.} \textit{In the case of white noise, with \( e \sim \mathcal{N} (e_0, \delta^2_e I) \), we define the noise level as}
\[ \text{"noise level"} = \frac{\delta_e}{[D]} \times 100\%, \]
with \([D]\) denoting the spatial mean of the absolute value of the measured data.

We remark that in this thesis we consider noise only of the form \( e \sim \mathcal{N} (e_0, \delta^2_e I) \), However generalisations are straightforward, see for example [22].

\section*{4.3.2 Approximation Error Samples}
Recall, in (4.3.5) we approximated the joint distribution of our principal parameters, \( x \), and the distribution of the approximation errors, \( \epsilon \) as Gaussians. However, whether or not this is truly the case in practice is of little concern. The prior distribution \( \pi(x, z) \) does not in
fact need to be normal nor does the marginal prior, \( \pi_{\text{prior}}(\tilde{x}) \) \[128\]. No matter what the prior density \( \pi(\tilde{x}, z) \) we start by drawing a set of \( q \in \mathbb{N} \) samples which we shall denote \( (\tilde{x}^{(\ell)}, z^{(\ell)}) \) and compute the approximation errors,

\[
\epsilon^{(\ell)} = \tilde{A}(\tilde{x}^{(\ell)}, z^{(\ell)}) - A(x^{(\ell)}, z_0), \quad \ell = 1, 2, \ldots, q,
\]

with \( x^{(\ell)} = P\tilde{x}^{(\ell)} \). We then make the Gaussian approximation of \( \pi(\epsilon, x) \) and compute the sample means and joint sample covariances from the ensemble.

\[
\begin{bmatrix}
\tilde{x}^{(\ell)} \\
\epsilon^{(\ell)}
\end{bmatrix}, \quad \ell = 1, 2, \ldots, q. \tag{4.3.14}
\]

In most cases this is the extent to which samples are needed for the approximation error approach, however, we present a new extension to methodology for estimating the auxiliary parameters, \( z \).

### 4.4 A Novel Extension of the Bayesian Approximation Error Approach to Estimate Auxiliary Unknowns

The new extension to the BAE approach developed here pertain to the subsequent estimation of the auxiliary unknowns, \( z \), which can often be of interest. With the standard BAE approach one can compute the MAP estimate for \( z \) as

\[
\nu_{\text{MAP}} = \max_z \pi(z | \nu = \nu_{\text{MAP}}), \quad \text{where} \quad \nu_{\text{MAP}} = \mathcal{D} - A(x_{\text{MAP}}, z_0). \tag{4.4.1}
\]

Such an approach has been implemented effectively to estimate auxiliary parameters. Worthy of special note is reconstruction of the domain boundary in electrical impedance tomography (EIT) where the boundary is unknown \[28\]. However, this approach may fail if the joint distribution of \( x \) and \( z \) cannot be well approximated by a Gaussian. In sections 5.2 and 5.4 we give examples where this is the case, and show that the estimates attained are misleading.

In an effort to improve the estimation of these auxiliary unknowns we consider a set of \( s \in \mathbb{N} \) nonlinear mappings

\[
F_i : \mathbb{R}^M \rightarrow \mathbb{R}^{m_i}, \quad m_i \leq M, \quad i = 1, 2, \ldots, s
\]

which act on the noisy residual \( \nu \). Each \( F_i(\nu) \) is then a random variable, and therefore there exists marginal probability distributions \( \pi(z | F) \). These mappings acting on \( \nu_{\text{MAP}} \) will be of particular significance, and we introduce the following notation,

\[
F_i^* := F_i(\nu_{\text{MAP}}) = F_i(\mathcal{D} - A_{z_0}(x_{\text{MAP}}) + e) \quad i = 1, 2, \ldots, s, \quad \text{and} \quad F^* = F(\nu_{\text{MAP}}).
\]

By using these mappings we hope to achieve a joint distribution of \( x \) and \( F \) that is well
approximated by a Gaussian. The updated MAP estimate for $z$ is

$$z_{\text{MAP}} = \max_z \pi(z|F = F^*).$$

We approximate the marginal density $\pi(z|F)$ as Gaussian, so that by theorem 4.2.3 we have

$$z_{\text{MAP}} = z_0 - P_{11}^{-1} P_{12} (F^* - F_0), \quad (4.4.2)$$

with the precision matrix $P$ defined as

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P_{21} & P_{22} \end{bmatrix} = \begin{bmatrix} \Gamma_{zz} & \Gamma_{zF} \\ \Gamma_{Fz} & \Gamma_{FF} \end{bmatrix}^{-1}, \quad (4.4.3)$$

and $F_0$ is the mean of $F$. All blocks of the covariance matrix in (4.4.3) along with the mean $F_0$ are constructed using sample statistics from augmented samples,

$$\begin{bmatrix} \bar{x}^{(\ell)} \\ z^{(\ell)} \\ \epsilon^{(\ell)} \\ F_1(\epsilon^{(\ell)}) \\ F_2(\epsilon^{(\ell)}) \\ \vdots \\ F_s(\epsilon^{(\ell)}) \end{bmatrix}, \; \ell = 1, 2, \ldots, q. \quad (4.4.4)$$

Notice that we carry out our samples of $F$ on the noise free residuals, that is, we apply $F$ only to the approximation error $\epsilon$. This is discussed more in section 4.4.2.

The assumption that each $F_i$ is nonlinear is natural, as any linear mapping fails to give any more insight on $z$ than simply using $\nu$. We give a proof of this result using only linear algebra.

**Theorem 4.4.1.** Let $z \in \mathbb{R}^n$ and $F(\epsilon) = F\epsilon$ be a linear mapping from $\mathbb{R}^M$ to $\mathbb{R}^m$ with $m \leq M$ and $\text{rank}(F) \leq M$. Assume the densities $\pi(z)$, $\pi(F)$, $\pi(z, F)$, and $\pi(z|F)$ are normal, and $\Gamma_z$ and $\Gamma_\epsilon$ are invertible. Then in the sense of quadratic forms

$$\Gamma_{z|F} \geq \Gamma_{z|\epsilon},$$

that is to say, conditioning $z$ on $F\epsilon$ is at best as useful as conditioning $z$ on $\epsilon$ only.

**Proof.** Suppose $z$ and $F\epsilon$ have means $z_0$ and $F_0 = F\epsilon_0$ and covariance matrices $\Gamma_z$ and $\Gamma_F = F\Gamma_\epsilon F^T$ respectively, and associated joint distribution

$$\begin{bmatrix} z \\ F\epsilon \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} z_0 \\ F_0 \end{bmatrix}, \begin{bmatrix} \Gamma_z & \Gamma_{zF} \\ \Gamma_{Fz} & \Gamma_F \end{bmatrix} \right).$$

Let $\text{rank}(F) \leq M$ and hence $\text{rank}(\Gamma_F) = r \leq m$, that is, the covariance matrix $\Gamma_F$ may only
be positive semidefinite. In any case, $\Gamma_F$ has a singular value decomposition,

$$\Gamma_F = V \Lambda V^T = \left( V \Lambda^{\frac{1}{2}} \right) \left( V \Lambda^{\frac{1}{2}} \right)^T = BB^T \in \mathbb{R}^{m \times m},$$

where $B, V \in \mathbb{R}^{m \times r}$ and $\Lambda \in \mathbb{R}^{r \times r}$.

In the case that $r < m$ we have $\Gamma_F$ singular, which is problematic in the construction of the conditional covariance matrix. To this end consider the substitution

$$\mathcal{F} = \Lambda^{-\frac{1}{2}} V^T (F_\epsilon - F_0) = \Lambda^{-\frac{1}{2}} V^T (\epsilon - \epsilon_0) = \Lambda^{-\frac{1}{2}} V^T F,$$

so that

$$\mathcal{F} \sim \mathcal{N}(0, I_r).$$

The joint distribution of $z$ and $\mathcal{F}$ is then

$$\begin{bmatrix} z \\ \mathcal{F} \end{bmatrix} \sim \mathcal{N} \left( \begin{bmatrix} z_0 \\ 0 \end{bmatrix} , \begin{bmatrix} \Gamma_{zz} & \mathbb{E} \left( Z \mathcal{F}^T V \Lambda^{-\frac{1}{2}} \right) \\ \mathbb{E} \left( \Lambda^{-\frac{1}{2}} V^T F Z \right) & I_r \end{bmatrix} \right),$$

where $Z = z - z_0$. Then by remark 4.2.4 we have

$$\Gamma_{z|F} = \Gamma_{z|\mathcal{F}} = \Gamma_z - \mathbb{E} (Z \mathcal{F}^T V \Lambda^{\frac{1}{2}}) I_r \mathbb{E} \left( \Lambda^{\frac{1}{2}} V^T F Z \right)$$

$$= \Gamma_z - \mathbb{E} (Z \mathcal{F}^T V A^{\frac{1}{2}}) \mathbb{E} \left( A^{\frac{1}{2}} V^T F Z \right)$$

$$= \Gamma_z - \mathbb{E} (Z \mathcal{F}^T) \Gamma^{\dagger}_F \mathbb{E} (F Z^T).$$

We will now show that the posterior covariance can only increase or stay constant (in terms of quadratic forms) when conditioning on $F_\epsilon$ rather than on $\epsilon$. Initially, suppose $F \in \mathbb{R}^{M \times M}$ and $\text{rank}(F) = \text{rank}(\Gamma_F) = M$, then $\Gamma_F$ is invertible with $\Gamma^{-1}_F = V \Lambda^{-1} V^T$. Let $E = \epsilon - \epsilon_0$, then we have

$$\Gamma_{z|F} = \Gamma_z - \mathbb{E} \left( Z \mathcal{F}^T \right) \Gamma^{\dagger}_F \mathbb{E} \left( F Z^T \right)$$

$$= \Gamma_z - \mathbb{E} \left( Z E^T \right) F^T F^{-1} F^{-1} F \mathbb{E} \left( E Z^T \right)$$

$$= \Gamma_z - \mathbb{E} \left( Z E^T \right) \Gamma^{\dagger}_\epsilon \mathbb{E} \left( E Z^T \right)$$

$$= \Gamma_z - \Gamma_z \Gamma^{\dagger}_\epsilon \Gamma_{\epsilon|z} = \Gamma_{z|\epsilon}. \quad (4.4.5)$$

Now, suppose $F \in \mathbb{R}^{m \times M}$ and $\text{rank}(\Gamma_F) = r < m \leq M$. To complete the proof we must show, in terms of quadratic forms,

$$\Gamma_{z|F} \Gamma^{\dagger}_\epsilon \Gamma_{\epsilon|z} \geq \Gamma_{z|F} \Gamma^{\dagger}_F \Gamma_{F|z},$$

$$= \Gamma_{z|F} F^T \left( \Gamma^{\dagger}_\epsilon F^T \right)^\dagger F \Gamma_{\epsilon|z},$$
which is equivalent to showing, in terms of quadratic forms,

\[ \Gamma^{-1}_\epsilon \geq F^T (F \Gamma_\epsilon F^T)^\dagger F. \] (4.4.6)

The covariance matrix \( \Gamma_\epsilon \) is positive definite, and hence we have a singular value decomposition

\[ \Gamma_\epsilon = V \mathcal{D} V^T, \]

with \( V, \mathcal{D} \in \mathbb{R}^{M \times M} \) and \( \mathcal{D} \) a diagonal matrix with positive values on the diagonal. We will also make use of the singular value decomposition of \( F \),

\[ F = UDV^T \]

where \( U \in \mathbb{R}^{M \times r}, \mathcal{D} \in \mathbb{R}^{r \times r} \) and \( V \in \mathbb{R}^{M \times r} \), with \( \mathcal{D} \) a diagonal matrix with positive values on the diagonal. Substituting these singular value decompositions into equation (4.4.6) we proceed,

\[ \Gamma^{-1}_\epsilon \geq F^T (F \Gamma_\epsilon F^T)^\dagger F \]

\[ \iff V \mathcal{D}^\dagger V = VDU^T (UDV^T VDU^T VDU^T)^\dagger UDV^T \]

\[ \iff I_M \geq \mathcal{D}_2 \mathcal{V}^T (UDV^T VDU^T)^\dagger UDV^T \]

\[ \iff I_M \geq C^T (CC^T)^\dagger C, \quad \text{with} \quad C = UDV^T \mathcal{D}^{\frac{1}{2}}. \]

Now we will make use of the singular value decomposition of \( C \),

\[ C = UDV^T, \]

where \( U \in \mathbb{R}^{M \times r}, D \in \mathbb{R}^{r \times r} \) and \( V \in \mathbb{R}^{r \times M} \), with \( D \) a diagonal matrix with positive values on the diagonal. Then we continue,

\[ C^T (CC^T)^\dagger C = VDU^T (UDV^T VDU^T)^\dagger UDV^T \]

\[ = VDU^T (UD^2 U^T)^\dagger UDV^T \]

\[ = VDU^T UDV^T \]

\[ = VDV^T = VI_M V^T, \]

which is the eigenvalue decomposition of a positive semi-definite matrix with \( r \) eigenvalues of 1 and \( k - r \) eigenvalues of 0, hence

\[ I_M \geq C^T (CC^T)^\dagger C, \]

and the proof is complete. \( \Box \)

We have not proven that there exists a mapping suitable in estimating auxiliary param-
eters. We have only shown such a mapping must be nonlinear to be useful. The mappings are problem specific. We make explicit the fact that if we take only the identity map, $F(\nu) = I(\nu) = \nu$, then we arrive at the typical Bayesian approximation error approach for estimating auxiliary unknowns. The selection of suitable mappings and the pre-sampling of such features are closely related to the field of pattern recognition \cite{133}. More precisely, the ideas of feature selection and learning are closely related to the outlined method. We now discuss the similarities.

4.4.1 Links with Pattern Recognition and Feature Selection

The ensemble generated for approximation error approach could be viewed as set of learning or training samples, terms ubiquitous in machine learning and pattern recognition. These fields are largely concerned with classifying objects based on (indirect) observations \cite{133}, and are effectively used in many fields, from image and speech recognition to medical diagnosis and statistical arbitrage detection \cite{134}, and importantly, in parameter estimation.

It will turn out that the pattern recognition approach will be a similar approach to the extended Bayesian approximation error approach previously developed for estimating auxiliary parameters. We will concern ourselves with the development of the set of mappings, or so called features, in the pattern recognition framework. We consider each feature as simply a mapping from the measurement space to the feature space, $F : \mathbb{R}^M \to \mathbb{R}^m$, with $m \leq M$, which is in line with \cite{133}. Such an approach is used so that important features can extracted from the data that might otherwise go unnoticed. As in the previous section we may have an arbitrary number of features, however we limit the discussion to just a single feature for brevity. Developing these features so as to extract important characteristics from the observations is termed feature extraction. We now introduce two associated definitions with regard to feature extraction.

**Definition 4.4.2.** A feature $F$ is relevant if

$$\mathbb{P}\left[\pi(z|\epsilon, F(\epsilon)) = \pi(z|\epsilon)\right] = 0,$$

where $\mathbb{P}$ denotes probability. That is $F$ is relevant if the distribution of $z$ conditioned on $\epsilon$ and the distribution of $z$ conditioned on $\epsilon$ and $F(\epsilon)$ are equal only on sets of measure 0.

A feature $F$ is irrelevant if it is not relevant.

This definition is comparative to the widely used definitions of strong relevance in the literature, see texts such as \cite{135}.

**Corollary 4.4.3.** Under the normality assumptions of theorem 4.4.1, any linear feature is irrelevant.

**Proof.** This is an immediate consequence of theorem 4.4.1. \qed

A key idea in pattern recognition is to infer some parameter from indirect noisy measurements of the parameter, which is similar to an inverse problem. However, the approach to
the problem is quite different. Inferring the parameter is generally carried out by comparing the data in the feature space, to the outputs predicted by $A$ generated from a set of known parameter. One of the most widely used algorithm for such a task are the nearest neighbour algorithms.

**The Nearest Neighbour Algorithm**

The $(k)$-nearest neighbour (kNN) algorithm is one of the most common techniques used in pattern recognition. The idea is to classify an unknown parameter based on the nearest training sample(s) in the feature space. We briefly outline the process here.

The training set is simply an ensemble of $q$ samples drawn from the prior $\pi(x, z)$ which we denote $(x^{(\ell)}, z^{(\ell)})$. Given our ensemble of $q$ samples, in the general setting, there is some mapping to the measurements $A: \mathcal{H}_1 \rightarrow \mathcal{H}_2$, where $\mathcal{H}_2$ is called the measurement space. It is here that we apply our features, with each feature mapping the measurement space to an associated feature space, denoted $\mathbb{F}$. This process is known as training and is outlined in Figure 4.4.1.

The general problem is given data, $\mathcal{D} = A(x, z) + e$, we wish to infer $(x, z)$ (in our case we assume $x$ and $z$ are independent, and that we have already found $x = x_{\text{MAP}}$, with $\mathcal{D} = A(x_{\text{MAP}}, z)$, and we would use the method only to infer $z$). Then the crux of the nearest neighbour algorithm is to assume find the closest prediction in the measurement space, say $F(A(x, z))$, and then take $(x, y)$ as $(x, z)$. That is, let

$$(x, z) = \min_{(x^{(\ell)}, z^{(\ell)})} \left\| F(\mathcal{D}) - F\left(A\left(x^{(\ell)}, z^{(\ell)}\right)\right) \right\|_W \quad \ell = 1, 2, \ldots, q,$$

then make the approximation $(x, z) \approx (x, z)$, see Figure 4.4.1. The choice of weighting on the norm has spurred much discussion and is often chosen as the *Mahalanobis norm*, that is $W = \Gamma_{x,z}^{-1}$, the precision matrix of $x$ and $z$. This norm gives a good frame of reference for distance as it takes correlation into account [136] and can also be seen as a *whitening operator* [22].

The accuracy of the approximation $(x, z) = (x, z)$, is usually proportional to number of training samples taken, meaning this approach can be a costly exercise. Our numerical experiments suggest taking the BAE approach with the addition of the features, $z_{\text{MAP}} = \max_z \pi(z | F = F^*)$, is far more accurate than the nearest neighbour for any reasonable number of samples. We also remark that more than one nearest neighbour can be taken into account (hence the name $k$-nearest neighbour), something very common in classification problems.

**4.4.2 A Note on the Effects of Noise**

Samples of our features in (4.4.4) are noiseless, that is, we only sample $F(\epsilon)$ rather than $F(\nu)$. This is done for two reasons:

- Changing the noise model does not require us to take a new set of samples.
Figure 4.4.1: The spaces involved with feature selection and nearest neighbour. The grey dots represent the training set of samples. The red dot in $H_2$ denotes the measured data, while the red dot in $F$ denotes the value of the data once the feature(s) have been applied. The solid black dot in $F$ is $F(A(x, z))$, and represents the nearest neighbour in the feature space, thus we make the approximation $(x, y) = (x, z)$.

- In all cases investigated the results have been preferable when compared to those in which noise was added at the sampling stage. We mention when taking the nearest neighbour approach, the weighting $W$ would need to be updated accordingly with the introduction of noise.

However, when possible, we should attempt to find features which are robust to the effects of noise, so that

$$
\int_{-\infty}^{\infty} F_i(\nu(e)) \pi(e) \, de = \mathbb{E}_e (F_i(\nu)) \approx F_i(e).
$$

That is, features should be chosen to be somewhat invariant to noise.

### 4.5 Selection of a Prior Model

Selection of a suitable prior model for $x$ is generally application based, in some application of EIT we can assume the conductivity to be fairly smooth and thus smoothness-promoting priors are commonly employed [137]. Conventional smoothness priors are implemented using gradients and hence in the finite dimensional setting the associated covariance matrices do not exist, as difference matrices have nontrivial nullspaces. That is, this type of prior density is improper (and non-informative) [22]. To this end, methods have been devised to facilitate gradient based approaches to smoothness-promoting priors which can be found in [22].

An alternative approach which has been gaining traction is the use of Whittle-Matérn priors, which can be dated back to the work of Matérn in 1947 [138]. For this approach our parameter of interest (conductivity) is modelled as

$$
x = x_0 + X,
$$
where $X$ is a centred Markov random field with the (Matérn) covariance function

$$\text{Cov} \left(X(z), X(w)\right) = \alpha^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\|z-w\|}{\beta}\right) \nu K_\nu \left(\frac{\|z-w\|}{\beta}\right), \quad z, w \in \Omega,$$

(4.5.1)

where $\Gamma$ denotes the Gamma function, $K_\nu$ is a modified Bessel function of the second kind of order $\nu$, $\nu > 0$ is the smoothness parameter, $\alpha^2$ is the marginal variance and $\beta$ is the correlation length. Correlation length shall be defined as the minimum (spatial) lag $\tau$ such that the autocorrelation function reaches the value $e^{-1}$. That is, for a given (discrete) signal $f$ of length $n$, with variance $\sigma_f$, the correlation length is given by

$$\beta = \min_\tau \{\tau : A_f(\tau) = e^{-1}\},$$

(4.5.2)

where

$$A_f(\tau) = \frac{1}{(n-\tau)\sigma_f^2} \sum_{j=1}^{n-\tau} f_j f_{j+\tau}, \quad \tau = 0, 1, \ldots, n-1.$$

(4.5.3)

Since $f$ is discrete we will in practice need to use interpolation to find $\tau$. For our purposes linear interpolation will suffice.

The choice of $e^{-1}$ to define the correlation length is consistent with the definition used in the description of rough surfaces in varying fields such as computer graphics [139] wave propagation [140], and more generally tribology [141] and nanometrology [142], the study of interacting surfaces and the study of measurements at small scales.

Random fields with these general Matérn covariances are formally solutions of the stochastic partial differential equations, see for example [23, 137]. However, in practice it is much simpler to consider a limiting case of the Whittle-Matérn prior for computations. That is the case of letting $\nu$, the smoothness parameter, tend towards infinity. In this manner we arrive at the squared exponential prior [23],

$$\text{Cov} \left(X(z), X(w)\right) = \alpha^2 \exp \left(\frac{-\|z-w\|^2}{2\beta^2}\right), \quad z, w \in \Omega.$$  

(4.5.4)

Two draws from such a prior distribution with different parameters along with the sum are shown in Figure 4.5.1

### 4.5.1 Hyperpriors

In this thesis, our parameter of interest is assumed to vary at two distinct scales. To accommodate this, let us formally define $x_L$ and $x_S$ to be the large scale component and the small scale component of $x$ respectively with

$$x = x_L + x_S.$$  

(4.5.5)
We also assign prior distributions to $x_L$ and $x_S$, which we shall denote by $\pi_{\text{prior}}^L(x_L)$ and $\pi_{\text{prior}}^S(x_S)$ respectively. Let us assume both of the components come from Gaussian distributions and are mutually independent, so that

$$x_L \sim \mathcal{N}(x_{L0}, \Gamma_L) \quad \text{and} \quad x_S \sim \mathcal{N}(0, \Gamma_S). \quad (4.5.6)$$

With the exponential prior previously defined in mind, let us define $\beta_L$ and $\beta_S$ as the correlation lengths at the large scale and fine scale respectively, and suppose $\beta_L \gg \beta_S$. Then we take the respective covariance matrices as

$$\Gamma_L = \alpha_L^2 \exp \left( -\frac{\|z - w\|^2}{2\beta_L^2} \right) \quad \text{and} \quad \Gamma_S = \alpha_S^2 \exp \left( -\frac{\|z - w\|^2}{2\beta_S^2} \right),$$

and we formulate our prior distribution for our multiscale parameter,

$$x \sim \mathcal{N}(x_{L0}, \Gamma_L + \Gamma_S), \quad (4.5.7)$$

a sample from which is shown in Figure 4.5.1.

In our setting, knowing $\alpha_S$, $\beta_S$, or both, is equivalent to knowing the variance and/or correlation length of the small scale characteristics of the conductivity. However, such information is not usually at hand, and a more physically motivated example is the case where $\alpha_S$ and $\beta_S$ are unknown. In the Bayesian setting this leads us to model $\alpha_S$ and $\beta_S$ as random variables. Taking this viewpoint implies that the prior distribution $x_S$ is determined by a realisation of $\alpha_S$ and $\beta_S$, i.e. we have a conditional prior on $\sigma_S$,

$$\pi_{\text{prior}}^S(x_S|\alpha_S, \beta_S) = \left( \frac{1}{2\pi} \right)^{\frac{d}{2}} \frac{1}{\sqrt{\det \Gamma_S}} \exp \left( -\frac{1}{2} x_S^T \Gamma_S^{-1} x_S \right),$$

that is, the prior density is conditioned on the knowledge of $\alpha_S$ and $\beta_S$. The densities $\pi(\alpha_S)$ and $\pi(\beta_S)$ are termed hyperpriors and are widely used in Bayesian inverse problems, see for example [22, 143, 144, 145, 146] among many others.

The selection of a suitable hyperprior distribution is problem specific, many different
families of distributions are used in the literature. We will limit ourselves to the use of the uniform distribution for the small scale marginal variance, $\alpha_S$, and the log-normal distribution for the small scale correlation length, $\beta_S$. We formalise this by writing

$$\alpha_S \sim \mathcal{U}(0, \bar{\alpha}_S) \quad \text{that is} \quad \pi_h(\alpha_S) = \begin{cases} \frac{1}{\bar{\alpha}_S} & 0 \leq x \leq \bar{\alpha}_S \\ 0 & \text{otherwise} \end{cases}, \quad (4.5.8)$$

where $\bar{\alpha}_S$ is the maximum value $\alpha_S$ can take. For the small scale correlation length,

$$\beta_S \sim \ln \mathcal{N}(\beta_{S0}, \sigma_{\beta_S}^2) \quad \text{that is} \quad \pi_h(\beta_S) = \frac{1}{\beta_S \sigma_{\beta_S} \sqrt{2\pi}} \exp\left(-\frac{(\ln \beta_S - \beta_{S0})^2}{2\sigma_{\beta_S}^2}\right), \quad (4.5.9)$$

where $\beta_{S0}$ is the mean of $\beta_S$ and $\sigma_{\beta_S}^2$ is the variance of $\beta_S$. Both of these marginal distributions along with the joint distribution are shown in Figure 4.5.2.

Figure 4.5.2: The marginal and joint distributions of $\alpha_S$ and $\beta_S$, with $\bar{\alpha}_S = 20$, $\beta_{S0} = -3.5$ and $\sigma_{\beta_S}^2 = 0.25$. 

83
By Bayes’ theorem (under the assumption of independence of $x_L$ and $x_S$) one could construct the posterior probability density

$$
\pi(x_L, x_S, \alpha_S, \beta_S | \mathcal{D}) \propto \pi^L_{\text{prior}}(x_L) \pi^S_{\text{prior}}(x_S | \alpha_S, \beta_S) \pi(\mathcal{D} | x_L, x_S) \pi(\alpha_S, \beta_S)
$$

and infer on the distributions of the four unknowns $x_L$, $x_S$, $\alpha_S$ and $\beta_S$ simultaneously. However such an approach can be expensive and often requires MCMC type algorithms, see [146] for details. However, this will not be our approach, as determining $\alpha_S$, $\beta_S$, and $x_S$ in this fashion would require a fine discretisation. Instead we will use the BAE approach outlined in section 4.3 to simply substitute $z = (\alpha_S, \beta_S)$ into equation (4.4.1) to quickly attain MAP estimates for $\alpha_S$ and $\beta_S$. We could in theory then determine $x_S$, which will be demonstrated in section 5.2.3. This final task will not be performed, as it would require us to refine our mesh sufficiently to reconstruct the small scale component. However, the ability to characterise $x_S$ by determining the associated amplitude, $\alpha_S$, and correlation length, $\beta_S$, in many settings is all that is required.

### 4.6 Non-linear Least Squares

This section gives a very brief outline of the optimisation techniques used to deal with non-linear least squares problems. From section 4.3, under certain assumptions we have

$$
 x_{\text{MAP}} = \min_x \left\{ \frac{1}{2} \left( \| L_1 (\mathcal{D} - A(x)) \|_2^2 + \| L_2 (x - x_0) \|_2^2 \right) \right\},
$$

(4.6.1)

where $(L_1^T L_1)$ and $(L_2^T L_2)$ are symmetric positive definite matrices. The respective interpretations are no longer of any importance, and we have done away with $\epsilon_0$ to save space, moreover, from this point on we will drop the subscript on $x$. Many methods have been devised to tackle such a problem, including but not limited limited to: line search methods, trust-region methods, and quasi-Newton methods, all of which are outlined in numerous texts, for example see [118]. Each of these classes of algorithms are useful in their own right, however we will only consider the so-called Gauss-Newton method, which is briefly outlined below.

The Gauss-Newton method is an iterative algorithm which attempts to minimise problems in the specific form of (4.6.1). Given an initial estimate for $x$ the method proceeds by the iterations

$$
x^{k+1} = x^k + \alpha^k \delta^k_x,
$$

where $\delta^k_x$ is the search direction at step $k$, and $\alpha^k$ is the step size at step $k$. Thus the method boils down to finding $\delta^k_x$ and $\alpha^k$.

Let $F$ be the finite dimensional representation of a nonlinear functional, with $F : \mathbb{R}^n \rightarrow \mathbb{R}$. Then assuming $F$ is twice differentiable with respect to $x$ we have

$$
F(x + s_x) = F(x) + \nabla F(x)^T s_x + \frac{1}{2} s_x^T \nabla^2 F(x) s_x + \mathcal{O} \left( \| s_x \|^3 \right).
$$

(4.6.2)
In the iterative scheme we replace (4.6.2) by
\[
F \left( x^k + s^k \right) = F(x^k) + \nabla F(x^k)^T s^k + \frac{1}{2} s^k \nabla^2 F(x^k) s^k + O \left( \| s_x \|^3 \right) =: m_k(s_x^k).
\]
To find our search direction we minimise \( m_k(s_x) \) by setting the derivative \( \nabla m_k(s_x) = 0 \), resulting in
\[
s_x^k = - \left( \nabla^2 F(x^k) \right)^{-1} \nabla F(x^k).
\]
(4.6.3)

Let us now consider calculations of the the derivatives involved in (4.6.3) where \( F \) is the functional of (4.6.1), and (without loss of generality) assume the noise has mean zero. Then we can stack \( F \), so that
\[
F(x) = \frac{1}{2} \left\| \begin{bmatrix} L_1 (D - A(x)) \\ L_2 (x - x_0) \end{bmatrix} \right\|_2^2 = \frac{1}{2} \| b - \mathfrak{A}(x) \|_2^2 = \frac{1}{2} \| r(x) \|_2^2 = \frac{1}{2} r(x)^T r(x),
\]
with
\[
b = \begin{bmatrix} L_1 D \\ L_2 x \end{bmatrix} \quad \text{and} \quad \mathfrak{A} = \begin{bmatrix} L_1 A(x) \\ L_2 x \end{bmatrix},
\]
and \( r(x) \) is called the residual [118]. The gradient is then
\[
\nabla F(x) = \frac{1}{2} \left( \frac{\partial}{\partial x} r(x) \right)^T r(x) = \frac{1}{2} \left( \frac{\partial}{\partial x} (b - \mathfrak{A}(x)) \right)^T (b - \mathfrak{A}(x)) = -\mathfrak{J}(x)^T (b - \mathfrak{A}(x)),
\]
where \( \mathfrak{J} \) is the Jacobian matrix of \( \mathfrak{A} \). We then make the so-called Gauss-Newton approximation,
\[
\nabla^2 F(x) \approx \mathfrak{J}(x)^T \mathfrak{J}(x),
\]
and find the Gauss-Newton search direction,
\[
s_x^k = \left( \mathfrak{J}(x_k)^T \mathfrak{J}(x_k) \right)^{-1} \mathfrak{J}(x_k) (b - \mathfrak{A}(x_k)).
\]
(4.6.4)
Such an approximation is often valid [118], and negates the need for the second derivatives. This also implies the calculation of the Jacobian is the principle task. Due to the previous stacking of \( F \) the Jacobian \( \mathfrak{J} \) is comprised of two blocks, one being \( L_2 \) and the other being \( J(x) \), the Jacobian of \( A \). We already have \( L_2 \) in hand, and thus the entire problem reduces down to computing \( J(x) \), which we will call the Jacobian from here.

**Remark 4.6.1.** In the problems dealt with in this thesis we also have the extra requirement that \( x > 0 \). There have been several approaches used to implement this concept, such as the interior point methods [147, 118]. However we simply set a threshold value of \( C > 0 \) such that if at any iteration we have \( x_i \leq 0 \) for \( i = 1, 2, \ldots, N \) we shift it to \( C \), as in [148]. In this thesis we typically set \( C = 0.1 \).
4.6.1 Calculation of the Jacobian

The problem we will be dealing with in the following chapters is that of estimating the electric conductivity of a medium based on measurements of the potential. As discussed previously, the equations governing electric potential in a body are given by

\[-\nabla \cdot (\sigma(x) \nabla u(x)) = f(x) \quad \text{in } \Omega\]

\[u = g_D \quad \text{on } \Gamma_D\]

\[\sigma \frac{\partial u}{\partial \nu} = g_N \quad \text{on } \Gamma_N.\]

The boundary conditions will not be critical in what follows. What is important is to recall that solving this problem for the potential, after discretisation, amounts to solving

\[K_{\sigma} u = F\]

where \(K_{\sigma}\) is the stiffness matrix with subscript \(\sigma\) to indicate (nonlinear) dependance on \(\sigma\) and here \(u\) denotes the values of the potential at the FEM nodes. For the time being we will only consider a finite element discretisation of the problem using the piecewise linear functions outlined in chapter 3. In this case the solution to the problem, assuming we have incorporated any conditions on \(K_{\sigma}\) so the system is uniquely solvable, is

\[u = K_{\sigma}^{-1} F.\]

Let \(\mathfrak{M}\) be the measurement operator, and assume our data measurement scheme consists of point-wise measurements of the potential, that is

\[A(\sigma) = \mathfrak{M}(u) = u|_{x \in \mathcal{B}},\]

where \(\mathcal{B}\) is the set of measurement points. Thus the Jacobian matrix we require is

\[J_{ij} = \frac{\partial A_i}{\partial \sigma_j}.\]

We will construct the Jacobian column-wise, then, for a given \(i\). Since \(\mathfrak{M}\) is linear (point-wise evaluation is a linear operation) and independent of \(\sigma\), we have

\[\frac{\partial A}{\partial \sigma_j} = \mathfrak{M} \frac{\partial u}{\partial \sigma_j} = \mathfrak{M} \frac{\partial}{\partial \sigma_j} (K_{\sigma}^{-1} F) = \mathfrak{M} \left( \frac{\partial K_{\sigma}^{-1}}{\partial \sigma_j} \right) F = -\mathfrak{M} K_{\sigma}^{-1} \left( \frac{\partial K_{\sigma}}{\partial \sigma_j} \right) K_{\sigma}^{-1} F. \quad (4.6.5)\]

This is based on the fact that \(F\) is independent of \(\sigma\), and for a square invertible matrix \(K_x\), with \(x\) a one dimensional variable we have

\[\frac{\partial K_{\sigma}^{-1}}{\partial x} = -K_{\sigma}^{-1} \left( \frac{\partial K_{\sigma}}{\partial x} \right) K_{\sigma}^{-1} \quad \text{for all } \sigma.\]
By construction we have
\[ u = K^{-1}F, \]
and thus the columns of the Jacobian given in (4.6.5) can be rewritten as
\[ \frac{\partial A}{\partial \sigma_j} = -\mathcal{M}K^{-1} \left( \frac{\partial K_{\sigma}}{\partial \sigma_j} \right) u, \tag{4.6.6} \]
meaning we must only compute \( \frac{\partial K_{\sigma}}{\partial \sigma_j} \).

Computation of \( \frac{\partial K_{\sigma}}{\partial \sigma_j} \) can be carried out in a number of ways, and is based mostly on how we implement our FEM to construct \( K_{\sigma} \). We outline here the method used in this thesis, which is efficient when we need to solve the forward problem multiple times, which is typical in inverse problems. For convenience we shall now denote the stiffness matrix simply by \( K \).

We consider the construction of the stiffness matrix derived in chapter 3,
\[ K_{ij} = \int_{\Omega} \left( \sum_{k=1}^{N} \sigma_k \phi_k(x) \right) \nabla \phi_j \cdot \nabla \phi_i \, dx = \sum_{K} \int_{K} \left( \sum_{k=1}^{N} \sigma_k \phi_k(x) \right) \nabla \phi_j \cdot \nabla \phi_i \, dx, \tag{4.6.7} \]
where \( K \) denotes a single element in the triangulation. Now assume our finite element basis functions are piecewise linear and are of the type given in chapter 3, i.e.
\[ \phi_i(x_j) = \begin{cases} 1, & i = j \\ 0, & \text{otherwise,} \end{cases} \]
where \( x_j \) is the \( j \)th node of our triangulation. Then the representation of the stiffness matrix in (4.6.7) can be replaced with
\[ K_{ij} = \sum_{K} \left( \mathcal{N}_k^T \mathcal{N}_K \int_{K} \left( \sum_{k=1}^{d+1} \sigma_k \phi_k(x) \right) \, dx \right), \tag{4.6.8} \]
since \( \nabla \phi_i \), for \( i = 1, 2, \ldots, N \), is constant in each element \( K \), and
\[ \mathcal{N}_k = [\nabla \phi_{i_1}, \nabla \phi_{i_2}, \ldots, \nabla \phi_{i_{d+1}}] \in \mathbb{R}^{d \times (d+1)}, \tag{4.6.9} \]
where the basis functions \( \phi_{i_1}, \phi_{i_2}, \ldots, \phi_{i_{d+1}} \) are the \( d + 1 \) basis functions which have support in the given element \( K \). To construct \( \mathcal{N} \) each basis function is usually mapped to a local or master element where the gradients are computed. For more on this see for example [86]. Moreover we can carry out the integral involved analytically,
\[ \int_{K} \left( \sum_{k=1}^{d+1} \sigma_k \phi_k(x) \right) \, dx = C|K| \sum_{k=1}^{d+1} \sigma_k, \]
where \( C \) is a (known) constant depending only on the dimension \( d \), and \( |K| \) denotes the area (volume) of an element. For the two dimensional case we have \( C = \frac{1}{3} \), and thus the stiffness
matrix can be written
\[
K_{ij} = \frac{1}{3} \sum_{K} N_K^T N_K \left( |K| \sum_{k=1}^{d+1} \sigma_k \right).
\]

Consider now the vectorised version of \( K \), which we shall denote by \( \vec{K} \),
\[
\vec{K} = \text{vec}(K) \quad \text{that is,} \quad K_{ij} = \vec{K}_{(j-1)N+i} \in \mathbb{R}^{N^2}.
\]

Then \( \vec{K} \) can be computed as
\[
\vec{K} = P\sigma, \quad (4.6.10)
\]
where \( P \in \mathbb{R}^{N^2 \times N} \) shall be called the \textit{pre-stiffness matrix} and of which the \( k \)th column is given by
\[
P((j-1)N + i, k) = \begin{cases} \sum_{K} \frac{|K|}{3} \int_{K} \nabla \phi_j \cdot \nabla \phi_i \, dx & \text{when } \phi_i, \phi_j \text{ and } \phi_k \text{ share support on } K, \\ 0, & \text{otherwise}. \end{cases}
\]

The pre-stiffness matrix is large, however it is sparse, and by using sparse indexing means the product \( P\sigma \) computes the bare minimum number of multiplications required to calculate the stiffness matrix, that is, using any other method to compute the stiffness matrix will use at least the same number of multiplications.

There are two main advantages to assembling the stiffness matrix in this way:

1. To compute the forward problem with different values of \( \sigma \) we only require one matrix multiplication, as the pre-stiffness matrix \( P \) needs to be only calculated once for a given mesh, regardless of boundary conditions.

2. The vectorisation operation is linear and invertible, with inverse operation, \( \mathcal{M}(\cdot) \), defined as
\[
\mathcal{M}(\text{vec}(K)) = \text{vec}(\mathcal{M}(K)) = K, \quad \text{with} \quad \left( \mathcal{M}(\vec{K}) \right)_{ij} = \vec{K}_{(j-1)N+i}.
\]
Furthermore, this operation is independent of \( \sigma \) and thus
\[
\frac{\partial K}{\partial \sigma_j} = \mathcal{M} \left( \frac{\partial}{\partial \sigma_j} (\text{vec}(K)) \right) = \mathcal{M} \left( \frac{\partial}{\partial \sigma_j} (P\sigma) \right) = \mathcal{M} (P(:,j)),
\]
and thus the \( i \)th column of the full Jacobian can be written as
\[
\frac{\partial A}{\partial \sigma_j} = -\mathcal{M}^{-1} \mathcal{M} (P(:,j)) u. \quad (4.6.11)
\]
4.7 Discussion

In this chapter we have given a brief overview of inverse problems in both the deterministic setting and the Bayesian framework. We have derived the Bayesian approximation error approach established in [22] and [21]. Furthermore we have introduced a new extension to the method which can help estimate auxiliary unknowns. This extension is a computationally efficient type of post-processing designed to cope well when the joint distribution of our the approximation errors and the auxiliary parameters cannot be well approximated by a Gaussian distribution.

We also covered several concepts which need to be understood in order to carry out inversions, such as Gaussian densities, (hyper-) priors, and non-linear optimisation. In the nonlinear optimisation section we showed that construction of the Jacobian matrix was a key task. We also constructed the Jacobian of the mapping from $\sigma$ to $\mathcal{D}$ for the problem of inferring the electrical conductivity given voltage measurements when using FEM. This will be used in the subsequent chapters, and analysed in more detail in chapter 6.
In this chapter we address the problem of finding a low dimension approximation of the electrical conductivity given measurements of the potential and in the presence of a heterogeneous background medium. That is, the conductivity is multiscaled, having characteristics on both a small a large scale, such as that in Figure 5.0.1, however, in this chapter we are only interested in recovery of the large scale component. Such an example can arise in a variety of physical settings, such as biomedical imaging, geophysical imaging, or industrial imaging. Take for example the problem of locating a reinforcement bar (rebar) in a block of concrete [149]. We are not interested in locating small air or moisture pockets, or the aggregate used, we are only interested in large scale structures, such as the rebar.

Figure 5.0.1: An example of a heterogeneous background.

It will be shown that simply ignoring the small scale components of the conductivity can render the estimates for the large scale component useless. The use of an extremely fine
discretisation to resolve the small scale structures is ruled out due to time and computing constraints. We hope to demonstrate that using the Bayesian approximation error approach to account for uncertainties is an attractive option to overcome these difficulties.

After first formulating the problem, we compare results using the BAE approach against results using Bayesian methods without incorporating the approximation errors. We begin by considering the one dimensional case, results of which motivate us to consider the two dimensional case. An application of the extension to the BAE approach given in chapter 4, to recover parameters which characterise the small scale behaviour will also be explored in one and two dimensions.

Consider the problem of finding the large scale electrical conductivity of a multiscale medium by injecting a current into the medium and measuring the potential on the boundary. Such a problem is governed by the Poisson equation,

\[ -\nabla \cdot (\sigma_\varepsilon(x) \nabla u_\varepsilon(x)) = 0 \text{ in } \Omega \]
\[ \sigma_\varepsilon \frac{\partial u_\varepsilon}{\partial \nu} = g_N \text{ on } \partial \Omega, \]

where \( \varepsilon \) is simply used to denote small scale behaviour. We will also assume, as in chapter 4, that \( \sigma_\varepsilon \) can be decomposed into a large scale component \( \sigma_L \) and a small scale component, \( \sigma_S \), so that

\[ \sigma_\varepsilon(x) = \sigma_L(x) + \sigma_S(x). \]

We take the data to be voltage at the boundary, i.e. \( D = u_\varepsilon|_{\partial \Omega} \). Such a problem is a simplification of the electrical impedance tomography (EIT) problem.

As noted in chapter 4, there has been earlier work on some related ideas. Most notably, Simon [109] considered the stochastic anomaly detection problem. In this problem the background medium was taken to be randomly heterogeneous with an exponential squared prior distribution with known variance and correlation length, and the anomaly was assumed to be circular and perfectly conducting. Similarly, we will consider the background media to be randomly heterogeneous with an exponential squared prior distribution.

We will further generalise the problem by taking the assumption that we are uncertain of the amplitude and correlation length of background media. To this end we attach a uniform prior distribution for the small scale marginal variance and a log-normal prior distribution for the small scale correlation length. Unlike in [109] we make no assumption on the shape of the large scale inclusions, we simply assume the large scale conductivity is from a exponential squared prior distribution with known variance and correlation length that is larger than that of the small scale.

A similar one dimensional problem which we will first test our ideas on, is to determine the large scale part of \( \sigma_\varepsilon \), that is \( \sigma_L \), given voltage measurements, and

\[ -d \frac{d}{dx} \left( \sigma_\varepsilon(x) \frac{d}{dx} u_\varepsilon(x) \right) = f \text{ in } (0,1) \]
\[ u_\varepsilon(0) = u_\varepsilon(1) = 0. \]
In one dimension to get meaningful results data must be taken from the interior of the domain, and thus we take our data as internal voltage measurements,

\[ \mathcal{D} = u^h(1 : N), \]

where \( u^h \) denotes the finite element solution to (5.0.2). Taking all points of the potential to make up our data is convenient (and clearly overly optimistic), moreover all forward solves and inversions are carried out on the same discretisation, an obvious inverse crime [22]. We reiterate that the one dimensional case is only used as a stepping stone on the way to a more realistic case considered in two dimensions. For the remainder of the chapter we drop the subscript \( \varepsilon \) on the full multiscale conductivity.

One of the benefits of posing our problem in the Bayesian framework is that we can also compute the uncertainty (or spread) of our estimates. We show in the form of approximate standard deviation confidence intervals.

In this chapter we propose the use of the BAE approach to simultaneously compensate for the approximation errors induced by both the neglect to model any small scale structure of the conductivity, and by the use of a reduced model. We also demonstrate the application of the novel extension to the BAE approach developed in chapter 4 to recover the correlation length and amplitude of the small scale component of the conductivity.

### 5.1 The One Dimensional Case

In this section we will look at a number of qualitatively different cases and attempt to compare results obtained with and without using the BAE approach. Here we use the BAE approach to approximately marginalise over the small scale behaviour, and accordingly we treat the small scale variance, \( \alpha_S \), and small scale correlation length, \( \beta_S \), as auxiliary unknowns. We will also show the extension of the BAE approach discussed in section 4.4 can be used to estimate the small scale marginal variance and correlation length. Moreover, once estimated, the small scale parameters can be incorporated into an updated prior distribution for the entire conductivity and then used to estimate the full conductivity.

We compare four cases in this section:

1. When the conductivity is devoid of small scale feature;
2. when the conductivity has relatively small variations at the small scale;
3. when the conductivity has relatively large amplitude small scale behaviour;
4. and lastly, a more physically motivated example.

For the first of these three examples we carry out the comparisons for the noise levels of 0.1%, 1% and 5% to give a more complete overview, with covariance matrix of the noise being a
multiple of the identity matrix. Before carrying out the comparisons we first must compute the approximation errors as outlined in section 4.3.2.

### 5.1.1 Setting up the Bayesian Approximation Error approach

In order to carry out any inversions we must first attach suitable values to all prior distributions involved. This means fixing the values for \( \sigma_{L0}, \alpha_L, \beta_L, \bar{\alpha}_S, \beta_{S0} \) and \( \sigma_{\beta S}^2 \). The following values were set for coarse scale parameters,

\[
\sigma_{L0} = 300, \quad \alpha_L = 15, \quad \beta_L = 0.4, \tag{5.1.1}
\]

and for the fine scale parameters we set

\[
\bar{\alpha}_S = 20, \quad \beta_{S0} = -4.5, \quad \sigma_{\beta S}^2 = 0.7^2. \tag{5.1.2}
\]

Other than the large scale mean value, \( \sigma_{L0} \), the above values were simply selected to ensure the total conductivity \( \sigma = \sigma_L + \sigma_S \) would exhibit behaviour on two distinct scales in both correlation length and amplitude, see Figure 5.1.1 for some examples of some draws from these prior distributions.

![Figure 5.1.1: Five draws from Left: \( \pi_{prior}^L(\sigma_L) \) Right: \( \pi_{prior}^S(\sigma_S|\alpha_S, \beta_S)\pi(\alpha_S, \beta_S) \).](image)

The more challenging and physically relevant case is when one has \( \alpha_L < \bar{\alpha}_S \) and \( \beta_S < \beta_L \), that is when the small scale variations can be larger than the large scale variations, and accordingly we place most of our focus on this scenario. The value of \( \sigma_{L0} \) is sufficiently high\(^1\) to enforce the positivity of \( \sigma \) while maintaining the independence of \( \sigma_L \) and \( \sigma_S \).

With all prior distributions set we are in a position to draw all the samples required to implement the Bayesian approximation error approach. We use \( A_0(\sigma_L^{(\ell)}) \) to denote \( A(\sigma_L^{(\ell)}, 0) \), which is our less accurate model, so that our approximation errors are

\[
e^{(\ell)} = A(\sigma_L^{(\ell)}, \sigma_S^{(\ell)}) - A_0(\sigma_L^{(\ell)}), \quad \ell = 1, 2, \ldots, q,
\]

\(^1\)So that the \( P(\sigma \leq 0) \approx 0 \).
and our ensemble of samples from associated prior distributions are

\[
\begin{bmatrix}
\sigma_{L}^{(\ell)} \\
\sigma_{S}^{(\ell)} \\
\alpha_{S}^{(\ell)} \\
\beta_{S}^{(\ell)} \\
\epsilon^{(\ell)}
\end{bmatrix}, \quad \ell = 1, 2, \ldots, q. \tag{5.1.3}
\]

We then make the Gaussian approximation, so that

\[
\begin{bmatrix}
\sigma_{L} \\
\sigma_{S} \\
\alpha_{S} \\
\beta_{S} \\
\epsilon
\end{bmatrix}
\sim N
\begin{bmatrix}
\sigma_{L0} \\
0 \\
\alpha_{S0} \\
\beta_{S0} \\
\epsilon_{0}
\end{bmatrix},
\begin{bmatrix}
\Gamma_{\sigma_{L}} & 0 & 0 & 0 & \Gamma_{\sigma_{L}\epsilon} \\
0 & \Gamma_{\sigma_{S}} & \Gamma_{\sigma_{S}\alpha_{S}} & \Gamma_{\sigma_{S}\beta_{S}} & \Gamma_{\sigma_{S}\epsilon} \\
0 & \Gamma_{\alpha_{S}\sigma_{S}} & \Gamma_{\alpha_{S}} & 0 & \Gamma_{\alpha_{S}\epsilon} \\
0 & \Gamma_{\beta_{S}\sigma_{S}} & 0 & \Gamma_{\beta_{S}} & \Gamma_{\beta_{S}\epsilon} \\
0 & \Gamma_{\epsilon\sigma_{L}} & \Gamma_{\epsilon\alpha_{S}} & \Gamma_{\epsilon\beta_{S}} & \Gamma_{\epsilon}
\end{bmatrix}.
\tag{5.1.4}
\]

With all necessary covariance matrices other than the covariance of the noise constructed from the samples. Figure 5.1.2 shows the statistics of the approximation error when the prior distributions are those described by the parameters given in (5.1.1) and (5.1.2), along with the varying noise levels. It is important to note that when the noise level is = 5% the noise dominates the approximation errors with regard to definition 4.3.13.

![Figure 5.1.2: Statistics of the approximation error.](image)

With all necessary covariance matrices other than the covariance of the noise constructed from the samples. Figure 5.1.2 shows the statistics of the approximation error when the prior distributions are those described by the parameters given in (5.1.1) and (5.1.2), along with the varying noise levels. It is important to note that when the noise level is = 5% the noise dominates the approximation errors with regard to definition 4.3.13.

We now have all the components needed to carry out inversions using the Bayesian approximation error approach.

### 5.1.2 Case 1: No Small Scale Behaviour

We first compare results for the case in which \( \sigma = \sigma_{L} \), that is the conductivity is entirely characterised on the large scale. This scenario can in fact be further decomposed into two distinct situations:
We assume a priori there is no small scale component to $\sigma$

We assume a priori there is small a scale component to $\sigma$

The first situation requires little attention: No small scale implies $\alpha_S = 0$ which implies $\epsilon = 0$ and $\Gamma_\epsilon = 0$. Consequently the results when using the BAE approach are identical to those computed without using the BAE approach.

Suppose then, our prior belief is that the conductivity will have small scale characteristics yet it turns out that $\sigma$, the true conductivity, is in fact devoid of any small scale behaviour. Thus we carry out our approximation error samples using the prior distributions described by the parameters given in (5.1.1) and (5.1.2). By this approach our model will allow for the presence of small scale behaviour, even though there is no small scale characteristics present.

In this case the MAP estimate obtained using the BAE approach are basically identical to the MAP estimates given using the standard noise model for all noise levels, see Figure 5.1.3. In all cases the true conductivity is well contained in the posterior distribution. However, it is clear that taking the approximation errors into account results in a significant increase in the approximate posterior variance estimates. The increased variance is simply allowing for small scale features (which are void in this case).

5.1.3 Case 2: Small Amplitude Small Scale Behaviour

We now compare results for the case in which $\alpha_S \leq \alpha_L$. With a noise level of 0.1% the MAP estimate obtained without using the BAE approach appears reasonable, but is in fact misleading: The true value is generally not contained well within the posterior error bars, see Figure 5.1.4. This is because the errors due to the small scale cannot be accounted for using the standard noise model. On the other hand, the MAP estimate given using the BAE approach results in a similar estimate, and moreover, the true large scale conductivity is well contained in the posterior error bars. For the higher noise levels both methods give similar estimates, with the BAE approach resulting in larger uncertainty bars as expected.

5.1.4 Case 3: Large Amplitude Small Scale Behaviour

In this case we take $\alpha_S > \alpha_L$, that is the amplitude of the small scale variations are larger than the variations at the large scale. In this case the small scale features can have a substantial effect on our estimates of the large scale conductivity. We see that in the case of 0.1% or 1% noise level using the standard noise model can result in a misleading posterior estimates for $\sigma_{L_{MAP}}$, the true large scale conductivity typically being far from the posterior distribution, see Figure 5.1.5.

On the other hand, with the Bayesian approximation error approach for these noise levels the small scale effects on $\sigma_{L_{MAP}}$ are minimal, with the true large scale conductivity being very well contained in the posterior distribution, see Figure 5.1.5.

For a noise level of 5% both estimates are similar meaning the approximation errors are dominated by the noise. We see the posterior variance in both cases is large, i.e. there is a
large amount of uncertainty is the estimates.

5.1.5 Case 4: A Slightly Different Example

In this section we look to a slightly more physically motivated example: Consider a simplified (one dimensional) problem of locating mineral deposits beneath the earth’s surface. Such a problem is sometimes referred to as subsurface resistivity imaging, and is related to electric resistivity tomography, see [150]. In this setting, the large scale prior distribution encodes what we expect to find (the range of conductivities/resistivities we expect), and the likely
Chapter 5. Estimation of the Large Scale Component

Figure 5.1.4: Case 2 results

Left Column: (top to bottom) $\sigma_{\text{LMAP}}$ using standard noise model for noise levels of 0.1\%; 1\%; 5\%, as well as 1, 2 and 3 approximate standard deviation bars.

Right: (top to bottom) $\sigma_{\text{LMAP}}$ using the BAE approach for noise levels of 0.1\%; 1\%; 5\%, as well as 1, 2 and 3 approximate standard deviation bars.

sizes of any deposits. While the small scale prior distribution would account for the fact that in most cases the conductivity of earth near its surface, the near surface, is far from homogeneous, small mineral deposits and other nonhomogenieties are almost always present.

Suppose a large deposit of iron exists (which we wish to find), but the surrounding clay soil also includes deposits of, say, gold and silicon, having conductivities greater than and smaller than that of iron respectively [151, 152]. We model the conductivity as,

$$\sigma = \sigma_{\text{Fe}} \chi(I_{\text{Fe}}) + \sigma_{\text{C}} \chi(I_{\text{C}}) + \sigma_{\text{Au}} \chi(I_{\text{Au}}) + \sigma_{\text{Si}} \chi(I_{\text{Si}}),$$
Figure 5.1.5: Case 3 results **Left Column:** (top to bottom) $\sigma_{\text{LMAP}}$ using standard noise model for noise levels of 0.1%; 1%; 5%, as well as 1, 2 and 3 approximate standard deviation bars **Right:** (top to bottom) $\sigma_{\text{LMAP}}$ using the BAE approach for noise levels of 0.1%; 1%; 5%, as well as 1, 2 and 3 approximate standard deviation bars.

with $\sigma_{\text{Fe}}$, $\sigma_{\text{C}}$, $\sigma_{\text{Au}}$, and $\sigma_{\text{Si}}$ being the conductivities of iron, clay, gold and silicon respectively, where $\sigma_{\text{Au}} > \sigma_{\text{Fe}} > \sigma_{\text{C}} > \sigma_{\text{Si}}$, with $I_{\text{Fe}}$, $I_{\text{C}}$, $I_{\text{Au}}$, and $I_{\text{Si}}$ denoting the (possibly disconnected) regions in which iron, clay, gold and silicon appear respectively, and $\chi$ denotes the characteristic (indicator) function.

We use the exponential squared prior for both the large and small scale distributions, although the conductivity is discontinuous. This avoids the possibility of overly optimistic estimates and illustrates the robustness of the BAE approach [132]. We set the following coarse
scale parameters,
\[ \sigma_{L0} = \sigma_C = 175, \quad \alpha_L = 30 \quad \beta_L = 0.3, \]
while the corresponding fine scale parameters we set as
\[ \tilde{\alpha}_S = 40, \quad \beta_{S0} = -4 \quad \sigma_{S}^2 = 1^2. \]

The choice \( \sigma_{L0} = \sigma_C \) is intuitive, as we should expect the majority of medium to consist of clay. We make overt the fact that the values chosen here do not coincide with the actual conductivity values of the specified materials, this example is constructed solely for the purposes of a qualitative demonstration. The other parameters were selected in a similar fashion to those used in the previous examples, but attention was given so that the large scale prior could capture variations of at least as large as \( \sigma_{Fe} - \sigma_C \) so that the small scale prior could capture variations of at least as large as \( \sigma_{Au} - \sigma_C \) and \( \sigma_{Si} - \sigma_C \), see Figure 5.1.6 for some typical samples from these priors. For this example we set
\[ \sigma_{Au} = \sigma_C + 100 \quad \sigma_{Fe} = \sigma_C + 50 \quad \sigma_{Si} = \sigma_C - 40. \]

Furthermore, we only consider a noise level of 0.2%.

![Figure 5.1.6: Five draws from Left: \( \pi_{prior}^L(\sigma_L) \) Right: \( \pi_{prior}^S(\sigma_S|\alpha_S, \beta_S)\pi_h(\alpha_S, \beta_S) \).](image)

The large scale conductivity profile was chosen to be quite a simple curve, and is shown in Figure 5.1.7, along with the total conductivity. As for the fine scale conductivity the profile was chosen ad hoc, and is meant to be void of any patterns or symmetries.

In this case simply ignoring the small scale attributes leads to large error in the reconstructions: Firstly, the value of \( \sigma_{LMAP} \) in \( I_{Fe} \) is grossly overestimated, and would likely be classified as something other than iron; secondly if \( \sigma_{LMAP} \) along with the approximated posterior confidence intervals suggest that there almost certainly exist a iron deposit where in fact there is nothing other than clay and gold; thirdly, the true coarse scale conductivity, \( \sigma_L \) lies outside the approximate posterior distribution. From a classification point of view, failure to take into account the small scale structure can lead to misclassification, see Figure 5.1.8.
Conversely, by implementing the BAE approach to accommodate the small scale structure, we see good agreement between $\sigma_{\text{LMAP}}$ and $\sigma_L$, with $\sigma_L$ being well contained within the posterior density. We do see however that in the regions $\sigma_{\text{MAP}}$ does a poor job of estimating $\sigma_L$ is regions where $\sigma_L$ is non smooth, see Figure 5.1.8. This is a direct result of the exponential squared prior, which heavily favours smooth functions. To avoid this issue, a prior which promotes jumps in the conductivity such as the total variation prior [153] could be used.

![Figure 5.1.7: Case 4 true conductivity profiles](image)

**Left:** The underlying coarse scale conductivity  
**Right:** The sum of the coarse scale and fine scale conductivities.

![Figure 5.1.8: Case 4 results](image)

**Left:** $\sigma_{\text{LMAP}}$ using standard noise model, as well as 1, 2 and 3 approximate standard deviation bars  
**Right:** $\sigma_{\text{LMAP}}$ using the BAE approach, as well as 1, 2 and 3 approximate standard deviation bars.

### 5.2 Recovery of $\alpha_S$ and $\beta_S$

In section 4.4, we showed that using the typical Bayesian approximation error approach the auxiliary parameters are estimated,

$$ z_{\text{MAP}} = \max_z \pi(z|\nu = \nu_{\text{MAP}}), $$

(5.2.1)
where \( z = (\alpha_S, \beta_S) \) and
\[
\nu_{\text{MAP}} = \mathcal{D} - A(\sigma_{\text{MAP}}, z_0) \in \mathbb{R}^M.
\] (5.2.2)

To test the ability of \( \nu \) to predict the values of \( z \) we examine the correlation matrix (Pearson product-moment correlations), \( \mathcal{P} \) [154]. Specifically we look at both the column of the correlation matrix corresponding to the correlations between \( \alpha \) and each component of \( \nu \), and the column of the correlation matrix corresponding to the correlations between \( \beta \) and each component of \( \nu \). We will also use the coefficient of multiple correlation [155], denoted \( \mathcal{R} \), to measure how well we can predict both \( \alpha \) and \( \beta \) using all components of \( \nu \) as predictors, which is analogous to the standard procedure to estimate the auxiliary unknowns given by (5.2.1). The coefficient of multiple correlation for predicting a variable \( z \) given the set of \( P \) predictors \( w = [w_1, w_2, \ldots, w_P] \) is
\[
\mathcal{R}_{z|w} = \left( c^T \mathcal{P}^{-1} c \right)^{\frac{1}{2}}, \quad 0 \leq \mathcal{R} \leq 1,
\]
where \( c = [\rho_{x_i, y}, \rho_{w_1, y}, \ldots, \rho_{w_P, y}]^T \) is the correlations between each predictor \( x_i, i = 1, 2, \ldots, N \) and \( y \), and \( \mathcal{P} \) is the correlation matrix \( \mathcal{P}_{ij} = \rho_{w_i, w_j} \). Explicitly, the coefficient of multiple correlation for predicting \( \alpha_S \) given \( \nu \) as the predictor is
\[
\mathcal{R}_{\alpha_S|\nu}^2 = \begin{bmatrix}
\rho_{\nu_1 \nu_1} & \rho_{\nu_1 \nu_2} & \cdots & \rho_{\nu_1 \nu_M} \\
\rho_{\nu_2 \nu_1} & \rho_{\nu_2 \nu_2} & \cdots & \rho_{\nu_2 \nu_M} \\
\vdots & \vdots & \ddots & \vdots \\
\rho_{\nu_M \nu_1} & \rho_{\nu_M \nu_2} & \cdots & \rho_{\nu_M \nu_M}
\end{bmatrix}
\begin{bmatrix}
\rho_{\nu_1 \alpha_S} \\
\rho_{\nu_2 \alpha_S} \\
\vdots \\
\rho_{\nu_M \alpha_S}
\end{bmatrix}.
\]
Geometrically, \( \mathcal{R} \) is a measure of how close to the true value of \( \alpha_S \) (resp. \( \beta_S \)) one can predict by taking linear (or affine) combinations of the set of predictors, \( \nu \). For more on the geometric reasoning among others [156, 155]. For a perfect predictor one has \( \mathcal{R} = 1 \), while in the case \( \mathcal{R} = 0 \) the predictor set gives no insight into the variable of interest.

The coefficient of multiple correlation for predicting \( \alpha_S \) given \( \nu \) and the coefficient of multiple correlation for predicting \( \beta_S \) given \( \nu \) are small, and furthermore the maximum correlation between \( \alpha_S \) and the set \( \nu_i \) for \( i = 1, 2, \ldots, M \) and between \( \beta_S \) and the set \( \nu_i \) for \( i = 1, 2, \ldots, M \) are also small. For example, when taking 10,000 samples the following values were achieved for the coefficients of multiple correlation,
\[
\mathcal{R}_{\alpha_S|\nu} = 0.5248, \quad \mathcal{R}_{\beta_S|\nu} = 0.1462
\] (5.2.3)
and for the maximal correlations,
\[
\max_i \{\rho_{\alpha_S \nu_i}\} = 0.4616, \quad \max_i \{\rho_{\beta_S \nu_i}\} = 0.0220 \quad \text{for} \ i \in 1, 2, \ldots, M.
\] (5.2.4)

To illustrate the lack of information \( \nu \) contains on \( \alpha \) and \( \beta \), we calculated \( z_{\text{MAP}} = (\alpha_{\text{MAP}}, \beta_{\text{MAP}}) \) for the same 10,000 samples in the absence of noise, the results are shown
in Figure 5.2.1. We see that the predictions are meaningless.

![Figure 5.2.1: Estimates of $\alpha_{\text{SMAP}}$ and $\beta_{\text{SMAP}}$ based on $\nu$, along with the true values. Left: $\alpha_{\text{SMAP}}$ against true values, $\alpha_S$ Right: $\beta_{\text{SMAP}}$ against true values, $\beta_S$.](image)

To summarise, $\nu$ is not a good predictor for $\alpha_S$ or $\beta_S$ in this setting. This is because the Gaussian approximation for the joint distribution of $\nu$ and $z$ is not a feasible one. In the next section we will use the extension to the Bayesian approximation error approach outlined in section 4.4 to estimate $\alpha_S$ and $\beta_S$.

### 5.2.1 Extending the Bayesian Approximation Error Approach to Find $\alpha_S$ and $\beta_S$

Recall that the extension outlined in section 4.4 was centred around the construction of a suitable set of (relevant) features of $\nu$ to predict $z = (\alpha_S, \beta_S)$. We now introduce two features, firstly

\[ F_{\alpha}(\nu) = \sum_{k=1}^{d-1} (|D_1 \nu|)_k \]  

(5.2.5)

where $D_1$ is the first difference matrix, and secondly,

\[ F_{\beta}(\nu) = \min_{\tau} \{ k : A_{D_2 \nu}(\tau) = e^{-1} \}, \]  

(5.2.6)

where $D_2$ is the second difference matrix, and $A_{D_2 \nu}(\tau)$ denotes the autocorrelation function of the second difference of $\nu$. Thus we can say, $F_\alpha$ is (approximately) proportional to the total variation of $\nu$, while $F_\beta$ is the correlation length of the second difference of $\nu$. The choice of these specific features were physically motivated. The total variation of the noisy residual gives a good measure of the level of variations. On the other hand, it is natural to use the autocorrelation of the noisy residual to estimate the correlation of the small scale. However, our numerical experiments indicated that determining the correlation of the second difference of the noisy residual in fact offered more insight. We also note, $F_\alpha, F_\beta : \mathbb{R}^d \rightarrow \mathbb{R}$.

In line with evaluating the usefulness of the standard BAE approach for estimating $\alpha_S$.
and $\beta_S$, we present coefficients of multiple correlation and the maximal correlations. Since $F_\alpha, F_\beta: \mathbb{R}^d \to \mathbb{R}$ we only have 2 correlations in total\footnote{Under the assumption the pairs $\alpha_S$ and $\beta_S$, and $F_\alpha$ and $F_\beta$ are independent}, namely the correlation between $\alpha_S$ and $F_\alpha$ and the correlation between $\beta_S$ and $F_\beta$. When taking 10,000 samples the following values were achieved,

$$R_{\alpha_S|F_\alpha} = \rho_{\alpha_S F_\alpha} = 0.9857, \quad R_{\beta_S|F_\beta} = 0.9759,$$

the predictive power of these features can be seen in Figure 5.2.2. Although the joint distributions are not Gaussian, the regression curve is linear, making estimation of the parameters easier.

5.2.2 Some Characteristics of $F_\alpha$ and $F_\beta$.

We make some interesting observations which are peculiar to the features $F_\alpha$ and $F_\beta$ chosen.

- The estimates attained from both $F_\alpha$ and $F_\beta$ are heteroscedastic, that is the variance increases for increasing values of the parameters, to accommodate for this, we allow for linearly changing variance, see Figure 5.2.3.

- In some cases $F_\alpha$ can predict values $\alpha_{\text{SMAP}} > \bar{\alpha}_S$, that is to say, $F_\alpha$ can predict $\alpha_{\text{SMAP}}$ is greater than the sharp maximum bound placed on $\alpha_S$ from the prior. In the case $\alpha_{\text{SMAP}} > \bar{\alpha}_S$ we could simply take $\alpha_{\text{SMAP}} = \bar{\alpha}_S$, while simultaneously adjusting the posterior variance, see Figure 5.2.3.

- When the small scale correlation length, $\beta_S$, is smaller than the discretisation size, $h$, that is $\beta_S < h$, the feature $F_\beta$ will generally predict $\beta_S = h$. This is inline with physical intuition: For a discrete signal we cannot find correlation lengths shorter than the discretisation.

Figure 5.2.2: Estimates of $\alpha_{\text{SMAP}}$ and $\beta_{\text{SMAP}}$ based on $F(\nu)$, along with the true values \textbf{Left}: $\alpha_{\text{SMAP}}$ against true values, $\alpha_S$ \textbf{Right}: $\beta_{\text{SMAP}}$ against true values, $\beta_S$. 
Chapter 5. Estimation of the Large Scale Component

5.2.3 Results of Extending the Bayesian Approximation Error

In this section we have used the same parameters and priors as outlined at the start of this chapter, in section 5.1.1. We apply the extension of the BAE approach outlined in section 4.4 using the features given in section 5.2.1. The results are given in Figures 5.2.4 and 5.2.5.

\[
\sigma = \sigma_L + \sigma_S \sim \mathcal{N}(\sigma_{L0}, \Gamma_L + \Gamma_S)
\]
with
\[ \Gamma_L = \alpha_L^2 \exp \left( -\frac{\| z - w \|^2}{2\beta_L^2} \right) \quad \text{and} \quad \Gamma_S = \alpha_{S\text{MAP}}^2 \exp \left( -\frac{\| z - w \|^2}{2\beta_{S\text{MAP}}^2} \right) \quad \forall z, w \in \Omega. \]

To carry out this step two things should be noted:

1. There will be additional time involved, as this stage requires an additional solution of a non-linear minimisation problem.

2. The discretisation needs to be sufficiently small to resolve the small scale conductivity.

We demonstrate the effectiveness of this method by using it to estimate the total conductivity, \( \sigma \), for the cases discussed in sections 5.1.4 and 5.1.5. The estimate for the total conductivity, \( \sigma_{\text{MAP}} \), are in good agreement with the true conductivity \( \sigma \), for the most part, and furthermore the true conductivity, \( \sigma \), is well contained in the posterior distribution, see Figure 5.2.6.

Figure 5.2.6: Estimates of the full conductivity Left: For the example of 5.1.4 Right: For the example of 5.1.5.
5.3 The Two Dimensional Case

The results obtained in the one dimensional setting motivate us to consider the two dimensional problem. The problem we consider was outlined at the start of the chapter in 5.0.1, and is related to EIT problem. We also give the problem a physical setting: Suppose we wish to assess the quality of a concrete block or locate (relatively) large internal components. In the first case large air or moisture pockets, or regions of unmixed aggregate would be indicative of poorly mixed and poor quality concrete, while in the second case we maybe looking for rebars. Electrical resistivity tomography (ERT) an inverse problem equivalent to EIT \cite{157} has been shown to be an attractive choice of imaging modality to tackle such a problem \cite{158, 149, 14}.

In the two dimensional example we use an underlying large scale conductivity which can be interpreted in either or both of the described scenarios, shown in Figure 5.3.1. As in the one dimensional example, we use a smoothness promoting prior while the actual underlying conductivity is discontinuous to look at the robustness of the method.

We assume we have prior knowledge on the types and rough bounds on the size of aggregates used, meaning we can place prior distributions on \( \alpha_S \) and \( \beta_S \). We interpret the problem as follows: Our large scale prior should account for any large internal structures (rebars, large moisture pockets etc), while the small scale prior should encode our prior information about the concrete mixture. In this scenario we take the (physically motivated) viewpoint that the any of the material constituting the large scale may also appear at the small scale and hence we allow the case \( \alpha_S = \alpha_L \), by setting \( \tilde{\alpha}_S > \alpha_L \), which will be outlined in section 5.3.2.

![Figure 5.3.1: Underlying large scale conductivity.](image)

In this scenario, our data will consist of point-wise potential values along the boundary. To avoid so called inverse crimes, we generate our data by solving for the potential using the linear finite element method on a mesh much finer than that on which we attempt to reconstruct the conductivity on. The mesh we use to calculate our data from will be denoted \( T_D \), and consists of 7200 elements and 4271 nodes, see Figure 5.3.2.

We will apply the BAE approach to marginalise over discretisation errors, small scale
behaviour, and combinations of both. Moreover, the amount of data used for the inversions also decreases when we use coarser discretisation, a typical set of data measurements is shown in Figure 5.3.3.

We aim to compare the results with and without the use of the BAE approach. To this end we will carry out the comparisons for noise levels of 0.05\% and 0.5\%. Such noise levels are warranted, as both levels dominate the errors between the potentials calculated on $\mathcal{T}_D$ and on the reference inversion mesh outlined below. We first present the case in which no approximation errors are present to provide us with a reference point to compare our later results.

### 5.3.1 The Reference Case

We first carry out inversions for the base case, with $\sigma$ having no small scale structure. We use $\mathcal{T}_h$ as our mesh for the inversions, shown in Figure 5.3.2, with the point values of the potential at each boundary node making up our data. The large scale prior used is the exponential
squared prior, as discussed in the previous chapter,

\[ \Gamma_L = \alpha_L^2 \exp \left( -\frac{\|z - w\|^2}{2\beta_L^2} \right), \quad \forall z, w \in \Omega \]

with \( \alpha_L = 15 \) and \( \beta_L = 0.2 \), see Figure 5.3.4 for some typical draws from this prior.

Figure 5.3.4: Several draws from \( \pi_{\text{prior}}^L(\sigma_L) \).

Throughout this section we take \( e \sim \mathcal{N}(0, \delta_e^2 I) \), with \( \delta_e = 10^{-7} \) and \( \delta_e = 10^{-6} \) corresponding to noise levels of 0.05% and 0.5% respectively. The MAP estimate, \( \sigma_{\text{MAP}} \) with the assumptions of section 4.2.3, can then be written as (inline with equation (4.3.12))

\[ \sigma_{\text{MAP}} = \min_{\sigma} \left\{ \frac{1}{\delta_e} \| (\mathcal{D} - A(\sigma) - e_0) \|^2_2 + \| L_L (\sigma_L - \sigma_{L0}) \|^2_2 \right\} \]

with \( L_L^T L_L = \Gamma_L^{-1} \), the inverse of the large scale prior covariance matrix. Minimisation is carried out using a Gauss-Newton scheme with a fixed step size of 0.1, and an initial homogeneous guess of 200.

In Figure 5.3.5, we see MAP estimate calculated in the reference case, which appears to be both qualitatively and quantitatively consistent. We remark that the squared exponential prior promotes smoothness. Consequently, in regions near discontinuities the true conductivity may fail to lie within the approximate posterior confidence intervals. However, in most other areas the true conductivity is well contained within the approximate posterior confidence intervals. For capturing such features a prior such as the total variation prior \[153\] should be used. This set of results will be used as a reference for the optimal reconstructions, as they were computed in the absence of any approximation errors. We also note the estimate took 10.9s to converge.

In Figure 5.3.5 we have also shown the projection of \( \sigma \) onto the first 200 principle components of \( \Gamma_{\sigma_L} \), to show how well the exponential squared prior could estimate the discontinuous conductivity. However, this is an overly optimistic estimate for two reasons:

1. The ratio between the first and two-hundredth singular values of \( \Gamma_{\sigma_L} \) is approximately \( 2 \times 10^7 \). That is, the chances of achieving such an estimate with the use of the exponential squared prior is less than 1 in \( 10^7 \). For more on principle component analysis (PCA) see for example \[159\].
2. There is no noise involved in carrying out the projection.

### 5.3.2 Setting up the Bayesian Approximation Error approach

The steps carried out here are only slightly more involved than in the one-dimensional case. We again assume both the large and small scale components of the conductivity are draws from exponential squared distributions, and hence

\[
\Gamma_S = \alpha_S^2 \exp \left( -\frac{\|z - w\|^2}{2\beta_S^2} \right), \quad \forall z, w \in \Omega.
\]

We take \(\alpha_S\) and \(\beta_S\) to be random variables drawn from uniform and log-normal (hyper-)prior distributions respectively, distributions which were outlined in section 4.5. Here we take

\[
\alpha_S \sim \mathcal{U}(0, 20) \quad \text{and} \quad \beta_S \sim \ln\mathcal{N}(-3.5, 1),
\]

that is \(\tilde{\alpha}_S = 20\), \(\beta_{S0} = -3.5\), and \(\sigma_{\beta_S}^2 = 0.5^2\). The reasoning behind these values is identical to the intuition used in the one-dimensional case. Figure 5.3.6 shows the distributions for \(\alpha_S\)
Table 5.3.1: Attributes of the meshes used, where $T_D$ denotes the mesh the data was generated on, $T_h$ denotes the reference inverse mesh, $T_H$ denotes the coarse mesh for inversions, and $T_{HH}$ denotes the coarsest mesh used for inversions.

<table>
<thead>
<tr>
<th>Mesh</th>
<th>Elements</th>
<th>Nodes</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T_D$</td>
<td>7200</td>
<td>3721</td>
</tr>
<tr>
<td>$T_h$</td>
<td>1800</td>
<td>961</td>
</tr>
<tr>
<td>$T_H$</td>
<td>200</td>
<td>121</td>
</tr>
<tr>
<td>$T_{H}$</td>
<td>50</td>
<td>36</td>
</tr>
</tbody>
</table>

and $\beta_S$, while Figure 5.3.7 shows some typical draws from $\pi(\sigma_S|\alpha_S, \beta_S)$.

Figure 5.3.6: The hyper-priors **Left:** $\pi_h(\alpha_S)$ **Right:** $\pi_h(\beta_S)$.

Figure 5.3.7: Several draws from $\pi_{prior}^S(\sigma_S|\alpha_S, \beta_S)$.

At this point we also fix each of the meshes and measurement points to be used for inversions. We label the coarsest mesh $T_{HH}$ and the coarse mesh $T_H$. The relevant information for the meshes used are given in Table 5.3.1 and they are shown in Figure 5.3.8.

With the prior distributions and discretisation set we carry out the sampling for the approximation error. There are 5 different scenarios we investigate, while for each case we also use noise levels of 0.05% and 0.5%.

1. Approximate marginalisation over modelling errors due to neglecting the small scale, that is, setting $\sigma_S = 0$. 
Figure 5.3.8: Coarse meshes **Left:** The coarse mesh $T_H$ **Right:** The coarsest mesh $T_{H'}$.

2. Approximate marginalisation over discretisation errors caused by using $T_H$ in place of $T_h$.

3. Approximate marginalisation over modelling and discretisation errors from neglecting the small scale along with using $T_H$ instead of $T_h$.

4. Approximate marginalisation over large discretisation errors from using $T_H$ rather than $T_h$.

5. Approximate marginalisation over modelling and large discretisation errors caused by neglecting the small scale along with using $T_{H'}$ in place of $T_h$.

The exact form of the approximation errors will be discussed in each case. However, in all cases we make the Gaussian approximation, giving

$$\begin{pmatrix}
\sigma_L \\
0 \\
\alpha_S \\
\beta_S \\
\epsilon \\
\epsilon_0
\end{pmatrix} \sim \mathcal{N}
\begin{pmatrix}
\sigma_{L0} \\
0 \\
\alpha_{S0} \\
\beta_{S0} \\
\epsilon_0 \\
\epsilon_0
\end{pmatrix},
\begin{pmatrix}
\Gamma_{\sigma_L} & 0 & 0 & 0 & \Gamma_{\sigma_L\epsilon} & 0 \\
0 & \Gamma_{\sigma_S} & \Gamma_{\sigma_S\alpha_S} & \Gamma_{\sigma_S\beta_S} & \Gamma_{\sigma_S\epsilon} & 0 \\
0 & \Gamma_{\alpha_S\sigma_S} & \Gamma_{\alpha_S\alpha_S} & 0 & \Gamma_{\alpha_S\epsilon} & 0 \\
0 & \Gamma_{\beta_S\sigma_S} & 0 & \Gamma_{\beta_S}\beta_S & \Gamma_{\beta_S\epsilon} & 0 \\
\Gamma_{\epsilon\sigma_L} & \Gamma_{\epsilon\sigma_S} & \Gamma_{\epsilon\alpha_S} & \Gamma_{\epsilon\beta_S} & \Gamma_\epsilon & 0 \\
0 & 0 & 0 & 0 & 0 & \Gamma_\epsilon
\end{pmatrix}, \quad (5.3.1)
$$

with all necessary covariance matrices other than the covariance of the noise constructed from the samples. Due to the symmetry of the boundary conditions and the fact that our approximation errors are equally distributed the approximation error is also symmetric. The statistics of the approximation error measured along the boundary $y = 0$ are shown in Figure 5.3.9.

In all cases, the underlying coarse scale conductivity is that shown in Figure 5.3.10. Moreover, in all cases in which we approximately marginalise over the small scale structures using the BAE approach, we use the multiscale conductivity shown in Figure 5.3.10.
5.3.3 Case 1

Initially we investigate what effect neglecting the small scale component of the conductivity has on our estimates. To this end we take our approximation error samples as

\[ \epsilon^{(\ell)} = \bar{A}(\bar{\sigma}_L^{(\ell)} + \bar{\sigma}_S^{(\ell)}) - \bar{A}_0(\sigma_L^{(\ell)}), \quad \ell = 1, 2, \ldots, q, \]

where \( \bar{\sigma}_L \) and \( \bar{\sigma}_S \) denote the more accurate finite element representations of \( \sigma_L \) and \( \sigma_S \) on \( T_h \) respectively, \( \bar{A}(\bar{\sigma}_L^{(\ell)}, \bar{\sigma}_S^{(\ell)}) \) denotes the more accurate model (computed on \( T_h \)) and \( \bar{A}_0(\sigma_L^{(\ell)}) \) denotes the accurate model with \( \sigma_S \) fixed at \( \sigma_S = 0 \) with a noise level of 0.05%. The inversions are initially carried out on \( T_h \), so that any effects of the approximation errors are due entirely to the fact that we do not account for the small scale behaviour of the conductivity.

In Figure 5.3.11 we show the MAP estimate we recover using the standard noise model (which took 10.9s to converge) and the MAP estimate attained using the approximation error approach (which took 11.0s to converge). The estimated conductivity using the standard noise model is infeasible. Moreover, the approximated confidence intervals fail to capture
the true conductivity in most places. Conversely, the MAP estimate recovered using the BAE approach appears comparable to that attained in the base case. We do see an increased approximate posterior variance which allows for variations caused by the small scale structure. We may conclude then that in this setting estimates attained using the standard noise model are susceptible to large errors if the statistics of the error induced by the small scale are not accounted for. The true conductivity used to generate the data is the multiscale conductivity shown in Figure 5.3.10.

![Figure 5.3.11: Case 1 results](image)

Top Row: (left to right) The estimate, $\sigma_{MAP}$, found using the standard error model; a cross-section along dotted line with 1, 2 and 3 approximate standard deviation bars

Bottom Row: (left to right) The estimate, $\sigma_{MAP}$, found using the BAE approach; a cross-section along dotted line with 1, 2 and 3 approximate standard deviation bars.

Figure 5.3.11: Case 1 results **Top Row:** (left to right) The estimate, $\sigma_{MAP}$, found using the standard error model; a cross-section along dotted line with 1, 2 and 3 approximate standard deviation bars **Bottom Row:** (left to right) The estimate, $\sigma_{MAP}$, found using the BAE approach; a cross-section along dotted line with 1, 2 and 3 approximate standard deviation bars.

We now investigate the same case but with an increased noise level, of 0.5%. The estimates took the same amount of time to converge, however the reconstructions attained without the use of BAE are far better than the corresponding results with 0.05% noise, see Figure 5.3.12. Moreover we see wider variances in the posterior which manage to capture the true peak and trough of the conductivity in the cross-section. This is due to the fact that the increased noise level dominates the approximation error, while still allowing for an adequate amount of information to be gathered from the data. The MAP estimate achieved using BAE is generally the same when using both levels of noise.
Chapter 5. Estimation of the Large Scale Component

Figure 5.3.12: Case 1 results with increased noise **Top Row:** (left to right) The estimate, $\hat{\sigma}_{\text{MAP}}$, found using the standard error model; a cross-section along dotted line with 1, 2 and 3 approximate standard deviation bars **Bottom Row:** (left to right) The estimate, $\hat{\sigma}_{\text{MAP}}$, found using the BAE approach; a cross-section along dotted line with 1, 2 and 3 approximate standard deviation bars. The results are similar is due to the fact that the covariance matrix of the approximation errors in this case is diagonally dominant.

### 5.3.4 Case 2

In this case we investigate what effect coarsening our discretisation has on the estimates. Here we take the approximation error samples to be

$$
e^{(\ell)} = \hat{A}(\hat{\sigma}_L^{(\ell)}) - \hat{A}(\hat{\sigma}_L^{(\ell)}) , \quad \ell = 1, 2, \ldots, q,$$

where $\hat{\sigma}_L$ denotes the less accurate finite element representation of $\sigma_L$ on $T_H$, $\hat{A}(\hat{\sigma}_L^{(\ell)})$ denotes the more accurate model (computed on $T_h$), and $\hat{A}(\hat{\sigma}_L^{(\ell)})$ denotes the less accurate model (computed on $T_H$). The inversions are carried out on $T_H$, with the true conductivity used to generate the data being only the large scale conductivity shown in Figure 5.3.10, that is, $\alpha_S = 0$ when simulating the synthetics data. We also take $\alpha_S = 0$ when sampling for the approximation error. In this way we isolate the effects of the approximation errors due to discretisation errors only.
Initially we take a noise level of 0.05%. The MAP estimate we recover using the standard noise model took 0.19s to converge and appears to show some qualitative similarities to the true large scale conductivity. However, quantitatively the estimate is somewhat error ridden, with values generally too high, the peaks and the trough of the true solution being clearly overestimated. Moreover, the true conductivity fails to lie within the approximate posterior confidence intervals in most regions, see Figure 5.3.13.

On the other hand, the MAP estimate attained by incorporating the approximation errors took 0.22s to converge, with the estimate being fairly accurate qualitatively and quantitatively. Again, we see an increased posterior variance, which is typical for the BAE approach. We conclude then that using the standard noise model can lead to estimates being prone to errors if one does not take into account the statistics of the error induced by a coarser discretisation.

Figure 5.3.13: Case 2 results

Top Row: (left to right) The estimate, $\sigma_{\text{MAP}}$, found using the standard error model; a cross-section along dotted line with 1, 2 and 3 approximate standard deviation bars

Bottom Row: (left to right) The estimate, $\sigma_{\text{MAP}}$, found using the BAE approach; a cross-section along dotted line with 1, 2 and 3 approximate standard deviation bars.

Increasing the noise level to 0.5% noise appears to change the results very little, see Figure 5.3.14. The reconstructions using the BAE approach and without the use of the BAE approach are almost identical to the respective estimates with the smaller noise level. Although the
Chapter 5. Estimation of the Large Scale Component 117

posterior variance has slightly increased in the case of not using the BAE approach, the true conductivity is still well outside the error bars.

Figure 5.3.14: Case 2 results with increased noise Top Row: (left to right) The estimate, $\sigma_{\text{MAP}}$, found using the standard error model; a cross-section along dotted line with 1, 2 and 3 approximate standard deviation bars Bottom Row: (left to right) The estimate, $\sigma_{\text{MAP}}$, found using the BAE approach; a cross-section along dotted line with 1, 2 and 3 approximate standard deviation bars.

5.3.5 Case 3

Next we use BAE approach to approximately marginalise over both discretisation error and our failure to take into account the small scale structure of the conductivity. This is achieved by taking our approximation error samples as

$$
\epsilon^{(\ell)} = \hat{A}(\hat{\sigma}^{(\ell)}_L + \hat{\sigma}^{(\ell)}_S) - \hat{A}_0(\hat{\sigma}^{(\ell)}_L), \quad \ell = 1, 2, \ldots, q,
$$

where $\hat{A}_0(\hat{\sigma}^{(\ell)}_L)$ denotes the less accurate model (computed on $T_H$) with $\sigma_S$ fixed at $\sigma_S = 0$. That is, we do not attempt to take into account any small scale features of the conductivity in our model, while the data is generated using the multiscaled conductivity shown in Figure 5.3.10. Furthermore, we carry out inversions on $T_H$.

In this case, with 0.05% noise, the MAP estimate recovered using the standard noise model
again took 0.19s to converge, and is not too dissimilar to the corresponding estimate found in the previous case, see Figure 5.3.15. The peaks and troughs are again overestimated. The estimate indeed finds the region of low conductivity, however the location is incorrect.

The MAP estimate attained using the BAE approach once again took 0.22s to converge. As in the one dimensional case we see the estimate is beginning to be pulled toward the prior mean. Besides the slight underestimation of the middle of the bar of high conductivity at the top of the domain, the estimate is still qualitatively and quantitatively fairly accurate, and the true conductivity is well captured in the approximate posterior distribution.

The results using a noise level of 0.5% are very similar to the case of 0.05% noise, as the approximation errors dominate both the noise levels, and thus are not shown.

Figure 5.3.15: Case 3 results \textbf{Top Row:}(left to right) The estimate, $\sigma_{\text{MAP}}$, found using the standard error model; a cross-section along dotted line with 1, 2 and 3 approximate standard deviation bars \textbf{Bottom Row:}(left to right) The estimate, $\sigma_{\text{MAP}}$, found using the BAE approach; a cross-section along dotted line with 1, 2 and 3 approximate standard deviation bars.
5.3.6 Case 4

Now we use the BAE approach in an attempt to marginalise over major discretisation error. We take our approximation error samples as,

\[ \epsilon^{(\ell)} = \bar{A}(\sigma^{(\ell)}_L) - A(\sigma^{(\ell)}_L), \quad \ell = 1, 2, \ldots, q, \]

where \( A(\sigma^{(\ell)}_L) \) denotes the far less accurate model (computed on \( T_H \)).

To isolate the effects of the approximation errors due to the large discretisation errors only we carry out inversions on \( T_H \) while the data is generated using only the large scale conductivity shown in Figure 5.3.10. Furthermore, \( \alpha_S = 0 \) when sampling for the approximation error, that is all the samples are also devoid of any small scale structure.

In this case the MAP estimate recovered using the standard noise model took 0.02s to converge and is completely meaningless, see Figure 5.3.16. The estimate shows little resemblance to the true conductivity, while the true conductivity is generally outside the posterior error bars.

The MAP estimate attained using the approximation error approach, also taking 0.02s to converge, is qualitatively informative. Quantitatively the accuracy of the estimate has deteriorated, with the reconstruction underestimating the bar of high conductivity at the top of the domain. However, the peak and trough in the true conductivity to the bottom of the domain are still well reconstructed in the MAP estimate. We do see that the posterior variance levels have grown considerably.

We do not show the case of a noise level of 0.05%, as the results are the same.

5.3.7 Case 5

Here we attempt to use the BAE approach to approximately marginalise over both major discretisation error and our failure to take into account the small scale component of the conductivity. In this case our approximation error samples are

\[ \epsilon^{(\ell)} = \bar{A}(\sigma^{(\ell)}_L + \sigma^{(\ell)}_S) - A_0(\sigma^{(\ell)}_L), \quad \ell = 1, 2, \ldots, q, \]

where \( A_0(\sigma^{(\ell)}_L) \) denotes the far less accurate model (computed on \( T_H \)) with \( \sigma_S \) fixed at \( \sigma_S = 0 \).

The results are shown in Figure 5.3.17. Similarly to the previous case, the MAP estimate recovered using the standard noise model offers very little insight. Moreover, the MAP estimate calculated by incorporating the approximation errors has further deteriorated. It has been pulled towards the prior mean significantly. We are still able to make out the areas of high and low conductivity in the estimate, which are indeed in the correct location. Once again we do not show the estimates when using a noise level of 0.05% as the results are the same.
5.4 Recovery of $\alpha_S$ and $\beta_S$

The data in two dimensions is collected at only the boundary, while in one dimension the measurements were taken throughout the domain. However, in both cases (one and two dimensions) the data is taken along a line. This leads us to expect that using the features $F_\alpha(\nu)$ and $F_\beta(\nu)$ outlined in (5.2.5) and (5.2.6) will be suitable for inferring $\alpha_S$ and $\beta_S$ respectively in two dimensions. The result from the sampling stage look promising when conditioning on $F_\alpha(\nu)$ and $F_\beta(\nu)$, see Figure 5.4.1. For comparison we also show in Figure 5.4.1 the estimates for $\alpha_S$ and $\beta_S$ conditioned on only $\nu$.

Using these features we attempt to recover $\alpha_S$ and $\beta_S$ with a noise level of 0.05% and no discretisation errors, i.e. we carry out inversions on $T_b$. However, the MAP estimate attained using the BAE approach shows some minor anomalies near the boundaries, which leads to large values of $F_\alpha(\nu)$, resulting in misleading estimates for $\alpha_S$. On the other hand $F_\beta(\nu)$ seems to be robust against these anomalies, with the estimates for $\beta_S$ being generally quite good. In Figure 5.4.2 the true values of $\alpha_S$ and $\beta_S$ with their respective prior distributions are compared.
Chapter 5. Estimation of the Large Scale Component

Figure 5.3.17: Case 5 results **Top Row:** (left to right) The estimate, \( \sigma_{\text{MAP}} \), found using the standard error model; a cross-section along dotted line with 1, 2 and 3 approximate standard deviation bars. **Bottom Row:** (left to right) The estimate, \( \sigma_{\text{MAP}} \), found using the BAE approach; a cross-section along dotted line with 1, 2 and 3 approximate standard deviation bars.

to their respective estimates \( \alpha_{\text{SMAP}} \) and \( \beta_{\text{SMAP}} \) along with the posterior distributions.

5.5 Discussion

In some of the results given in this chapter it would appear as though simply increasing the noise level can alleviate most of the problems arising from not taking into account approximation errors. That is, in some cases it may appear as though simply taking \( c \Gamma_e \) with \( c > 1 \) as an updated noise covariance gives similar results to those attained using the BAE approach. Generally this is ill advised [22] as \( c \) would need to be chosen to be larger than \( \max \text{diag} (\Gamma_e) \). This would imply all measurements are equally corrupted by a larger amount of noise, and we may lose information carried in the data, such as the loss of multiscale structure, which will be addressed in the next chapter.

Coincidentally, the approximation error covariance matrix in all cases considered in the two dimensional setting are generally diagonally dominant, as seen in Figure 5.5.1. Moreover, the approximation error covariance matrices become more diagonally dominant when we coarsen
Figure 5.4.1: Estimates of $\alpha_{\text{SMAP}}$ and $\beta_{\text{SMAP}}$ along with the true values. **Top Row:** (left to right) $\alpha_{\text{SMAP}}$ against true values, $\alpha_S$, based on $\nu$; $\beta_{\text{SMAP}}$ against true values, $\beta_S$, based on $\nu$. **Bottom Row:** (left to right) $\alpha_{\text{SMAP}}$ against true values, $\alpha_S$, based on $F(\nu)$; $\beta_{\text{SMAP}}$ against true values, $\beta_S$, based on $F(\nu)$.

Figure 5.4.2: Comparison of MAP estimates and true values for auxiliary unknowns, along with comparison of prior and posterior distributions based on the feature extraction method for the two dimensional example. **Left:** Estimate for $\alpha_S$. **Right:** Estimate for $\beta_S$.

Our discretisation. This implies that the approximation errors become less correlated when we reduce the dimension of the problem. However, this situation is peculiar to our problem.
That is, in general the covariance of the approximation errors are not diagonally dominant, and moreover, with a coarser discretisation the correlation of the approximation errors usually increases [128].

Figure 5.5.1: The approximation error covariance matrices  **Left**: When marginalising over small scale behaviour  **Centre**: When marginalising over small scale structure and discretisation errors  **Right**: When marginalising over small scale structure and large discretisation errors.

In this chapter we have demonstrated the effectiveness of the BAE approach. The one dimensional case was first considered, where we also showed the extension to the method outlined in section 4.4 can be applied to characterise the small scale behaviour of the conductivity via correlation length and variance. We compared the BAE approach to use of the standard noise model in the two dimensional setting. The BAE approach was used to approximately marginalise over both errors resulting from neglect the small scale component of the conductivity and errors due to discretisation. This lead to consistent estimates at a heavily reduced computational cost. We saw that when the approximation errors dominated the noise, in terms of definition 4.3.1, the estimates using the standard noise model often contained artefacts, while the reconstruction obtained using the approximation error approach only slightly deteriorated.
In this chapter we provide a proof of concept for the use of multiscale finite element methods (MsFEM) with Gauss-Newton based optimisation to estimate the MAP estimate of the multiscale conductivity associated with the multiscale Poisson equation. As stated previously, in this thesis we have deemed the large number of samples required for Markov chain Monte Carlo (MCMC) methods for multiscale inverse problems prohibitive. We have instead attempted to use gradient based inversion techniques to and a local Gaussian assumption to infer a MAP estimate and approximate posterior. That said, we should note several authors have indeed carried out MCMC inversions using MsFEM. Most of these efforts have been aimed towards the inverse problem associated with subsurface flows, that is estimating subsurface permeability. Foremost has been the work of Efendiev, [160, 77, 161, 18] and more recently Parno [19]. Our approach will be significantly different, as we will focus on gradient based methods rather than sampling based inversions. It is worth noting a similar proof of concept study has been carried out for heterogeneous multiscale methods (HMM), [162, 163], although the problems considered were effectively of a very low dimension.

We will endeavour to use MsFEM to estimate the full multiscale conductivity. Such a problem could be seen as special case of a so-called large scale inverse problem. We are faced with the task of finding large dimensional parameters, with the addition of such parameters having an underlying multiscale structure.

It was shown in chapter 4, that used certain assumptions, finding the solution of an inverse, in the form of the MAP estimate, revolves around solving the minimisation of problem,

\[
\sigma_{\text{MAP}} = \min_{\sigma} \left\{ \frac{1}{2} \left( \| L_1 (D - A(\sigma)) \|_2^2 + \| L_2 (\sigma - \sigma_0) \|_2^2 \right) \right\}. \tag{6.0.1}
\]

It was also shown in section 4.6 that calculation of the Jacobian of our forward model with
respect to our parameter,
\[ J_{ij} = \frac{\partial A_i}{\partial \sigma_j}, \]
was the central ingredient to carrying out the minimisation when using the Gauss-Newton method. In this chapter we consider the calculation of the Jacobian of the forward model based on MsFEM. The calculations are largely carried out using linear algebra. To the best of our knowledge both gradient based inversions using MsFEM and construction of the associated Jacobian have not been considered before.

We consider the problem of determining the multiscale conductivity \( \sigma(x) \), given measurements of the potential \( u \), such that
\[
-\frac{d}{dx} \left( \sigma(x) \frac{d}{dx} u(x) \right) = f \quad \text{in} \ (0,1) \quad \text{(6.0.2)}
\]
\[ u(0) = u(1) = 0, \]
with \( \sigma(x) \) varying over multiple scales. We take internal voltage measurements,
\[ \mathcal{D} = u^h(1 : N), \]
where \( u^h \) denotes the multiscale finite element solution to (6.0.2). We avoid committing inverse crimes by generating our synthetic data on a more finely refined discretisation.

Once we compute the Jacobian we provide an example of the application of MsFEM to the inverse problems outlined above using the Gauss-Newton method. We also compare the results when using the standard noise model and when using the Bayesian approximation error (BAE) approach.

In this chapter we propose the use of MsFEM with gradient based methods to carry out versions for the conductivity. Our approach also entails constructing of the Jacobian matrix, which can be used to approximate posterior confidence intervals and sensitivity analysis.

\section*{6.1 The Jacobian}

In chapter 3 we showed the MsFEM solution, \( u^h \), can be written as
\[ u^h = B\alpha, \quad \alpha = \mathcal{K}^{-1} F, \]
where \( B \) is the matrix of the multiscale basis functions, \( \mathcal{K} \) is the multiscale stiffness matrix and \( F \) is a vector corresponding to any forcing terms and the implementation of boundary conditions. As our data is taken as point measurements of the potential we have
\[ A(\sigma) = \mathcal{M}(u) = \mathcal{M}u, \]
where the measurement operator $\mathcal{M}$ is linear and independent of $\sigma$, and the forward problem is solved using MsFEM. Then following similar working to that in section 4.6.1, for a fixed column of the Jacobian, we have
\[
\frac{\partial A}{\partial \sigma_j} = \mathcal{M} \left( \frac{\partial u}{\partial \sigma_j} \right) = \mathcal{M} \left( \frac{\partial \mathcal{M} (B, \mathcal{X}^{-1} F)}{\partial \sigma_j} \right) = \mathcal{M} \left( \frac{\partial B}{\partial \sigma_j} \right) + B \mathcal{X}^{-1} \left( \frac{\partial \mathcal{X}}{\partial \sigma_j} \right) u.
\] (6.1.1)

It is worth comparing the form of (6.1.1) to the Jacobian calculated in section 4.6.1: Were we to choose linear finite element basis functions not dependent on $\sigma$ in (6.1.1), we would arrive at the exact same form of the Jacobian.

We must now determine how the basis functions and the multiscale stiffness matrix change with respect to $\sigma_j$. The first of these tasks is straight forward: Recall, from chapter 3, $B = [\psi_1, \psi_2, \ldots, \psi_N]$, with each $\psi_\ell$, for $\ell = 1, 2, \ldots, N$, satisfying
\[
-\nabla \cdot (\sigma \nabla \psi_\ell) = 0 \quad \text{in } K,
\psi_\ell = \phi_\ell \quad \text{on } \partial K, \quad \forall K \in T_H, K \subset S_\ell,
\] (6.1.2)
where $S_\ell$ is the support of $\psi_\ell$. By the same working as in section 4.6.1, we have
\[
\frac{\partial B}{\partial \sigma_j} = \left[ \frac{\partial \psi_1}{\partial \sigma_j}, \frac{\partial \psi_2}{\partial \sigma_j}, \ldots, \frac{\partial \psi_N}{\partial \sigma_j} \right]
= - \begin{bmatrix}
K_1^{-1} \mathcal{M} (P_1(:,j)) \psi_1 & 0 & \cdots & 0 \\
0 & K_2^{-1} \mathcal{M} (P_2(:,j)) \psi_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & K_N^{-1} \mathcal{M} (P_N(:,j)) \psi_N
\end{bmatrix}
\] (6.1.3)
up to some permutation, where $K_\ell$ and $P_\ell$, for $\ell = 1, 2, \ldots, N$, are the the stiffness matrix and pre-stiffness matrix used in solving for $\psi_\ell$ over $S_\ell$ respectively. The fact that $\phi_\ell$ are compactly supported results in the paddings by 0. As outlined in chapter 3, the multiscale basis functions are usually solved for over each coarse element in their respective support rather than solving over their entire support. As the multiscale basis functions have Dirichlet boundary conditions then the basis functions along the boundaries exhibit no change in relation to $\sigma$. Furthermore as the triangulation $T_H$ is composed of non-overlapping elements by construction, we have
\[
\frac{\partial \psi_\ell}{\partial \sigma_j} = K_\ell^{-1} (\mathcal{M} (P_\ell(:,j))) \psi_\ell = \sum_{K \subset S_\ell} \left( K_K^{-1} (\mathcal{M} (P_K(:,j))) \psi_\ell|_K \right), \quad \ell = 1, 2, \ldots, N.
\]
That is to say, no matter if we construct the multiscale basis functions coarse element-wise or over their respective supports, we end up at (6.1.3).

We consider now the way in which the multiscale stiffness matrix changes with respect to
the conductivity. Recall the multiscale stiffness matrix is constructed

\[ X_{ij} = \sum_K \int_K \sigma \nabla \phi_i \cdot \nabla \psi_j \, dx = \sum_K (N^T M D Q). \]  

(6.1.4)

Of the matrices in (6.1.4), only \( D \) and \( Q \) depend on \( \sigma \). We again carry out the analysis for a fixed column of the Jacobian. In this manner we have

\[ \frac{\partial X}{\partial \sigma_i} = \frac{\partial}{\partial \sigma_i} \left( \sum_K (N^T M D Q) \right) = \sum_K \left( N^T M \left( \frac{\partial D}{\partial \sigma_i} \right) Q + D \left( \frac{\partial Q}{\partial \sigma_i} \right) \right). \]  

(6.1.5)

Recall from chapter 3, the (block) matrix \( Q \) is in fact made of point-wise evaluation of the multiscale basis functions,

\[ Q = \begin{bmatrix} \Psi_1 \\ \Psi_2 \\ \vdots \\ \Psi_{N_k} \end{bmatrix} \in \mathbb{R}^{(d+1)N_k \times (d+1)} \quad \text{with} \quad \Psi_{\ell}(m, n) = \psi_m(x_n), \]

for \( m, n = 1, 2, \ldots, d + 1 \) and \( \ell = 1, 2, \ldots, N_k \). Note that indices here are local, in the sense that for each \( \ell \) (corresponding to a fine scale element, \( k \)) the indices \( m \) and \( n \) must again pass through 1 to \( d + 1 \). Hence to calculate \( \frac{\partial Q}{\partial \sigma_i} \) we need to calculate \( \frac{\partial \Psi_{\ell}}{\partial \sigma_i} \) for \( \ell = 1, 2, \ldots, N_k \), with \( N_k \) denoting the number of fine elements in a coarse element, which is essentially trivial:

\[ \frac{\partial}{\partial \sigma_i} (\Psi_{\ell}(m, n)) = \frac{\partial}{\partial \sigma_i} (\psi_m(x_n)) = \frac{\partial \psi_m}{\partial \sigma_i} (x_n) = \frac{\partial B}{\partial \sigma_j} (n, m), \]

(6.1.6)

with \( \frac{\partial B}{\partial \sigma_j} \) already computed as above.

For the final step of constructing the Jacobian we consider the representation of the (diagonal block) matrix \( D \) from chapter 3:

\[ D_{\ell\ell} = \left( \frac{|J_{k_\ell}|}{6} \sum_{r=1}^{d+1} \sigma_{\ell_r} \right) I_d \quad \ell = 1, 2, \ldots, N_k \]

where \( I_d \) is the \( d \) by \( d \) identity matrix. By again considering a fixed column \( i \) of the Jacobian, we have

\[ \frac{\partial D_{\ell\ell}}{\partial \sigma_j} = \begin{cases} \frac{|J_{k_\ell}|}{6} I_d & \text{if } x_j \text{ is a corner node of element } k_\ell \\ 0_d, & \text{otherwise}, \end{cases} \]

where \( 0_d \) is the \( d \) by \( d \) zero matrix. Thus up to some permutation, and depending on the
Chapter 6. Inversions using Multiscale Finite Element Methods

The number of small scale elements \( x_j \) is a corner of,

\[
\frac{\partial D}{\partial \sigma_j} = \begin{pmatrix}
\frac{|J_{k_1}|}{6} I_d & 0 & 0 & \cdots & 0 \\
0 & \frac{|J_{k_2}|}{6} I_d & 0 & \cdots & 0 \\
0 & 0 & 0 & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
0 & 0 & 0 & \cdots & 0
\end{pmatrix}.
\]

(6.1.7)

With (6.1.6) and (6.1.7) in hand we have computed all derivatives involved in (6.1.5), and thus have computed the Jacobian.

### 6.2 Gradient Based Inversions Using MsFEM and BAE

In this section we show that MsFEM can be used with gradient based methods, namely the Gauss-Newton algorithm, to reconstruct the conductivity, \( \sigma \), in both the case that \( \sigma \) is void of multiscale structure, and when \( \sigma \) does indeed exhibit multiscale characteristics.

When the conductivity is multiscaled, we will again make the assumption that

\[
\sigma = \sigma_L + \sigma_S,
\]

where \( \sigma_L \) and \( \sigma_S \) denote the large scale and small scale components of \( \sigma \) respectively. We take

\[
\sigma_L \sim \mathcal{N}(\sigma_{L0}, \Gamma_L) \quad \text{and} \quad \sigma_S \sim \mathcal{N}(\sigma_{S0}, \Gamma_S),
\]

(6.2.1)

with

\[
\Gamma_L = \alpha^2_L \exp \left( -\frac{\|z - w\|^2}{2\beta^2_L} \right) \quad \text{and} \quad \Gamma_S = \alpha^2_S \exp \left( -\frac{\|z - w\|^2}{2\beta^2_S} \right), \quad \forall z, w \in [0, 1].
\]

(6.2.2)

The particular values for \( \alpha_L, \beta_L, \alpha_S, \) and \( \beta_S \), corresponding to the large scale amplitude and correlation length and the small scale amplitude and correlation length respectively will be given in each of the following examples, along with the respective means, \( \sigma_{L0} \) and \( \sigma_{S0} \).

In the following examples we take \( \mathcal{T}_H \) to be the division of the unit interval into 6 equal length elements, and \( \mathcal{T}_h \) to be a further refinement of \( \mathcal{T}_H \), with each coarse element containing 14 fine elements. In this way we will have 7 multiscale basis functions, with the finite element counterpart having 85 linear basis functions, that is the finite element method (FEM) inversions are carried out on \( \mathcal{T}_h \). Moreover, the data is generated using linear FEM on 337 nodes.

### 6.3 In the Absence of Small Scale Behaviour

In this section we compare reconstructions for the conductivity between the use of linear FEM and the use of MsFEM when the conductivity is devoid of any small scale features. That is
we take the prior assumption that

\[ \sigma \sim \mathcal{N}(\sigma_{L0}, \Gamma_L), \]

with \( \sigma_{L0} = 4, \alpha_L = 2 \) and \( \beta_L = 0.4 \). Draws from this distribution are shown in Figure 6.3.1.

![Figure 6.3.1: Five draws from \( \pi_{prior}^L(\sigma_L) \).](image)

We compare the results for different levels of additive noise, with the noise covariance assumed to be of the form \( \Gamma_e = \delta_e I \) and noise levels of 0.1%, 1%, 5%, 10% and 15%. In some cases we compare the estimate through the Gauss-Newton iterations and also look at how the multiscale basis functions converge as \( \sigma \) converges.

- Firstly, we consider the case of 0.1% noise. In this case using FEM gives consistent reconstructions, an very narrow confidence intervals. On the other hand, using MsFEM with 7 multiscale basis functions appears to perform poorly, see Figure 6.3.2. This is because the true conductivity varies only at the large scale, resulting in almost linear multiscale basis functions. This leads to discrepancies between the data and the simulations larger than the noise level of 0.1%. In Figure 6.3.2 we also show that to remedy this problem we can use more multiscale basis functions, resulting in estimates in line with those of FEM.

By looking at the Gauss-Newton iterations in Figure 6.3.2 we see that when we use 7 multiscale basis functions the first few iterations are consistent with the iterations when using FEM, while the later iterations differ. On the other hand, when using 22 multiscale basis functions were used the iterations appear identical.

- Next we consider the case when the noise level is 1%. In this case using FEM results in a better reconstruction again. The MsFEM estimate appears similar, however in terms of the posterior variance it is in fact a poor estimate, with the true conductivity poorly support by the posterior variance towards the centre of the interval, see Figure 6.3.3. This is most likely due to the same reasons as the previous case. The Gauss-Newton iterations in Figure 6.3.2 show both methods taking similar paths while converging,
Figure 6.3.2: Results for 0.1\% noise Top: potential and data Second Row: (left to right) $\sigma_{\text{MAP}}$ using FEM, as well as 1, 2 and 3 approximate standard deviation bars (they are very narrow); Gauss-Newton convergence Third Row: (left to right) $\sigma_{\text{MAP}}$ using MsFEM with 7 basis functions, as well as 1, 2 and 3 approximate standard deviation bars; Gauss-Newton convergence; convergence of multiscale basis functions Bottom Row: (left to right) $\sigma_L$ using MsFEM with 22 basis functions, as well as 1, 2 and 3 approximate standard deviation bars; Gauss-Newton convergence; convergence of basis functions.

however as noted earlier, the estimate computed using MsFEM does not converge to the correct value towards the centre of the domain.

- Finally, we look at the three cases of 5\%, 10\% and 15\% noise levels. The results for these three cases are shown in Figure 6.3.4. We see that using FEM and using MsFEM results in similar MAP estimates for the conductivity, and similar posterior variance estimates. We do not analyse the iterations of the reconstruction estimates as they are
Figure 6.3.3: Results for 1% noise **Top:** potential and data **Second Row:** (left to right) $\sigma_{\text{MAP}}$ using FEM, as well as 1, 2 and 3 approximate standard deviation bars; Gauss-Newton convergence **Third Row:** (left to right) $\sigma_{\text{MAP}}$ using MsFEM with 7 basis functions, as well as 1, 2 and 3 approximate standard deviation bars; Gauss-Newton convergence; convergence of multiscale basis functions.

identical. Moreover, we do not look at the convergence of the basis functions, as they are inline with the previous two cases.

For both methods, when using noise levels of more than 15% the estimates rapidly deteriorate. Approximate posterior confidence intervals grow substantially, and MAP estimates often do not resemble the true conductivity.

### 6.4 Presence of Small Scale Behaviour

In this section we compare estimates based on linear FEM compared to estimates based on MsFEM when the conductivity possesses small scale characteristics. We take the prior assumption that the small scale and large scale components are mutually independent, and

$$\sigma \sim \mathcal{N}(\sigma_0, \Gamma_{\sigma L} + \Gamma_{\sigma S}),$$
Figure 6.3.4: Results for 5%, 10% and 15% noise (left to right) Top Row: potential and data Second Row: (left to right) $\sigma_{\text{MAP}}$ using FEM, as well as 1, 2 and 3 approximate standard deviation bars Third Row: (left to right) $\sigma_{\text{MAP}}$ using MsFEM with 7 basis functions, as well as 1, 2 and 3 approximate standard deviation bars.

with $\sigma_{L0} = \sigma_{S0} = 6$, $\alpha_L = 2$, $\beta_L = 0.4$, $\alpha_S = 3$, and $\beta_S = 0.06$. Several draws from these distributions are shown in Figure 6.4.1. We also compare the results for several noise levels, namely: 0.1%, 1%, 5%, 10% and 15%.

Figure 6.4.1: Draws Left: From the large scale prior Right: From the small scale prior.
As shown in chapter 3, MsFEM is not as accurate as the associated fine scale FEM, specifically it appears as though MsFEM consistently underestimates the peak values achieved in the potential, as seen in Figure 3.7.1. To this end we apply the BAE approach to approximately marginalise over this error. To carry out the approximation error samples we introduce a new intermediate discretisation of the domain, say $\mathcal{T}_{h_I}$, with 169 nodes, and calculate our approximation errors as

$$
\epsilon^{(\ell)} = \tilde{A} \left( \tilde{\sigma}^{(\ell)} \right) - A_{MS} \left( \sigma^{(\ell)} \right), \quad \ell = 1, 2, \ldots, q,
$$

where $\tilde{A}$ the forward model calculated using FEM on $\mathcal{T}_{h_I}$, $\tilde{\sigma}$ is the representation of $\sigma$ on $\mathcal{T}_{h_I}$, $A_{MS}$ denotes the the forward model computed with MsFEM defined on $\mathcal{T}_H$ and refined to $\mathcal{T}_h$, and $\sigma$ is the projection of the conductivity projected onto $\mathcal{T}_h$. Hence our ensemble of samples from the associated prior distributions are

$$
\begin{bmatrix}
\tilde{\sigma}^{(\ell)}
\sigma^{(\ell)}
\epsilon^{(\ell)}
\end{bmatrix}, \quad \ell = 1, 2, \ldots, q, \quad (6.4.1)
$$

and by the Gaussian assumption we have

$$
\begin{bmatrix}
\tilde{\sigma} \\
\sigma \\
\epsilon
\end{bmatrix} \sim \mathcal{N}
\begin{bmatrix}
\tilde{\sigma}_0 \\
\sigma_0 \\
\epsilon_0
\end{bmatrix},
\begin{bmatrix}
\Gamma_{\tilde{\sigma}} & \Gamma_{\tilde{\sigma}\sigma} & \Gamma_{\tilde{\sigma}\epsilon} & 0 \\
\Gamma_{\sigma\tilde{\sigma}} & \Gamma_{\sigma} & \Gamma_{\sigma\epsilon} & 0 \\
\Gamma_{\epsilon\tilde{\sigma}} & \Gamma_{\epsilon\sigma} & \Gamma_{\epsilon} & 0 \\
0 & 0 & 0 & \Gamma_{\epsilon}
\end{bmatrix}, \quad (6.4.2)
$$

The approximation error statistics are shown in Figure 6.4.2, along with the noise levels. It important to note that for noise levels of 5% and 10% the noise dominates the approximation error with respect to definition 4.3.1. We see that the mean of the approximation error is positive, meaning that as stated earlier MsFEM often underapproximates the potential. We also see that the mean of the error is much larger towards the centre of the coarse elements, while the variance of the approximation error seems to be somewhat independent of this effect.

Firstly we investigate the case of 0.1% noise level, the results of which are shown in Figure 6.4.3. In this case we see that using FEM gives a consistent estimate. The true conductivity and MAP estimate being very close across the entire domain relative to the posterior variance.

On the other hand, the reconstruction using MsFEM without taking into account the approximation errors is misleading. The true conductivity is generally not contained within the posterior error bars. We also notice that towards the centre of the domain the MAP estimate has converged to an incorrect minimum very near to zero.

We remark that the area in the centre of the domain sees an increase in the approximate posterior standard deviation in all cases (for all noise levels). This inline with physical intuition, as the potential in this area is almost constant, implying there is little current flow through this region. This means we get very little information on the conductivity in this
location, hence leading to the increased in the variance.

Using MsFEM with the BAE approach results in good agreement between the MAP estimate and the true conductivity, with the true conductivity being well contained in the approximate posterior error bounds. We point out that similarly to using FEM the estimate is least accurate in the centre of the domain. However we see the posterior variance accommodates this trait and has larger uncertainty bounds in this region.

The multiscale basis functions appear to converge faster with the use of the BAE approach. Additionally when using the BAE approach with MsFEM is seems as though the converged multiscale basis functions more closely resemble linear basis functions than the converged multiscale basis functions of MsFEM with the standard error model.

Now we consider a noise level of 1%, the results of which are shown in Figure 6.4.4. In this case we also see good results using FEM or when using MsFEM with BAE: In both cases the true conductivity is well estimated by the corresponding MAP estimates, and is generally very well contained by the poster error bars. Using MsFEM without incorporation of the approximation errors produces an adequate estimate, with associated error bars generally containing the true conductivity well. Again, all methods result in larger variance towards the centre of the domain. Although the true conductivity is contained within the approximated error bars for all methods, we see that the error bars also allow for zero and negative values. This is due to the Gaussian approximation made in calculating the posterior covariance matrix, see equation (4.2.10).

The multiscale basis functions again converge faster using the BAE approach, with the converged multiscale basis functions of the BAE approach being even closer to linear FEM basis functions than those of the same approach in the previous case, see Figure 6.4.3.

Finally, we compare results with 5% and 10% noise levels. Only MAP estimates and the posterior variances are compared, as the convergence plots offer little insight. The results are given in Figure 6.4.5. All three methods result in fairly similar estimates and approximated uncertainty levels. With 5% noise the estimates have significantly deteriorated from the
Figure 6.4.3: Results for 0.1% noise Top: potential and data Second Row:(left to right) $\sigma_{\text{MAP}}$ using FEM, as well as 1, 2 and 3 approximate standard deviation bars (they are very narrow); Gauss-Newton convergence Third Row:(left to right) $\sigma_{\text{MAP}}$ using MsFEM with 7 basis functions, as well as 1, 2 and 3 approximate standard deviation bars; Gauss-Newton convergence; convergence of multiscale basis functions Bottom Row:(left to right) $\sigma_L$ using MsFEM with 7 basis functions and the BAE approach, as well as 1, 2 and 3 approximate standard deviation bars; Gauss-Newton convergence; convergence of basis functions.

1% noise case, however, the true conductivity is well supported in all approximate posterior distributions. At a noise level of 10% we see that the small scale behaviour is drowned out by the noise, and an averaged version of the conductivity appears to have been reconstructed.
Figure 6.4.4: Results for 1% noise Top: potential and data Second Row:(left to right) \( \sigma_{MAP} \) using FEM, as well as 1, 2 and 3 approximate standard deviation bars (they are very narrow); Gauss-Newton convergence Third Row:(left to right) \( \sigma_{MAP} \) using MsFEM with 7 basis functions, as well as 1, 2 and 3 approximate standard deviation bars; Gauss-Newton convergence; convergence of multiscale basis functions Bottom Row:(left to right) \( \sigma_L \) using MsFEM with 7 basis functions and the BAE approach, as well as 1, 2 and 3 approximate standard deviation bars; Gauss-Newton convergence; convergence of basis functions.
Figure 6.4.5: Results for 5% and 10% noise (left to right) **Top**: potential and data **Second Row**: $\sigma_{\text{MAP}}$ using FEM, as well as 1, 2 and 3 approximate standard deviation bars **Third Row**: $\sigma_{\text{MAP}}$ using MsFEM with 7 basis functions, as well as 1, 2 and 3 approximate standard deviation bars **Bottom Row**: $\sigma_{L}$ using MsFEM with 7 basis functions and the BAE approach, as well as 1, 2 and 3 approximate standard deviation bars.
6.5 Differences in the Jacobians

Here we examine the differences (or lack thereof) between Jacobian matrices associated with the FEM and MsFEM implementation of the problem at hand. The discussion is motivate by the similar structure of the uncertainties in the approximate posterior densities. We derived the approximate posterior covariance matrix in chapter 4 as
\[
\Gamma_{\text{post}} = (J\Gamma_e^{-1}J^T + \Gamma_x^{-1})^{-1}.
\]

The prior distribution used in all cases is identical, on the other hand, we would expect the Jacobian matrices to differ when using FEM and when using MsFEM, to this end we introduce the notation \(J_{\text{FEM}}\) and \(J_{\text{Ms}}\) to denote the Jacobian matrices associated with FEM and MsFEM respectively. Furthermore we introduce the approximate posterior covariance matrices,
\[
\Gamma_{\text{post}}^{\text{FEM}} = (J_{\text{FEM}}^T\Gamma_e^{-1}J_{\text{FEM}} + \Gamma_x^{-1})^{-1}, \quad \text{and}
\]
\[
\Gamma_{\text{post}}^{\text{Ms}} = (J_{\text{Ms}}^T\Gamma_e^{-1}J_{\text{Ms}} + \Gamma_x^{-1})^{-1},
\]
to represent the posterior covariance matrices resulting from using FEM, MsFEM and respectively. Due to the similarity in the approximate posterior confidence intervals in section 6.4, we expect \(J_{\text{FEM}} \approx J_{\text{Ms}}\), or since \(\Gamma_e = \delta_e^2 I\) we may even have \(J_{\text{FEM}}^TJ_{\text{FEM}} \approx J_{\text{Ms}}^TJ_{\text{Ms}}\). In what follows compare the Jacobians only for the case of a noise level of 1%, as the MAP estimates attained by the two methods are not too dissimilar for this noise level.

Recall the Gauss-Newton search direction given in equation (4.6.3) depends only on the Jacobian and the noise model. Although the noise models are identical, from Figure 6.4.4 we see after the first iteration of the Gauss-Newton algorithm the estimates are different, clearly \(J_{\text{FEM}}(\sigma_0)\) and \(J_{\text{Ms}}(\sigma_0)\) must be different. Indeed, both \(J_{\text{FEM}}(\sigma_0)\) and \(J_{\text{Ms}}(\sigma_0)\) are plotted as surfaces in Figure 6.5.1. We see that although the overall values are similar, \(J_{\text{Ms}}(\sigma_0)\) has a layered structure. For \(n = 1, 2, \ldots, 6\), the \(n\)th vertical band corresponds to the rate of change of the calculated potential with respect to the conductivity within the \(n\)th coarse element. This is most likely due to the fact that for a homogeneous conductivity, the basis functions are in equivalent to linear FEM basis functions.

Figure 6.5.2 shows both Jacobians evaluated at the second through fourth iterations. The banded structure of \(J_{\text{Ms}}\) quickly vanishes, as the estimate for the conductivity becomes non-homogeneous. The Jacobians also appear become more and more alike after each iteration. As the noise models are identical, and in this case \(\Gamma_e = \delta_e^2 I\) we may also be interested in how \(J_{\text{FEM}}^TJ_{\text{FEM}}\) and \(J_{\text{Ms}}^TJ_{\text{Ms}}\) vary. To this end, Figure 6.5.3 shows the absolute differences,
\[
|J_{\text{FEM}}^TJ_{\text{FEM}} - J_{\text{Ms}}^TJ_{\text{Ms}}| \quad \text{and} \quad |J_{\text{FEM}} - J_{\text{Ms}}|
\]
at the initial estimate and once the algorithm has converged. We see that once converge, there is a relatively large difference in the Jacobian localised around a certain column. It is hard to
Figure 6.5.1: Jacobians at the initial (homogeneous) estimate, $\sigma_0$  
**Left:** $J_{\text{FEM}}$  
**Right:** $J_{\text{Ms}}$.

Figure 6.5.2: Jacobians at the 2nd through 4th iteration (left to right) 
**Top Row:** $J_{\text{FEM}}$  
**Bottom Row:** $J_{\text{Ms}}$.

say why this occurs, but we point out that this column corresponds to the the derivative of all measured potential with respect to the conductivity at the point where the conductivity is closest to zero.

### 6.6 Discussion

We remark that calculation of the Jacobian matrix representing the derivative of the potential calculated using MsFEM with respect to the conductivity included finding the rate at which the multiscale basis functions change with respect to the conductivity. This immediately makes it possible to carry out sensitivity analysis [164] of the multiscale basis functions, which could be of interest for both inverse problems and forward simulations.

In this chapter we have shown that indeed it is possible to solve multiscale inverse prob-
lems using MsFEM with gradient based optimisation techniques, namely the Gauss-Newton algorithm. We were able to construct the Jacobian based on the implementation of MsFEM given in chapter 3, and briefly looked at the differences in this Jacobian, and the counterpart arising from the use of linear FEM.

We showed that the method can work both when the conductivity does not possess any multiscale characteristics, and more importantly, when the conductivity does indeed have multiscale behaviour. In the case of a multiscale conductivity we have mitigated the issue that in some cases MsFEM may underestimate the potential by incorporating approximation errors into our noise model. This modification has lead to consistent results, which are generally in line with the estimates recovered using FEM on the fine scale. Moreover by using the BAE approach the approximated posterior distribution generally well contained the true conductivity.

Figure 6.5.3: Absolute differences of Jacobians 

**Top Left:** $|J_{FEM} - J_{Ms}|$ at initial estimate  

**Top Right:** $|J_{FEM} - J_{Ms}|$ once converged  

**Bottom Left:** $|J^T_{FEM}J_{FEM} - J^T_{Ms}J_{Ms}|$ at initial estimate 

**Bottom Right:** $|J^T_{FEM}J_{FEM} - J^T_{Ms}J_{Ms}|$ once the estimates have converged.
In real world problems parameters of interest may possess multiscale behaviour, substantially adding to the difficulty of any inference of the parameters. Compounding such a problem is the fact that in many settings the smaller scale behaviour of the parameters can at best only be estimated. If these uncertainties along with any uncertainties related to discretisation are not taken into account estimates for the parameters may be rendered useless or misleading.

We have attempted to address such topics when dealing with the estimation of the spatially varying multiscale conductivity of the Poisson equation. More precisely, we have focused on presenting novel approaches to the following two problems: Recovery of a coarse scale homogenised version of the conductivity and recovery of the entire multiscale conductivity.

The Bayesian framework was used for both problems. In this framework we have treated all unknown parameters, including those corresponding to the small scale behaviour of the conductivity, as random variables. Accordingly, all inversions carried out to estimate the conductivity were based on Bayes’ theorem. Moreover, use Bayesian paradigm enabled us to decompose the multiscale conductivity into a large and small scale component, each being associated with separate prior probability distributions. The use of hyper-prior models allowed more flexibility, with the amplitude and correlation length of the small scale component also treated as random variables.

By using the Bayesian approximation error (BAE) approach we were able to approximately marginalise over the effects of the small scale structure of the conductivity. In particular, this meant we were able to use a drastically coarsened mesh to carry out our inversions, treating the small scale component as an auxiliary unknown, and compensating for any discretisation errors.

We also showed that a novel extension to the BAE approach based on feature extraction can be implemented to estimate auxiliary unknowns at almost no extra cost. This method was then applied to infer the amplitude and correlation length of the small scale component of the conductivity. The extension formulated also allowed us to quantify the uncertainty in
the estimates for these small scale parameters.

To address the problem of recovering the entire multiscale conductivity we showed that use of multiscale finite element methods (MsFEM) with gradient based inversion techniques and the BAE approach may be a feasible alternative to the use of computationally heavy sampling based methods.

Some of the methods in this thesis have only been tested in one dimension. The provided application of the extension to the BAE approach for estimating auxiliary unknowns developed appeared more useful to the one dimensional setting, with only limited success in the two dimensional case, however, this could most likely be solved by the use of a better suited feature. Moreover the use of MsFEM for inversions was only carried out in one dimension. With regard to the extension to the BAE approach a future direction for work could be to develop features better equipped for two or three dimensions. There may also be the possibility to use such a feature extraction style approach to give improved estimates for the primary parameters. Application of MsFEM to higher dimensional inverse problems is also worth investigating.

It would also be interesting to look into developing the theory of stochastic homogenisation to accommodate for the use of hyper-priors. This could further reduce computational cost when reconstructing the coarse scale conductivity.

In conclusion, feasible coarse scale conductivity estimates are possible and computationally efficient to obtain by using the BAE approach. Moreover, by using the BAE approach in conjunction with MsFEM it is also possible to obtain reasonable estimates for the full multiscale conductivity.
APPENDIX A

MATHEMATICAL APPENDIX

A.1 Definitions

Definition A.1.1 (Sobolev Spaces). For $1 \leq p \leq \infty$ and for every $s \in \mathbb{N}$, we define the Sobolev space as

$$W^{s,p}(\Omega) = \left\{ u \in L^p(\Omega), \partial^\alpha u \in L^p(\Omega), \forall \alpha \in \mathbb{N}^d, |\alpha| \leq s \right\},$$

with norm

$$\|u\|_{W^{s,p}(\Omega)} = \left( \sum_{|\alpha| \leq s} \int_\Omega |\partial^\alpha u|^p \, dx \right)^{\frac{1}{p}}.$$

Definition A.1.2 (Hilbert Spaces). We define the Hilbert space as $H^s(\Omega) = W^{s,2}(\Omega)$, with inner product

$$\langle u, v \rangle = \sum_{|\alpha| \leq s} \int_\Omega (\partial^\alpha u)(\partial^\alpha v) \, dx.$$

Definition A.1.3 ($H^{-1}(\Omega)$). We denote by $H^{-1}(\Omega)$ the dual of $H^1_0(\Omega)$ with the norm

$$\|F\|_{H^{-1}(\Omega)} = \sup \left\{ |F(x)| : u \in H^1_0(\Omega), \|u\|_{H^1(\Omega)} = 1 \right\}.$$

Definition A.1.4 ($H^{\frac{1}{2}}(\partial\Omega)$). We define the space $H^{\frac{1}{2}}(\partial\Omega)$ as

$$H^{\frac{1}{2}}(\partial\Omega) = \left\{ u \in L^2(\Omega), \exists v \in H^1(\Omega), u = \gamma v \right\},$$

where $\gamma : H^1(\Omega) \to L^2(\Omega)$ is the trace mapping, with norm defined

$$\|u\|_{H^{\frac{1}{2}}(\partial\Omega)} = \inf \left\{ \|v\|_{H^1(\Omega)} : v \in H^1(\Omega), u = \gamma v \right\}.$$
Definition A.1.5 (Coercivity). Let $\mathcal{H}$ be a real Hilbert space, and let $B : \mathcal{H} \times \mathcal{H}$ be a bilinear form, then $B$ is coercive if there exists $c > 0$ such that

$$B(u, u) \geq c \|u\|_{\mathcal{H}} \quad \forall u \in \mathcal{H}.$$ 

Theorem A.1.6 (Riesz Representation). Let $\mathcal{H}$ be Hilbert space, with dual $\mathcal{H}^*$. For every bounded linear functional $F \in \mathcal{H}^*$, there exists a uniquely determined element $v \in \mathcal{H}$ such that

$$F(u) = \langle u, v \rangle \quad \forall u \in \mathcal{H},$$

and $\|F\| = \|v\|$.

Theorem A.1.7 (Lax-Milgram). Let $B$ be a bounded, coercive bilinear form on a Hilbert space $\mathcal{H}$, with dual $\mathcal{H}^*$. Then for every bounded linear functional $F \in \mathcal{H}^*$ there exists a unique element $v \in \mathcal{H}$ such that

$$B(u, v) = F(u) \quad \text{for all } u \in \mathcal{H}.$$


