Recovery Methods for Evolution and Nonlinear Problems

Tristan Martin Pryer

Thesis submitted for the degree of Doctor of Philosophy

31st October 2010

Declaration

I hereby declare that this thesis has not been and will not be submitted in whole or in part to another University for the award of any other degree.

Signature:

Dedication

This Thesis is dedicated to my loving girlfriend Verity Webster.

Acknowledgments

This piece of research would not have been possible without the support of many people.

First and foremost is my supervisor Omar Lakkis whose advice, criticisms and encouragement aided me unmeasurably.

To other members of the Sussex faculty including (but not limited to) Erik Burman, Vanessa Styles and Holger Wendland I offer my extreme gratitude for their useful ideas.

I am indebted to my postgraduate peers agmonst whom include Chandrasekhar Venkataraman, James McMichen, Raquel Barriera, Lavinia Sarbu and Mohammad Shahrokhi-Dehkordi for their insightful comments.

I am grateful to the administrative staff of the Mathematics department at Sussex. Especially to Louise Winters for the early morning chit chats and tea.

Finally, I wish to thank my entire family, especially my mother, father and brother for providing the support I needed.

Abstract

Functions in finite dimensional spaces are, in general, not smooth enough to be differentiable in the classical sense and "recovered" versions of their first and second derivatives must be sought for certain applications. In this work we make use of recovered derivatives for applications in finite element schemes for two different purposes. We thus split this Thesis into two distinct parts.

In the first part we derive energy-norm aposteriori error bounds, using gradient recovery (ZZ) estimators to control the spatial error for fully discrete schemes of the linear heat equation. To our knowledge this is the first completely rigorous derivation of ZZ estimators for fully discrete schemes for evolution problems, without any restrictive assumption on the timestep size. An essential tool for the analysis is the elliptic reconstruction technique introduced as an aposteriori analog to the elliptic (Ritz) projection.

Our theoretical results are backed up with extensive numerical experimentation aimed at (1) testing the practical sharpness and asymptotic behaviour of the error estimator against the error, and (2) deriving an adaptive method based on our estimators.

An extra novelty is an implementation of a coarsening error "preindicator", with a complete implementation guide in ALBERTA (versions 1.0–2.0).

In the second part of this Thesis we propose a numerical method to approximate the solution of second order elliptic problems in nonvariational form. The method is of Galërkin type using conforming finite elements and applied directly to the nonvariational (or nondivergence) form of a second order linear elliptic problem. The key tools are an appropriate concept of the "finite element Hessian" based on a Hessian recovery and a Schur complement approach to solving the resulting linear algebra problem. The method is illustrated with computational experiments on linear PDEs in nonvariational form.

We then use the nonvariational finite element method to build a numerical method for

fully nonlinear elliptic equations. We linearise the problem via Newton's method resulting in a sequence of nonvariational elliptic problems which are then approximated with the nonvariational finite element method. This method is applicable to general fully nonlinear PDEs who admit a unique solution without constraint.

We also study fully nonlinear PDEs when they are only uniformly elliptic on a certain class of functions. We construct a numerical method for the Monge–Ampère equation based on using "finite element convexity" as a constraint for the aforementioned nonvariational finite element method. This method is backed up with numerical experimentation.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The study of partial differential equations (PDEs) began in the 18th century with the work of Euler, Lagrange, Laplace and d'Alembert who modelled various physical phenomena. The analysis of these models has remained a major part of mathematical study up to the present day. There are a variety of methods by which a PDE can be solved by hand. However, it is not currently possible to solve every PDE using such methods. This is especially true when we look at the solutions of nonlinear PDEs. In fact there is no general procedure available to solve nonlinear PDEs.

A powerful methodology used to approximate the solution of PDEs in general comes from numerical analysis techniques, such as finite difference, finite volume or finite element schemes. Currently it is probably the fastest developing area of numerical analysis, made possible by the rapid increase in the speed of personal computers.

Finite difference approximations have a long history, for example Bernoulli, Newton and Euler all made use of them, but numerical methods in general were virtually unknown before 1950. In a finite difference method we generate a structured (uniform) grid over the domain in which we are interested. The partial derivatives are then approximated by finite difference operators. Finite volume methods are constructed using a similar grid (although it no longer has to be structured). In this case we consider integrals of divergence terms from the PDE and, making use of the divergence theorem, consider them as surface integrals over small "volumes" of the nodes.

This Thesis details work we have undertaken relating to finite element methods. As with the finite difference and finite volume schemes, we require the domain to be divided

into a finite number of regions or elements which then constitute a mesh, although unlike finite difference schemes the mesh is not necessarily regular. We may take these elements to be triangles or quadrilaterals in the plane or tetrahedrons or hexahedrons in space. The solution is approximated with functions over each of the elements and the local contributions are assembled over the entire problem domain. The result is a linear system which can be solved using existing linear algebra techniques.

The mathematical basis of the finite element method comes from the work of Rayleigh and Ritz who introduced the variational calculus procedures fundamental to the method in 1877–1909.

Finite element methods (FEM) arguably constitute one of the most successful method families in numerically approximating second order elliptic PDEs that are given in variational (also known as divergence) form. The reasons behind the finite element methods success in such a framework are twofold: (1) the weak form is suitable to apply functional analytic frameworks (Lax–Milgram Theorem e.g.), and (2) the discrete functions need to be differentiated only once at most.

We are particularly interested in the concept of *recovery methods* and how we may apply them in the finite element framework. We use the term recovery methods to describe the representation of derivatives of a piecewise polynomial function in a higher regularity space than they naturally exist. The extra regularity is aimed at either obtaining a higher approximation order, for example the gradient recovery class of aposteriori estimators, or representing an object that would not otherwise exist as a function, like Hessian recovery.

Broadly this Thesis can be read in two parts. In §2 and §3, we are concerned with recovery operators and their use as aposteriori estimators for elliptic and parabolic problems respectively. In §4 and §5, we detail a finite element method for nonvariational second order elliptic problems from the case of the linear problem to the fully nonlinear problem.

In §2 we give an introduction to the concept of recovery operators as aposteriori estimators for linear elliptic problems. We begin by introducing the core notation used throughout the Thesis. We introduce the finite element method for the model elliptic problem and review some basic convergence properties. We then briefly look at aposteriori estimation for elliptic problems from the gradient recovery point of view.

In §3 we move onto studying recovery operators on parabolic problems. The aim of

this chapter is to bridge the gap between the practical use of ZZ estimators in adaptivity for evolution equations [ZW98, Pic03] and the rather mature error control theory via recovery for stationary equations. We focus on the linear heat equation as a model problem. Leykekheman & Wahlbin [LW06] are to our knowledge the only researchers to have explored this issue in depth. While obtaining satisfactory error bounds for spatially discrete schemes, they must assume unrealistically small timesteps for the fully discrete case. In this chapter we thoroughly analyse the fully discrete backward Euler schemes. More specifically, we provide reliable error bounds. The efficiency and asymptotic exactness of the bounds is dealt with computationally.

Our main analytical tool to tackle the fully discrete scheme's difficulties is the *elliptic reconstruction* [LM06], which provides a way to take advantage of elliptic aposteriori error estimates based on gradient recovery following Ainsworth & Oden's exposition [AO00].

The elliptic reconstruction technique, introduced under this name by Makridakis & Nochetto [MN03], involves the splitting the error into two parts; a *parabolic error* and an *elliptic error*, through the use of the *elliptic reconstruction* of the discrete solution, defined in (3.25). This allows us to utilise existing elliptic aposteriori estimators for the elliptic part and standard parabolic energy estimates to control the second part. Despite this technique being initially introduced to derive sharp bounds for lower order spatial error norms, such as $L_2(\Omega)$ [MN03, LM06, LMP10] and $L_\infty(\Omega)$ [DLM09], we apply it here as an analysis tool in an energy-norm framework, where a direct approach leads to a highly complicated analysis for the fully discrete scheme.

In fact, the single most interesting feature of the elliptic reconstruction is that the parabolic error's energy norm term is of a higher order (with respect to the spatial mesh-size parameter) than the elliptic error [LM06]. In this chapter we show, rigorously, that the full energy error can be accounted for only by the elliptic error, as long as data and timestep are resolved sufficiently well (cf. Lemma 3.4.1.1). This crucial observation was also used by Georgoulis & Lakkis to obtain residual aposteriori estimates for nonconforming methods [GL09]. Note that it is part of the adaptive methods practitioner's folklore to employ heuristic versions of this argument. By way of example, we quote Ziukas & Wiberg: "the [full parabolic] discretisation in energy norm can be bounded by the [elliptic error] estimator" [ZW98].

Although we treat the case of the heat equation for simplicity in this chapter, our results can be extended to cover more general elliptic operators, even time-dependent ones, by using appropriate elliptic gradient recovery techniques [FV06] and a more careful time-step analysis [GL09, cf.].

In §4 we move onto studying how we may use recovery methods to approximate the Hessian of a finite element function. We make use of this notion by departing from the basis set out in §2 and considering second order elliptic boundary value problems (BVPs) in nonvariational form

$$\text{find } u \text{ such that } \boldsymbol{A}{:}\mathrm{D}^2 u = f \text{ in } \Omega \text{ and } u|_{\partial\Omega} = g, \tag{1.1}$$

for which one may not always be successful in applying the standard FEM (with reference to §4.1 for the notation). Indeed, the use of the standard FEM requires (1) the coefficient matrix $\boldsymbol{A} : \Omega \to \mathbb{R}^{d\times d}$ to be (weakly) differentiable and (2) the rewriting of the second order term in divergence form, an operation which introduces an advection (first order) term:

$$\boldsymbol{A}{:}\mathrm{D}^2 u = \operatorname{div} \boldsymbol{A}\nabla u - (\operatorname{div} \boldsymbol{A})\,\nabla u. \tag{1.2}$$

Even when coefficient matrix $\boldsymbol{A}$ is differentiable on Ω, this procedure could result in the problem becoming advection–dominated and unstable for conforming FEM, as we demonstrate numerically using Problem (4.146).

Our main motivation for studying linear elliptic BVPs in nonvariational form is their important role in pure and applied mathematics. Important examples of nonvariational problems include the fully nonlinear BVP that is approximated via a Newton method, which becomes an infinite sequence of linear nonvariational elliptic problems [Boh08], or the Kolmogorov equations arising in the area of stochastics [SST08].

In this chapter, we propose and test a direct discretisation of the strong form (1.1) that makes no special assumption on the derivative of $\boldsymbol{A}$. The key concept is an appropriate definition of a *finite element Hessian* given in §4.1.3. The finite element Hessian has been used earlier in different contexts, such as anisotropic mesh generation [AV02, CSX07, VMD$^+$07] and *finite element convexity* [AM09]. The finite element Hessian is related also to the finite element (discrete) elliptic operator appearing in the analysis of evolution problems, see §3.3 and [Tho06] for details.

The method we propose is quite straightforward, and we are surprised that it is not easily available in the literature. It consists of discretising, via a Galërkin procedure, the BVP (1.1) directly without writing it in divergence form.

The main difficulty of our approach is having to deal with a somewhat involved linear algebra problem that needs to be solved as efficiently as possible (this is especially important when we apply this method in the linearisation of nonlinear elliptic BVPs). We overcame this difficulty in §4.2, by combining the notion of the distributional Hessian of a piecewise smooth function v,

$$\left\langle \mathrm{D}^2 v \,|\, \phi \right\rangle = -\int_\Omega \nabla v \otimes \nabla \phi + \int_{\partial\Omega} \nabla v \otimes \boldsymbol{n}\phi \quad \forall\, \phi \in \mathrm{C}^\infty(\overline{\Omega}), \tag{1.3}$$

with equation (1.1) resulting in a system of equations that are larger, but easier to handle numerically, once discretised.

It is worth noting that there are alternatives to our approach, most notably the standard finite difference method and its variants. The reason we are interested in a Galërkin procedure is the ability to use an unstructured mesh, essential for complicated geometries where the finite difference method leads to complicated, and sometimes prohibitive modifications (especially in dimension 3 and higher), and the potential of dealing with adaptive methods using available finite element code. Furthermore, our method has the potential to approach the iterative solution of fully nonlinear problems where finite difference methods can become clumsy and demanding [KT92, LR05, Obe08, CS08].

We make use of relatively standard techniques to derive an apriori bound in the $\mathrm{H}^{-1}(\Omega)$ norm. We numerically demonstrate convergence in this norm by making use of Lemma 3.9 from [LP10d] on the computation of $\mathrm{H}^{-1}(\Omega)$ norms. This observation (also given in Lemma 4.7.0.15) allows us to compute the $\mathrm{H}^{-1}(\Omega)$ norm of a function with as much accuracy as a (standard) finite element method allows for its energy norm.

We also study the scheme from an aposteriori framework. In this case we again find that 'standard' techniques yield a bound in the $\mathrm{H}^{-1}(\Omega)$ norm provoking the observation that this is the natural norm for the problem. To make computations simpler in this case we are able to apply a duality argument to derive $\mathrm{L}_2(\Omega)$ aposteriori bounds.

To finish this chapter we give a brief interlude on the numerical approximation of quasilinear PDEs in nonvariational form. We do this to make the method in §5 more accessible.

In §5 we make further use of the finite element Hessian by using it in a method to approximate fully nonlinear elliptic PDEs.

Fully nonlinear PDEs arise in many areas, including differential geometry (Monge–Ampère equation), mass transportation (Monge–Kantorovich problem), dynamic programming (Bellman equation) and fluid dynamics (geostrophic equations).

It is difficult to pose numerical methods for fully nonlinear equations for three main reasons. The first more obvious one is the strong nonlinearity on the highest order derivative. The second is the fact that a fully nonlinear equation does not always admit a classical solution even if the problem data is sufficiently smooth. The third is that the problem may not admit a unique solution, but multiple, then even if we can construct a numerical approximation it is difficult to know which solution we are approximating.

Regardless of the problems, numerical simulation of fully nonlinear second order elliptic equations have been the brunt of much recent study, particularly for the case of Monge–Ampère of which [DG06, FN08b, LR05, Obe08, OP88] are selected examples.

For general fully nonlinear equations some methods have been presented. In [Boh08] the author presents a C^1 finite element method and goes to great lengths to show stability and consistency of the scheme. The basis of this argument comes from Stetter [Ste73]. The practical relavence of this approach is questionable, however, since the C^1 finite elements are complicated and computationally expensive, the minimal order of the polynomial basis that falls under the framework is 5, using the Argyris element for example. In addition the C^1 finite elements are only available in 2D.

In [FN07, FN08b, FN08a] the authors give a method in which they approximate the general second order fully nonlinear PDE by a sequence of fourth order quasilinear PDEs. These are nonlinear biharmonic equations which allow the authors to numerically discretise via mixed finite elements for example. In fact for the Monge–Ampère equation, which admits two solutions, one convex and one concave, this methods allows for the approximation of both solutions via the correct choice of a parameter. Although computationally less expensive than C^1 finite elements, the mixed formulation still results in an extremely large algebraic system and perhaps the method can be further numerically improved. There is also the possibility of developing a new concept of "weak solution" for the fully nonlinear PDE, which the authors have named the "vanishing moment method". A major

advantage to this method over the current viscosity solution technique is the constructive nature of the procedure.

The method we propose in §5 consists of applying a Newton linearisation to the fully nonlinear PDE. This results in a sequence of linear nonvariational PDEs. At this point the problem falls under the framework of the finite element method proposed in §4. We numerically study various problems that are specifically constructed to be uniformly elliptic.

The major problem with our technique is the nonuniqueness of the problem. The method itself breaks down unless some constraints are passed down from the continuous level. It was observed numerically that at the discrete level the convexity constraint is violated and the sequence of linear operators lose ellipticity.

We solve this problem by formulating each Newton step as a semidefinite programming problem [VB96]. This is an optimisation problem which includes the areas of linear programming and convex quadratic programming with convex constraints.

We study the Monge–Ampère equation and give various numerical examples showing the method with added constraint is robust.

We finish in §6 with a summary of work and possible future directions the aforementioned projects may take.

Chapter 2

Recovery operators in linear elliptic problems

In this chapter we will present a brief summary of the conforming finite element method applied to a model elliptic PDE. We will provide a notion of aposteriori estimation in this simple case using gradient recovery estimators.

We will show analytically under certain assumptions that these estimators are asymptotically exact and give examples of when these estimators should not be used or should be supplemented by additional data [FV06].

The material presented in this chapter is well known and is meant as an introduction to the chapters studied hereafter.

2.1 Elliptic model problem

Let $\Omega \subset \mathbb{R}^d$ be a bounded d-dimensional domain with boundary $\partial\Omega$. Consider the linear second order partial differential equation (PDE) supplemented with Dirichlet boundary conditions

$$\begin{aligned} \mathscr{L}u := -\operatorname{div} \boldsymbol{A}\nabla u = f \text{ in } \Omega \\ u = 0 \text{ on } \partial\Omega, \end{aligned} \tag{2.1}$$

where $f : \overline{\Omega} \to \mathbb{R}$ and $\boldsymbol{A} : \overline{\Omega} \to \mathbb{R}^{d\times d}$. We assume that the matrix $\boldsymbol{A}(\boldsymbol{x})$ is bounded and uniformly positive definite, that is for each $\boldsymbol{\xi} \in \mathbb{R}^d \,/\, \{\mathbf{0}\}$ and $\boldsymbol{x} \in \Omega$ there exists a $\lambda > 0$

such that

$$\boldsymbol{\xi}^{\intercal}\boldsymbol{A}(\boldsymbol{x})\boldsymbol{\xi} \geq \lambda \left|\boldsymbol{\xi}\right|^2 . \tag{2.2}$$

This is a strict ellipticity condition. We introduce the vector space $\mathrm{C}^{\infty}(\Omega)$ of infinitely differentiable functions. In this space differentiability is understood in the classical sense. Also the vector subspace $\mathrm{C}_0^{\infty}(\Omega)$ of $\mathrm{C}^{\infty}(\Omega)$ of functions with compact support on Ω.

The standard methodology of finite elements then requires us to construct a *weak formulation* of (2.1). With this in mind we introduce the Sobolev space notation [Cia78, Eva98] with $q \in [1, \infty)$

$$\mathrm{L}_q(\Omega) := \left\{ \phi : \int_{\Omega} |\phi|^q < \infty \right\}, \tag{2.3}$$

$$\mathrm{L}_{\infty}(\Omega) := \left\{ \phi : \sup_{\boldsymbol{x}\in\Omega} |\phi(\boldsymbol{x})| < \infty \right\}, \tag{2.4}$$

$$\mathrm{W}_q^k(\Omega) = \{\phi \in \mathrm{L}_q(\Omega) : \mathrm{D}^{\boldsymbol{\alpha}} u \in \mathrm{L}_q(\Omega) \text{ for } |\boldsymbol{\alpha}| \leq k\}, \tag{2.5}$$

$$\mathrm{W}_{\infty}^k(\Omega) = \{\phi \in \mathrm{L}_{\infty}(\Omega) : \mathrm{D}^{\boldsymbol{\alpha}} u \in \mathrm{L}_{\infty}(\Omega) \text{ for } |\boldsymbol{\alpha}| \leq k\}, \tag{2.6}$$

$$\mathrm{H}^k(\Omega) := \mathrm{W}_2^k(\Omega), \tag{2.7}$$

where $\boldsymbol{\alpha} = \{\alpha_1, ..., \alpha_d\}$ is a multi-index, $|\boldsymbol{\alpha}| = \sum_{i=1}^{d} \alpha_i$, derivatives $\mathrm{D}^{\boldsymbol{\alpha}}$ are understood in a weak sense and by integrability is meant Lesbegue. We pay particular attention to the space $\mathrm{L}_2(\Omega)$ which is equipped with the inner product

$$\langle v, w \rangle := \int_{\Omega} vw. \tag{2.8}$$

This induces the norm

$$\|v\|^2 := \|v\|_{\mathrm{L}_2(\Omega)}^2 = \langle v, v \rangle . \tag{2.9}$$

The Hilbert spaces $\mathrm{H}^k(\Omega)$ are equipped with norms and semi-norms

$$\|v\|_k^2 := \|v\|_{\mathrm{H}^k(\Omega)}^2 = \sum_{|\boldsymbol{\alpha}|\leq k} \|\mathrm{D}^{\alpha} v\|^2 \tag{2.10}$$

$$\text{and } |v|_k^2 := |v|_{\mathrm{H}^k(\Omega)}^2 = \sum_{|\boldsymbol{\alpha}|=k} \|\mathrm{D}^{\alpha} v\|^2 \tag{2.11}$$

respectively. Note that for each $k \in \mathbb{N}$, $\mathrm{H}^k(\Omega)$ is the Banach completion of $\mathrm{C}^{\infty}(\Omega)$ with respect to its norm $\|\cdot\|_k$. With that in mind we denote the spaces $\mathrm{H}_0^k(\Omega)$ to be the Banach completion of $\mathrm{C}_0^{\infty}(\Omega)$ with respect to the norm $\|\cdot\|_k$, that is

$$\mathrm{H}_0^k(\Omega) := \overline{\mathrm{C}_0^{\infty}}\big|_{\|\cdot\|_k}. \tag{2.12}$$

Further, we wish to make use of the corresponding dual spaces of $\mathrm{H}_0^k(\Omega)$. To that end we define $\mathcal{D}(\Omega)$, the space of distributions to be the dual space of $\mathrm{C}_0^\infty(\Omega)$. We associate with these spaces a *duality pairing* between $d \in \mathcal{D}(\Omega)$ and $\phi \in \mathrm{C}_0^\infty(\Omega)$ which we denote $\langle d \,|\, \phi \rangle$. In fact, if d is locally integrable then we can represent d by

$$\langle d \,|\, \phi \rangle = \int_\Omega d\phi \tag{2.13}$$

and we see that $\langle \cdot \,|\, \cdot \rangle$ is nothing but an extension of the standard $\mathrm{L}_2(\Omega)$ inner product. Now we may define

$$\mathrm{H}^{-k}(\Omega) := \operatorname{dual} \mathrm{H}_0^k(\Omega) \tag{2.14}$$

and equip it with the norm

$$\|v\|_{-k} := \|v\|_{\mathrm{H}^{-k}(\Omega)} = \sup_{0 \neq \phi \in \mathrm{H}_0^k(\Omega)} \frac{\langle v \,|\, \phi \rangle}{|\phi|_k}. \tag{2.15}$$

2.2 Weak formulation and discretisation

Henceforth we will use the convention that the vector of partial derivatives, $\mathrm{D}u$, of a function $u : \Omega \to \mathbb{R}$ is a row vector, while the gradient of u, ∇u is the derivative's transpose, i.e., $\nabla u = (\mathrm{D}u)^\intercal$. We will make use of the slight abuse of notation as is standard practice in the literature and denote the Hessian of a function $\mathrm{D}^2 u := \nabla \mathrm{D} u$ to be a $d \times d$ matrix.

To make the transition to the weak formulation we test (2.1) with a smooth function $\phi \in \mathrm{H}_0^1(\Omega)$ over the domain of interest. Applying Green's Theorem (A.1.0.7) and noting that both $u, \phi \in \mathrm{H}_0^1(\Omega)$, gives the problem: Find $u \in \mathrm{H}_0^1(\Omega)$ such that

$$\int_\Omega \mathrm{D}\phi \boldsymbol{A} \nabla u = \int_\Omega f\phi \quad \forall\, \phi \in \mathrm{H}_0^1(\Omega). \tag{2.16}$$

The term weak formulation heuristically refers to the fact we have "weakened" the differentiability requirements on u to solve the problem. For simplicity of notation we introduce the following shorthand:

$$a(u, \phi) = \int_\Omega \mathrm{D}\phi \boldsymbol{A} \nabla u, \tag{2.17}$$

$$l(\phi) = \int_\Omega f\phi. \tag{2.18}$$

With this new notation, (2.16) becomes: Find $u \in \mathrm{H}_0^1(\Omega)$ such that

$$a(u, \phi) = l(\phi) \quad \forall\, \phi \in \mathrm{H}_0^1(\Omega). \tag{2.19}$$

In fact we may induce a norm from the bilinear form $\|v\|_a^2 := a(v, v)$ called the *energy norm* which for our problem is equivalent to the norm $\|v\|_1$, that is there exist constants α, β such that

$$\beta^{1/2} \|v\|_1 \leq \|v\|_a \leq \alpha^{1/2} \|v\|_1 . \tag{2.20}$$

Note for clarity of presentation we are restricting ourselves to homogeneous Dirichlet boundary conditions however this can be extended to non-trivial boundary values. Indeed, if we are provided with additional problem data $u(\boldsymbol{x}) = g(\boldsymbol{x})$ on $\partial\Omega$ we "shift" the space in which we seek the solution to $\mathrm{H}_g^1(\Omega) := \mathrm{H}_0^1(\Omega) + g$.

In order to construct a finite element approximation of the problem (2.19) we employ a conforming h-version Galërkin procedure whereby we replace $\mathrm{H}_0^1(\Omega)$ by a finite dimensional subspace $\mathbb{V} \subset \mathrm{H}_0^1(\Omega)$ consisting of continuous piecewise polynomials of degree p on a partition of Ω.

To that end we let $\mathscr{T}$ be a conforming, not necessarily quasi-uniform, triangulation of Ω, that is,

1. $K \in \mathscr{T}$ means K is an open simplex (segment for $d = 1$, triangle for $d = 2$ or tetrahedron for $d = 3$),
2. for any $K, J \in \mathscr{T}$ we have that $\overline{K} \cap \overline{J}$ is a full subsimplex (i.e., it is either $\emptyset$, a vertex, an edge, a face, or the whole of $\overline{K}$ and $\overline{J}$) of both $\overline{K}$ and $\overline{J}$ and
3. $\bigcup_{K \in \mathscr{T}} \overline{K} = \overline{\Omega}$.

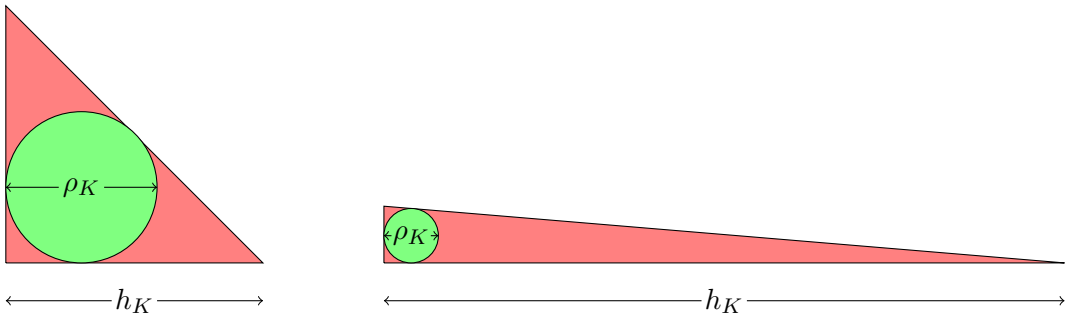
The shape regularity of $\mathscr{T}$ is defined as the number

$$\mu(\mathscr{T}) := \inf_{K \in \mathscr{T}} \frac{\rho_K}{h_K}, \tag{2.21}$$

where ρ_K is the radius of the largest ball contained inside K and h_K is the longest side of K.

2.2.0.1 Example (shape regularity of elements)**.** The shape regularity of a triangle is merely the ratio $\frac{\rho_K}{h_K}$. On the left we have an isotropic triangle with a relatively high shape

regularity. In the following figure on the right is an anisotropic triangle with a relatively low shape regularity.



An indexed family of triangulations $\{\mathscr{T}_k\}_k$ is called *shape regular* if

$$\mu := \inf_k \mu(\mathscr{T}_k) > 0. \tag{2.22}$$

We will use henceforth the usual convention where $h : \Omega \to \mathbb{R}$ denotes the *meshsize function* of $\mathscr{T}$, i.e.,

$$h(x) := h_{\mathscr{T}}(x) := \max_{K \ni x} h_K. \tag{2.23}$$

With a triangulation $\mathscr{T}$ as described above, and an integer $p \geq 1$ fixed, we may now consider the *finite element space*

$$\mathbb{V} := \left\{ \Phi \in \mathrm{H}_0^1(\Omega) : \ \Phi|_K \in \mathbb{P}^p \, \forall \, K \in \mathscr{T} \right\}; \tag{2.24}$$

where $\mathbb{P}^k$ denotes the linear space of polynomials in d variables of total degree no higher than a positive integer k. The *finite element approximation* to (2.1) is the function $U \in \mathbb{V}$ such that

$$a(U, \Phi) = \langle f, \Phi \rangle \quad \forall \, \Phi \in \mathbb{V}. \tag{2.25}$$

2.3 Apriori analysis

This small section is a brief compilation of standard results for the standard conforming h-version finite element method of degree p (2.25).

2.3.0.2 Lemma (Galërkin orthogonality)**.** *Let $u \in \mathrm{H}_0^1(\Omega)$ be the weak solution of (2.19) and $U \in \mathbb{V}$ be the finite element solution of (2.25). Then*

$$a(u - U, \Phi) = 0 \quad \forall \, \Phi \in \mathbb{V}. \tag{2.26}$$

Proof of 2.3.0.2. Note that (2.19) holds for each $\Phi \in \mathbb{V} \subset \mathrm{H}_0^1(\Omega)$. Taking the difference of (2.19) and (2.25) yields the desired result. □

2.3.0.3 Lemma (quasioptimality of the FE approximation)**.** *Let u and U be defined as in Lemma 2.3.0.2 then*

$$\|u - U\|_a \leq \min_{V \in \mathbb{V}} \|u - V\|_a . \tag{2.27}$$

Proof of 2.3.0.3. Due to the Galërkin orthogonality result of Lemma 2.3.0.2

$$\begin{aligned} \|u - U\|_a^2 &\leq a(u - U, u - U) \\ &\leq a(u - U, u) \\ &\leq a(u - U, u - V) \\ &\leq \|u - U\|_a \|u - V\|_a \qquad \forall\, V \in \mathbb{V}. \end{aligned} \tag{2.28}$$

dividing through by $\|u - U\|_a$ and noting that V is arbitrary yields the desired result. □

2.3.0.4 Definition (Lagrange interpolant)**.** A given function with sufficient regularity, $v \in \mathrm{H}^{p+1}(\Omega)$ for example may be approximated by the *Lagrange interpolant* $\Lambda^{\mathbb{V}} : \mathrm{H}^{p+1}(\Omega) \to \mathbb{V}$. This is achieved by representing v as a continuous piecewise polynomial function, $\Lambda^{\mathbb{V}} v$, which coincides with v at the Lagrange nodes $\boldsymbol{x}_i$. Moreover the interpolant satisfies the following error bounds

$$\begin{aligned} \left\|v - \Lambda^{\mathbb{V}} v\right\| &\leq C h^{p+1} |v|_{p+1} \\ \left|v - \Lambda^{\mathbb{V}} v\right|_1 &\leq C h^{p} |v|_{p+1} . \end{aligned} \tag{2.29}$$

2.3.0.5 Lemma (energy norm apriori error bound)**.** *Let $u \in \mathrm{H}_0^1(\Omega)$ be the weak solution of (2.19) and $U \in \mathbb{V}$ be the finite element solution of (2.25). Then*

$$\|u - U\|_a \leq C h^p |u|_{p+1} . \tag{2.30}$$

Proof Choosing $V = \Lambda^{\mathbb{V}} u$ in Lemma 2.3.0.3 and using the properties of the Lagrange interpolant from Definition 2.3.0.4 yields the desired result. □

2.3.0.6 Remark ($\mathrm{H}^1(\Omega)$ apriori error bound)**.** Due to the equivalence of the energy norm and the $\mathrm{H}^1(\Omega)$ norm it follows that

$$\|u - U\|_1 \leq C h^p |u|_{p+1} . \tag{2.31}$$

2.3.0.7 Lemma ($\mathrm{L}_2(\Omega)$ apriori error bound)**.** *Let u and U be defined as in Lemma 2.3.0.5 then the following error bound holds*

$$\|u - U\| \leq Ch^{p+1} \left|u\right|_{p+1}. \tag{2.32}$$

Proof The proof is an Aubin–Nitsche duality argument. We introduce the dual problem of (2.1): Given a generic $g \in \mathrm{L}_2(\Omega)$ find $w \in \mathrm{H}_0^1(\Omega)$ such that

$$a(\phi, w) = \langle g, \phi \rangle \qquad \forall\, \phi \in \mathrm{H}_0^1(\Omega). \tag{2.33}$$

Since $\mathrm{L}_2(\Omega) = \mathrm{dual}\,\mathrm{L}_2(\Omega)$, i.e., $\mathrm{L}_2(\Omega)$ is its own dual we can represent the norm in the following form

$$\|v\| = \sup_{\phi \in \mathrm{L}_2(\Omega)} \frac{\langle v, \phi \rangle}{\|\phi\|}. \tag{2.34}$$

Indeed by virtue of Cauchy–Bunyakovskĭ–Schwarz inequality

$$\sup_{\phi \in \mathrm{L}_2(\Omega)} \frac{\langle v, \phi \rangle}{\|\phi\|} \leq \sup_{\phi \in \mathrm{L}_2(\Omega)} \frac{\|v\| \, \|\phi\|}{\|\phi\|} \leq \|v\|, \tag{2.35}$$

also

$$\sup_{\phi \in \mathrm{L}_2(\Omega)} \frac{\langle v, \phi \rangle}{\|\phi\|} \geq \frac{\langle v, v \rangle}{\|v\|} \geq \|v\|. \tag{2.36}$$

Let $W \in \mathbb{V}$ be the finite element solution to the dual problem (2.33). Testing the error with a generic function g and using the definition of the dual solution and the Galërkin orthogonality property from Lemma 2.3.0.2 we have

$$\begin{aligned} \langle u - U, g \rangle &= a(w, u - U) \\ &= a(w - W, u - U) \\ &\leq \alpha \left\|w - W\right\|_1 \left\|u - U\right\|_1 \\ &\leq Ch \left|w\right|_2 h^p \left\|u\right\|_{p+1} \\ &\leq Ch^{p+1} \left\|u\right\|_{p+1} \|g\|, \end{aligned} \tag{2.37}$$

where we have used the regularity result

$$\left|w\right|_2 \leq C \|g\| \tag{2.38}$$

from Theorem A.1.0.11. Dividing through by $\|g\|$ and noting g was generic yields the desired result. □

2.4 Recovery operators as aposteriori estimators

2.4.1 A general introduction to aposteriori estimation

The content of §2.3 allows us to infer rates of convergence on the error committed by the finite element solution. However the upper bound itself is uncomputable (in general) due to the dependence on the exact solution. The aim of aposteriori analysis is to derive bounds for the error which are explicitly computable from the problem data, which for our model problem consists of $\boldsymbol{A}$, f and Ω, that is we wish to find an *estimator functional* $\mathscr{E}$ such that

$$\|u - U\|_{\mathscr{X}} \leq \mathscr{E}[U, \boldsymbol{A}, f, \mathscr{X}, \mathbb{V}], \tag{2.39}$$

where $\mathscr{X}$ is the space in which we wish to estimate the error. From the literature on elliptic aposteriori estimation [AO00, BR78, Cia78, Ver96, BX03a, ZZ87] there is a variety of ways to compute upper (and lower) bounds for the error in the norm $\|\cdot\|_{\mathscr{X}}$ of some function space $\mathscr{X}$ (e.g., $\mathrm{H}_0^1(\Omega)$, $\mathrm{L}_2(\Omega)$ and $\mathrm{L}_\infty(\Omega)$).

This estimate of the error can then be decomposed into local contributions and used to drive adaptive algorithms. This is achieved by either refining the mesh in a local neighbourhood of where the estimate is high, h–adaptivity, locally increasing the degree of polynomial used in the approximation, p–adaptivity, or some combination of both, hp–adaptivity.

2.4.2 Recovery operators

One way of deriving an estimator functional consists of applying a *gradient postprocessing operator (postprocessor)*, say G, to the approximate solution U and then proving that $\|G[U] - \nabla U\|$ is equivalent to the error $\|\nabla u - \nabla U\|$. *Gradient recovery operators* form a subclass of gradient postprocessors.

Gradient recovery aposteriori error estimators have been widely used since their dissemination in the engineering and scientific computation community by Zienkiewicz & Zhu [ZZ87], for which we will often refer to them shortly as *ZZ estimators*. Since their introduction they have constituted the most serious rival to *residual estimators* introduced earlier on [BR78]. The key to the ZZ estimators success is their implementation's simplicity, mild dependence upon the problem's data, and striking superconvergence and

asymptotic exactness properties. On the other hand, residual estimators are a bit more involved in implementation and cost more to compute, but they are easier to handle from the mathematical analysis view-point in deriving rigorous upper and lower bounds.

Recovery operators can be built in a variety of ways such as local weighted averaging (where the gradient is sampled from neighbouring elements) [Pic03], discrete $\mathrm{L}_2(\Omega)$-projection (using least squares fitting) [ZZ87] or global $\mathrm{L}_2(\Omega)$-projection (where a full discrete problem is solved) [BX03a]. The fundamental idea behind these approaches is to build an approximation $G[U]$ of ∇u which is more regular than the piecewise discontinuous gradient ∇U; the extra regularity is aimed at obtaining a higher approximation order.

2.4.2.1 Definition (stars and patches)**.** Given a triangulation $\mathscr{T}$ of Ω we may wish to study localised neighbourhoods of elements, to that end we introduce the notion of stars and patches. Given an element K, the patch of K which we denote $\hat{K}$ is defined as the set of all elements sharing a common subsimplex with K. In symbols

$$\hat{K} = \left\{ \bigcup_{L\in\mathscr{T}} L : \ \overline{L} \cap \overline{K} \neq \emptyset \right\} \tag{2.40}$$

A star is associated to a given degree of freedom $\boldsymbol{x}_i$ of $\mathbb{V}$ and is (at least in the conforming case) the set of all elements sharing that degree of freedom. In symbols

$$\tilde{\boldsymbol{x}}_i = \left\{ \bigcup_{K\in\mathscr{T}} K : \ \boldsymbol{x}_i \in \overline{K} \right\} \tag{2.41}$$

2.4.2.2 Definition (gradient recovery operator [AO00])**.** A *gradient recovery operator* on $\mathbb{V}$ is a linear operator $G : \mathbb{V} \to \mathbb{V}^d$ which enjoys the following properties:

Consistency we have, with $\mathit{\Lambda}^{\mathbb{V}} : \mathrm{C}^0(\Omega) \to \mathbb{V}$ denoting the Lagrange interpolant (see Definition 2.3.0.4),

$$G[\mathit{\Lambda}^{\mathbb{V}} v]\big|_K = \nabla v|_K \quad \forall\, v \in \mathbb{P}^{p+1},\, K \in \mathscr{T}. \tag{2.42}$$

Local bound there exists a $C_{\mathrm{G}} > 0$ such that

$$\|G[V]\|_{\mathrm{L}_\infty(K)} \leq C_{\mathrm{G}} \, \|\nabla V\|_{\mathrm{L}_\infty(\hat{K})} \quad \forall\, V \in \mathbb{V},\, K \in \mathscr{T}, \tag{2.43}$$

where $\hat{K}$ is the *patch* generated by K (2.40).

For simplicity, we assume that the operator is only locally dependent on ∇U, noting nonetheless, that global methods, such as the global $\mathrm{L}_2(\Omega)$-projection proposed by Bank & Xu [BX03a, BX03b], do exist and can be covered by the theory.

Under certain regularity assumptions recovery estimators are shown to be asymptotically exact. For instance, Zlámal [Zlá77] shows that if $w \in \mathrm{H}^{p+1}(\Omega)$, with reference to (2.1) and (2.25), its finite element approximation $W \in \mathbb{V}$ satisfies the following *superconvergence property*:

$$\left\|\nabla(W - \Lambda^{\mathbb{V}} w)\right\| = \mathrm{O}(h^{p+\zeta}) \text{ for some } \zeta \in (0, 1]\,. \tag{2.44}$$

A review of superconvergence results is given in [KN87]. If (2.44) is satisfied then the recovered gradient also satisfies the following superconvergence property [AO00]:

$$\|\nabla w - G[W])\| = \mathrm{O}(h^{p+\zeta}) \text{ for some } \zeta \in (0, 1]\,. \tag{2.45}$$

The reach of Zlámal's result is appreciated by stating the following consequence.

2.4.2.3 Lemma (gradient recovery aposteriori estimate [AO00])**.** *Let $\mathbb{V}$ be the finite element space defined in (2.24) and $G : \mathbb{V} \to \mathbb{V}^d$ a gradient recovery operator according to Definition 2.4.2.2. If u, U are the solutions of (2.19) and (2.25), respectively, and* (2.45) *holds then the recovery operator is asymptotically exact, in the sense that*

$$\lim_{h_{\mathscr{T}} \to 0} \frac{\|\nabla U - G[U]\|}{\|\nabla(U - u)\|} = 1. \tag{2.46}$$

Thus, there exist $\delta_0 \geq 0$, such that $\delta_0(h) \to 0$ as $h \to 0$ and

$$(1 - \delta_0)\,\|\nabla u - G[U]\| \leq \|\nabla(U - u)\| \leq (1 + \delta_0)\,\|\nabla u - G[U]\| \tag{2.47}$$

for all partitions $\mathscr{T}$ of Ω satisfying $h_{\mathscr{T}} < h_0$.

Proof Utilising the lower triangle inequality and (2.45) we see

$$\begin{aligned} \Big|\, \|\nabla U - G[U]\| - \|\nabla(U - u)\| \,\Big| &\leq \|\nabla u - G[U]\| \\ &= \mathrm{O}(h^{p+\zeta}). \end{aligned} \tag{2.48}$$

Note that from Remark 2.3.0.6 $\|\nabla(U - u)\| = O(h^p)$ so

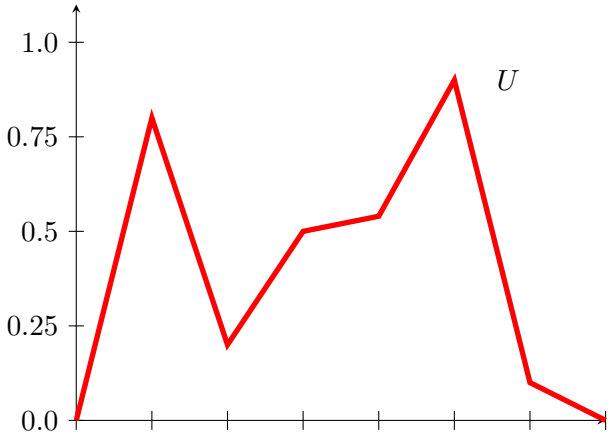
$$\begin{aligned} \frac{\|\nabla U - G[U]\|}{\|\nabla(U - u)\|} - 1 &= \frac{\|\nabla U - G[U]\| - \|\nabla(U - u)\|}{\|\nabla(U - u)\|} \\ &= \frac{O(h^{p+\zeta})}{O(h^p)} \\ &= O(h^{\zeta}), \end{aligned} \tag{2.49}$$

giving the desired result. □

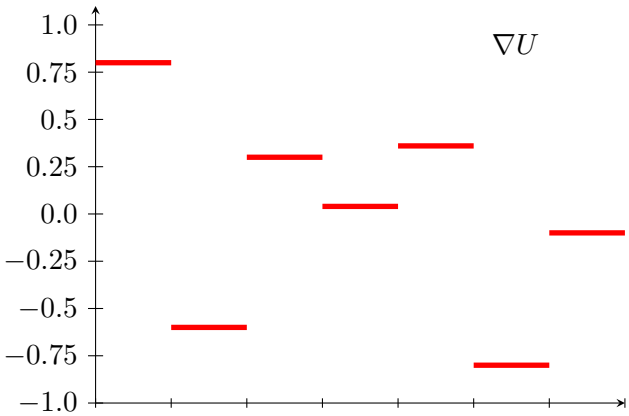
2.4.2.4 Remark (recovery in absence of regularity)**.** Lacking Zlámal's regularity assumption, recovery-based estimators are empirically observed to be efficient, reliable estimators, even on meshes with low shape-regularity [Car04a].

For more details about recovery-based estimators we refer to the available literature [BX03a, BX03b, XZ04, LZ99, AO00].

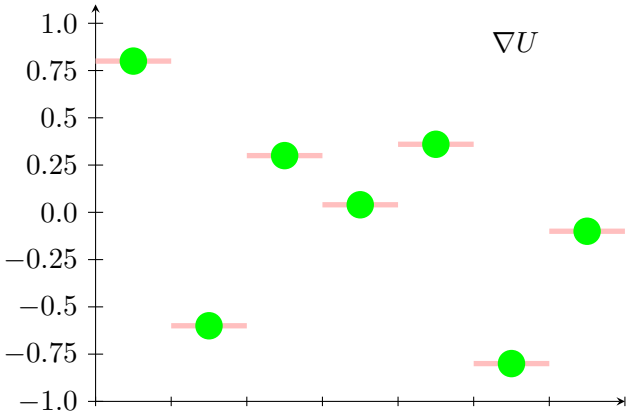
2.4.2.5 Example (a simple recovery estimator for $d = 1$ [AO00, §4.1])**.** For clarity we give a simple example extracted from the Ainsworth Oden book [AO00]. Let V be a continuous piecewise linear function as follows



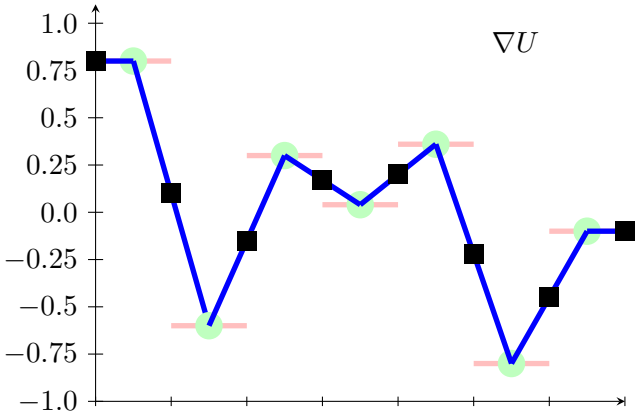
It follows that the gradient of the function, ∇U, is a discontinuous piecewise constant function



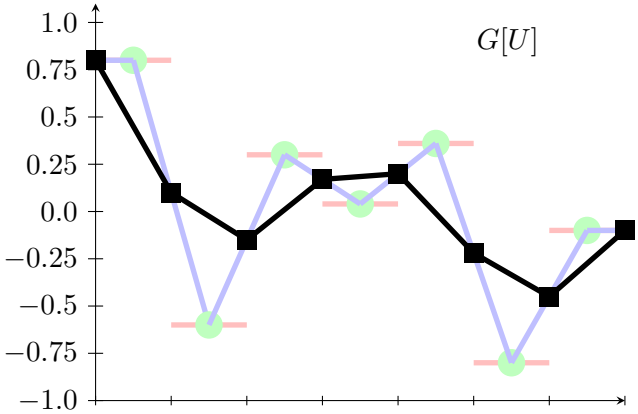
The recovered gradient is built by taking the values of the gradient at the barycenter of the elements.



At each element node we then sample the average of the gradients of the two elements common to this node.



We now have enough information to construct a continuous piecewise linear approximation $G[U]$ of ∇U.



2.4.2.6 Definition (gradient recovery aposteriori estimator functional)**.** Lemma 2.4.2.3 then justifies the use of the *recovery estimator* in the $\mathrm{H}_0^1(\Omega)$-norm (and by equivalence the energy norm) by defining, for the next two chapters, the *gradient recovery aposteriori estimator functional*

$$\mathscr{E}[V] := \mathscr{E}\left[V, \mathrm{H}_0^1(\Omega), \mathbb{V}\right] := \|G[V] - \nabla V\|\,, \text{ for } V \in \mathbb{V}, \tag{2.50}$$

where G is a gradient recovery operator given in Definition 2.4.2.2.

As a recovery estimator—in contrast with residual estimators—generally depends only on operator data (see §3.94 for example), we drop the term f from the estimator functional arguments.

2.4.3 Limitations of recovery estimators

We must be cautious using these estimators since there are certain cases where the recovery estimator $\mathscr{E}[U] = \|G[U] - \nabla U\|$ is not reliable.

2.4.3.1 Example ($\mathrm{L}_2(\Omega)$ orthogonal f [FV06])**.** Consider the case when f is $\mathrm{L}_2(\Omega)$-orthogonal to $\mathbb{V}$. In this case then the finite element solution $U = 0$ hence $\nabla U = 0$ and $G[U] = 0$. The estimate

$$\|\nabla u - \nabla U\| \leq C\,\|\nabla U - G[U]\| \tag{2.51}$$

then yields

$$\|\nabla u\| \leq 0 \tag{2.52}$$

which in general cannot be true.

This unreliability is due to the fact that

$$\langle \nabla u - \nabla U, \nabla \phi\rangle = \langle \nabla u - G[U], \nabla \phi\rangle + \langle G[U] - \nabla U, \nabla \phi\rangle\,. \tag{2.53}$$

By considering the recovery estimator alone we ignore the term $\|\nabla u - G[U]\|$. This term should be superconvergent in many cases however in this example above would be non-zero.

In the paper [FV06], Fierro and Veeser give analytical upper and lower bounds on the term $\|\nabla u - G[U]\|$. In most cases this is of a higher order.

2.4.3.2 Example (under resolution of data)**.** This problem usually arises in the context of adaptivity (see §3.7).

Suppose a function u solves the PDE (2.1) and is non zero only on a very small area of the domain. If we were to calculate the finite element solution of the PDE on an under resolved (coarse) finite element space we may find our finite element solution $U = 0$ which would again give $\nabla U = 0$ and hence $G[U] = 0$. The recovery estimator $\mathscr{E}[U] = 0$ and hence an automated adaptive algorithm would terminate immediately.

2.4.3.3 Example (oversmoothing of ∇U)**.** Suppose u, which is again the solution of (2.1), has a discontinuous gradient. In this case the recovery procedure would smear the discontinuity "polluting" a localised neighbourhood and in this case we may find that $G[U]$ is a worse approximation of ∇u than ∇U.

Chapter 3

Recovery operators as aposteriori estimators for linear parabolic problems

3.1 Introduction

In contrast to the large amount of work on recovery operators used in stationary elliptic problems, as described in §2.4, very little progress has been made on evolution problems with the one notable exception by Leykekhman and Wahlbin [LW06] who must make impractical assumptions on the timestep for the fully discrete schemes.

In this chapter we will rigorously analyse the finite element approximation of the heat equation from an aposteriori viewpoint employing gradient recovery ZZ estimators. We will study both semidiscrete and fully discrete schemes and show that the recovery estimators introduced from §2 can be justifiably used.

We will numerically show the estimators arising from the fully discrete analysis are asymptotically exact, as opposed to the residual based estimators which greatly overestimate the error. We will also propose an adaptive algorithm based on the arising estimators from Theorem 3.5.2.4 and numerically demonstrate the efficiency of the explicit and implicit timestepping schemes. We also give a thorough implementation guide of a coarsening estimator in ALBERTA which arises from the analysis.

This chapter is organised as follows. In §3.2 we introduce the model problem, and

its discretisations via conforming finite elements in space and backward Euler in time. In §3.3 we review a well known apriori error analysis showing convergence rates for the spatial energy norms in the semidiscrete scheme. In §3.4 we prepare the aposteriori analysis by introducing the elliptic reconstruction technique and illustrating its use for the spatially semidiscrete problem. This paves the way to tackle the fully discrete problem in §3.5, where our main results are given in Theorem 3.5.2.4. In §3.6, using numerical tests, we study the practical behaviour of the estimators and in §3.7 we explore the adaptive schemes based on our estimators. As we have used the finite element toolbox ALBERTA [SS05] for the tests, we have taken the opportunity to implement a *coarsening preindicator*, arising from the fully discrete analysis, previously unavailable and fully described in §3.8. This estimator predicts the "information loss" error that will occur under coarsening of the mesh at each timestep of the adaptive method and is crucial in an adaptive code to control information loss during coarsening.

3.2 Set up

3.2.1 The model problem

Let $\Omega \subset \mathbb{R}^d$ be a bounded polyhedral domain and consider the Laplace operator with homogenous Dirichlet boundary conditions denoted by

$$\begin{array}{rccl} \mathscr{A}: & \mathrm{H}_0^1(\Omega) & \to & \mathrm{H}^{-1}(\Omega) \\ & u & \mapsto & \mathscr{A}u := -\Delta u := -\operatorname{div}\nabla u = -\sum_{i=1}^d \partial_i^2 u. \end{array} \tag{3.1}$$

With reference to §2.1 for notation. In addition we make use of the standard notation for spaces of functions whose smoothness differs in the $\boldsymbol{x}$ and t variables [Eva98].

We let $T > 0$, the model parabolic problem consists in finding a function $u \in \mathrm{L}_2(0,T;\mathrm{H}_0^1(\Omega))$ and $\partial_t u \in \mathrm{L}_2(0,T;\mathrm{H}^{-1}(\Omega))$ such that

$$\begin{aligned} \partial_t u(t) + \mathscr{A}u(t) &= f(\cdot,t), \text{ for all } t \in (0,T], \\ u(\boldsymbol{x},0) &= u_0(\boldsymbol{x}), \text{ for } \boldsymbol{x} \in \overline{\Omega}, \\ u(\boldsymbol{x},t) &= 0, \text{ for } (\boldsymbol{x},t) \in \partial\Omega \times (0,T]. \end{aligned} \tag{3.2}$$

We consider the case where $u_0 \in \mathrm{L}_2(\Omega)$ and $f \in \mathrm{L}_2(0,T;\mathrm{L}_2(\Omega))$ for which the problem (3.2) admits a unique solution [Eva98].

Problem (3.2) is understood in the following weak form

$$\begin{aligned} \langle \partial_t u(t), \phi \rangle + a(u(t), \phi) &= \langle f(t), \phi \rangle \quad \forall\, \phi \in \mathrm{H}^1_0(\Omega),\, t \in (0, T] \\ u(\cdot, 0) &= u_0(\cdot), \end{aligned} \tag{3.3}$$

where $a(\phi, \psi) := \langle \nabla\phi, \nabla\psi \rangle$. The form $a(\cdot, \cdot)$ is clearly bounded

$$a(\phi, \psi) \leq \beta \, \|\phi\|_1 \, \|\psi\|_1 \quad \forall\, \phi, \psi \in \mathrm{H}^1_0(\Omega) \tag{3.4}$$

and coercive

$$a(\phi, \phi) \geq \alpha \, \|\phi\|_1^2 \quad \forall\, \phi \in \mathrm{H}^1_0(\Omega), \tag{3.5}$$

where $\alpha = (1 + C_{\mathrm{P}}^2)^{-1}$ and C_{P} is the Poincaré constant. From §2 the bilinear form defines an inner product on $\mathrm{H}^1_0(\Omega)$ and hence we can denote the energy norm $\|\cdot\|_a^2 := a(\cdot, \cdot)$. These observations justify our use of $\|\cdot\|_a$ (instead of $\|\cdot\|_{\mathrm{H}^1(\Omega)}$) as the norm of $\mathrm{H}^1_0(\Omega)$ to be with the implied dual norm on $\mathrm{H}^{-1}(\Omega)$.

3.2.2 Spatial discretisation

As in §2 let $\mathscr{T}$ be a conforming, not necessarily quasiuniform, triangulation of Ω, fix an integer $p \geq 1$, and consider the *finite element space*

$$\mathbb{V} := \left\{ \Phi \in \mathrm{H}^1_0(\Omega) : \ \Phi|_K \in \mathbb{P}^p \, \forall\, K \in \mathscr{T} \right\}; \tag{3.6}$$

The *spatially discrete finite element solution* in $\mathbb{V}$, is the function $U : [0, T] \to \mathbb{V}$ such that

$$\begin{aligned} \langle \partial_t U, \Phi \rangle + a(U, \Phi) &= \langle f, \Phi \rangle \quad \forall\, \Phi \in \mathbb{V}, \\ U(\boldsymbol{x}, 0) &= U^0 := \mathit{\Pi}^{\mathbb{V}} u_0(\boldsymbol{x}) \quad \forall\, \boldsymbol{x} \in \Omega, \end{aligned} \tag{3.7}$$

where $\mathit{\Pi}^{\mathbb{V}} : \mathrm{L}_2(\Omega) \to \mathbb{V}$ is a suitable projector (or an interpolator if the data u_0 is in a higher regularity subspace of $\mathrm{L}_2(\Omega)$, e.g., $\mathscr{T}$-wise continuous).

3.2.2.1 Definition (discrete Laplacian)**.** The *discrete Laplacian*, $A : \mathbb{V} \to \mathbb{V}$ is defined, through the Riesz representation in $\mathbb{V}$, by

$$\langle AV, \Phi \rangle = a(V, \Phi) \quad \forall\, \Phi \in \mathbb{V}, \tag{3.8}$$

3.2.2.2 Definition ($\mathrm{L}_2(\Omega)$-projection operator)**.** Let $v \in \mathrm{L}_2(\Omega)$ then the $\mathrm{L}_2(\Omega)$-*projection operator*, $P^{\mathbb{V}} : \mathrm{L}_2(\Omega) \to \mathbb{V}$, is defined through the following

$$\left\langle P^{\mathbb{V}} v, \Phi \right\rangle = \langle v, \Phi \rangle \qquad \forall\, \Phi \in \mathbb{V}, \tag{3.9}$$

that is $P^{\mathbb{V}} v - v \perp \mathbb{V}$.

We will often write the scheme (3.7) in its *pointwise form*

$$\partial_t U + AU = P^{\mathbb{V}} f \text{ and } U(0) = U^0. \tag{3.10}$$

The pointwise form is convenient as it allows for a more compact notation.

3.2.3 Fully discrete scheme

We subdivide the time interval $[0, T]$ into a partition of N consecutive adjacent subintervals whose endpoints are denoted $t_0 = 0 < t_1 < \ldots < t_N = T$. The n-th timestep is defined as $\tau_n := t_n - t_{n-1}$. In this chapter we will consistently use the shorthand $F^n(\cdot) := F(\cdot, t_n)$ for a generic time function F. A similar notation is used for time dependant function spaces.

The backward Euler method consists in finding a sequence of functions, $U^n \in \mathbb{V}^n$, such that for each $n = 1, \ldots, N$ we have:

$$\begin{aligned} \frac{1}{\tau_n} \left\langle U^n - \Lambda^n U^{n-1}, \Phi \right\rangle + a(U^n, \Phi) = \langle f^n, \Phi \rangle \qquad & \forall\, \Phi \in \mathbb{V}^n, \\ U^0 = \Pi^0 u_0, & \end{aligned} \tag{3.11}$$

where $\Lambda^{\mathbb{V}} : \mathrm{C}^0(\Omega) \to \mathbb{V}$ denotes the Lagrange interpolation operator (see Definition 2.3.0.4), $\Lambda^n := \Lambda^{\mathbb{V}^n}$, and Π^0 is defined as $\Pi^{\mathbb{V}}$. Note our nonrestrictive use of the Lagrange interpolator as a "data-transfer" operator from a finite element space to the next. We do this to reflect exactly what we do in practical computations (where interpolation is faster than averaging). All our analysis applies, however to a different data-transfer operator, including the $\mathrm{L}_2(\Omega)$ projector, if necessary.

As with the semidiscrete scheme the fully discrete scheme can be written in a pointwise form as follows:

$$\frac{U^n - \Lambda^n U^{n-1}}{\tau_n} + A^n U^n = P^n f^n \text{ and } U^0 = \Pi^0 u_0, \tag{3.12}$$

where $A^n = A^{\mathbb{V}^n}$ and $P^n = P^{\mathbb{V}^n}$ (cf. (3.8)).

3.3 Apriori analysis

Before submerging ourselves in the aposteriori error analysis we give an apriori analog to the method we employ in the aposteriori case.

We begin this section by introducing the Ritz projection, a crucial operator in the apriori analysis of the model problem.

3.3.0.1 Definition (Ritz (elliptic) projection)**.** Given $v \in \mathrm{H}_0^1(\Omega)$, the *Ritz projection* $R^h v : \mathrm{H}_0^1(\Omega) \to \mathbb{V}$ is nothing but the finite element solution of the corresponding elliptic problem, that is,

$$a\Big(R^h v, \Phi\Big) = a(v, \Phi) \qquad \forall\, \Phi \in \mathbb{V}. \tag{3.13}$$

3.3.0.2 Remark (error bound for the Ritz projection)**.** Given $v \in \mathrm{H}^s(\Omega) \cap \mathrm{H}_0^1(\Omega)$ for some $s \in 1, \ldots, p+1$, (recall p is the degree of $\mathbb{V}$) from §2.3 it is clear that the Ritz projection satisfies the following apriori error bound

$$\left\| R^h v - v \right\|_a \leq C h^{s-1} \left| v \right|_s . \tag{3.14}$$

3.3.0.3 Theorem (apriori error bound for the semidiscrete scheme)**.** *Let u and U be the solutions of (3.3) and (3.7) respectively then the following error bound holds, given that $u \in \mathrm{H}^s(\Omega) \cap \mathrm{H}_0^1(\Omega)$ for some $s \in 1, \ldots, p+1$*

$$\left\| (U-u)(t) \right\|^2 + \int_0^t \left\| (U-u)(s) \right\|_a^2 \, \mathrm{d}s \leq C h^{2s-2} \left\| u \right\|^2_{\mathrm{L}_2(0,t;\mathrm{H}^s(\Omega))} . \tag{3.15}$$

Proof We begin by splitting the error into two parts via the Ritz projection, the *parabolic* (ρ) and *elliptic* (ϵ), that is

$$e(t) = (u-U)(t) = \rho(t) - \epsilon(t) = (R^h u - U)(t) - (R^h u - u)(t). \tag{3.16}$$

Using the pointwise form of the schemes

$$\partial_t \rho + A\rho = \partial_t [R^h u - U] + A[R^h u - U] = \partial_t R^h u + A R^h u - P^{\mathbb{V}} f = \partial_t [R^h u - u] = \partial_t \epsilon. \tag{3.17}$$

This readily implies that

$$\partial_t e + Ae = A\epsilon. \tag{3.18}$$

Testing (3.18) with e gives

$$\frac{1}{2}\,\mathrm{d}_t \|e\|^2 + \|e\|_a^2 = a(\epsilon, e)\,. \tag{3.19}$$

Applying Young's inequality and integrating from 0 to t we see

$$\begin{aligned} \|e\|^2 + \int_0^t \|e\|_a^2 &\leq \int_0^t \|\epsilon\|_a^2 \\ &\leq \int_0^t \left\| R^h u - u \right\|_a^2 \\ &\leq C h^{2s-2} \|u\|^2_{\mathrm{L}_2(0,t;\mathrm{H}^s(\Omega))}\,, \end{aligned} \tag{3.20}$$

as claimed. □

3.3.0.4 Remark (optimality of Theorem 3.3.0.3 in $\mathrm{L}_2(0,t;\mathrm{L}_2(\Omega))$)**.** Note that for long integration times this method of apriori analysis leads to suboptimal estimates in the pointwise in time $\mathrm{L}_2(\Omega)$ error. In this work we are not concerned with estimation in this norm we are interested in $\mathrm{H}_0^1(\Omega)$ estimates. This method provides us with with an "optimal" constant, in the sense it *only* depends on the "elliptic error". The reasoning for which will become apparent in the next section.

3.4 Semidiscrete aposteriori estimate

To make the link between the parabolic problem and the elliptic recovered gradient estimates from §2.4 we utilise the elliptic reconstruction technique [MN03, LM06]. To make the discussion more accessible, we first do this for the spatially semidiscrete scheme. We divide the error into two parts—one called the elliptic error the other parabolic error—via the *elliptic reconstruction of the discrete solution*. This is an aposteriori analog of the apriori analysis already presented in §2.3.

3.4.0.5 Assumption (elliptic aposteriori error estimates)**.** We will consider henceforth the blanket assumption that for a fixed h_0, there are some $c_0 < C_0$, such that for any $\mathbb{V}$ with mesh-size $h < h_0$, for w and W solutions of the corresponding elliptic problem, find $w \in \mathrm{H}_0^1(\Omega)$ such that

$$a(w, \phi) = \langle g, \phi \rangle \quad \forall\, \phi \in \mathrm{H}_0^1(\Omega) \tag{3.21}$$

and its finite element approximation, find $W \in \mathbb{V}$ such that

$$a(W, \Phi) = \langle g, \Phi \rangle \quad \forall\, \Phi \in \mathbb{V} \tag{3.22}$$

and $\mathscr{E}$ defined in Definition 2.4.2.6 the following bounds are true

$$c_0\mathscr{E}[W] \leq \|\nabla[W-w]\| \leq C_0\mathscr{E}[W]\,. \tag{3.23}$$

Optionally, we will assume *asymptotic exactness*, in which case

$$C_0 \leq 1 + B(h_0) \text{ and } c_0 \geq 1 + \beta(h_0), \tag{3.24}$$

for some continuous functions B and β that vanish at 0.

3.4.0.6 Remark (efficiency of the elliptic estimator)**.** The lower bound in (3.23) is not needed for the theory to be developed herein, as we will prove only upper bounds. Nonetheless, this property is required for the efficiency of the parabolic estimators in practical situations.

Because the elliptic error can be directly bounded under the blanket Assumption 3.4.0.5, it is enough to show that the full error can be bounded in terms of the elliptic error only. This result is in accordance with the fact that the parabolic error on uniform meshes is of higher h-order in the energy norm with respect to the elliptic (and thus the full) error, as observed by Lakkis & Makridakis [LM06].

3.4.0.7 Definition (elliptic reconstruction)**.** The *elliptic reconstruction operator* is defined as $\mathscr{R} : \mathbb{V} \to \mathrm{H}^1_0(\Omega)$ such that

$$\mathscr{A}[\mathscr{R}V] = AV, \tag{3.25}$$

where A is the discrete elliptic operator defined in (3.8). In weak form, equation (3.25) reads

$$a(\mathscr{R}V, \Phi) = \langle AV, \Phi\rangle \qquad \forall\, \Phi \in \mathrm{H}^1_0(\Omega), \tag{3.26}$$

and it is well defined in virtue of the elliptic problem's well posedness. We will refer to the function $\mathscr{R}V$ as the *elliptic reconstruction* of V, while the elliptic reconstruction operator $\mathscr{R}$ will be called the *reconstruction operator* (or just the *reconstructor*) from $\mathbb{V}$.

If $U(t)$ denotes the solution of (3.10) at time t, we shall indicate by $\omega(t)$ its reconstruction $\mathscr{R}U(t)$.

Thus, setting $g(t) := AU(t)$, we then see $U(t)$ is the finite element solution corresponding to the elliptic problem of finding $\omega(t) \in \mathrm{H}^1_0(\Omega)$ such that $\mathscr{A}\omega(t) = g(t)$.

3.4.1 The error and its splitting

For the whole of this section we shall consider u to be the solution of (3.2), understood in the weak sense, and U its semidiscrete approximation given by (3.10). The corresponding *semidiscrete error* is defined by

$$e(t) := U(t) - u(t), \tag{3.27}$$

and can be split, using the elliptic reconstruction $\omega = \mathscr{R}U$ from Definition 3.4.0.7, as follows:

$$e(t) = (\omega(t) - u(t)) - (\omega(t) - U(t)) =: \rho(t) - \epsilon(t). \tag{3.28}$$

We shall refer to ϵ and ρ here defined as the *elliptic (reconstruction) error* and the *parabolic error* respectively.

Using this notation we have the estimate

$$\|\nabla[U - u](t)\| \leq \|\nabla\rho(t)\| + \|\nabla\epsilon(t)\|, \tag{3.29}$$

where, following the remarks made in Definition 3.4.0.7 and Assumption 3.4.0.5, the elliptic error can be bounded by the computable elliptic aposteriori estimator functional $\mathscr{E}$:

$$\|\epsilon(t)\|_a = \|\nabla\epsilon(t)\| \leq C_0\mathscr{E}[U(t)]. \tag{3.30}$$

It is therefore sufficient to bound the error's energy norm using the elliptic error's energy norm.

3.4.1.1 Lemma (elliptic energy bound for parabolic semidiscrete error)**.** *If e, ϵ are defined as in §3.4.1 then, for each $t \in [0, T]$, we have*

$$\|e(t)\|^2 + \int_0^t \|e(s)\|_a^2 \, \mathrm{d}s \leq \|e(0)\|^2 + \int_0^t \|\epsilon(s)\|_a^2 + 2\left\langle P^{\mathbb{V}}f(s) - f(s), e(s)\right\rangle \, \mathrm{d}s. \tag{3.31}$$

Proof From the the exact problem (3.2), the semidiscrete scheme (3.10), and the splitting (3.28) we have

$$\partial_t e + \mathscr{A}\rho = \partial_t[U - u] + \mathscr{A}[\omega - u] = \partial_t U + AU - \partial_t u - \mathscr{A}u = P^{\mathbb{V}}f - f. \tag{3.32}$$

Testing with e we obtain

$$\langle \partial_t e, e\rangle + a(\rho, e) = \left\langle P^{\mathbb{V}} f - f, e\right\rangle \tag{3.33}$$

and thus

$$\frac{1}{2}\,\mathrm{d}_t\|e\|^2 + \|e\|_a^2 = \left\langle P^{\mathbb{V}} f - f, e\right\rangle - a(\epsilon, e)\,, \tag{3.34}$$

where we use $\mathrm{d}_t = \frac{\mathrm{d}}{\mathrm{d}t}$. Integration from 0 to t yields

$$\|e(t)\|^2 + 2\int_0^t \|e\|_a^2 = \|e(0)\|^2 + 2\int_0^t \left\langle P^{\mathbb{V}} f - f, e\right\rangle - 2\int_0^t a(\epsilon, e) \quad \forall\, t \in [0,T]\,. \tag{3.35}$$

Hence, by Young's inequality on $a(\epsilon, e)$, we have

$$\|e(t)\|^2 + 2\int_0^t \|e\|_a^2 \le \|e(0)\|^2 + 2\int_0^t \left\langle P^{\mathbb{V}} f - f, e\right\rangle + \int_0^t \|e\|_a^2 + \int_0^t \|\epsilon\|_a^2\,, \tag{3.36}$$

whereby the claim is verified. □

3.4.1.2 Proposition (L_2 simplification rule)**.** *If* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^N$, $N \in \mathbb{N}$, $c \in \mathbb{R}$ *and* $f, g \in \mathrm{L}_2(D)$, *for some measurable domain* D, *are such that*

$$|\mathbf{a}|^2 + \|f\|^2 \le c^2 + \mathbf{a}^{\intercal}\mathbf{b} + \int_D fg, \tag{3.37}$$

then

$$\left(|\mathbf{a}|^2 + \|f\|^2\right)^{1/2} \le |c| + \left(|\mathbf{b}|^2 + \|g\|^2\right)^{1/2}, \tag{3.38}$$

where all the vector norms are Euclidean, and the function norms $\mathrm{L}_2(D)$.

Proof

Denote by $\boldsymbol{\alpha} := (|\mathbf{a}|\,, \|f\|)$ and $\boldsymbol{\beta} := (|\mathbf{b}|\,, \|g\|)$.

If $|\boldsymbol{\alpha}| \le |\boldsymbol{\beta}|$ then (3.38) is trivially satisfied. Otherwise we have $|\boldsymbol{\alpha}| > |\boldsymbol{\beta}|$ whereby (3.37) and the Cauchy–Bunyakovskǐ–Schwarz inequality imply that

$$\begin{aligned} |\boldsymbol{\alpha}|^2 &\le c^2 + |\mathbf{a}|\,|\mathbf{b}| + \|f\|\,\|g\| + |\boldsymbol{\beta}|(|\boldsymbol{\alpha}| - |\boldsymbol{\beta}|) \\ &\le c^2 + 2\,|\boldsymbol{\alpha}|\,|\boldsymbol{\beta}| - |\boldsymbol{\beta}|^2\,. \end{aligned} \tag{3.39}$$

Hence $(|\boldsymbol{\alpha}| - |\boldsymbol{\beta}|)^2 \le c^2$, and thereby

$$|\boldsymbol{\alpha}| \le |c| + |\boldsymbol{\beta}|\,, \tag{3.40}$$

as claimed. □

3.4.1.3 Theorem (aposteriori semidiscrete error estimate)**.** *With u and U as defined by (3.2) and (3.7), respectively, and an estimator functional $\mathscr{E}$ as defined in (2.50), we have*

$$\left(\|U(t)-u(t)\|^2+\int_0^t\|U-u\|_a^2\right)^{1/2} \leq \|U(0)-u(0)\|+C_0\,\|\mathscr{E}[U]\|_{\mathrm{L}_2[0,T]}+2\left\|P^{\mathbb{V}}f-f\right\|_{\mathrm{L}_2(0,T;\mathrm{H}^{-1}(\Omega))}. \quad (3.41)$$

Proof Using Lemma 3.4.1.1 we have

$$\|e(t)\|^2+\int_0^t\|e\|_a^2\leq\|e(0)\|^2+\int_0^t\|\epsilon\|_a^2+2\int_0^t\left\langle P^{\mathbb{V}}f-f,e\right\rangle. \quad (3.42)$$

Using Proposition 3.4.1.2, we obtain

$$\left(\|e(t)\|^2+\int_0^t\|e\|_a^2\right)^{1/2}\leq\left(\|e(0)\|^2+\int_0^t\|\epsilon\|_a^2\right)^{1/2}+2\left(\int_0^t\left\|P^{\mathbb{V}}f-f\right\|_{\mathrm{H}^{-1}(\Omega)}^2\right)^{1/2}. \quad (3.43)$$

Assumption (3.23) and the discussion in §3.4.1 ensure then that

$$\left(\|e(t)\|^2+\int_0^t\|e\|_a^2\right)^{1/2}\leq\left(\|e(0)\|^2+C_0^2\int_0^t\mathscr{E}[U]^2\right)^{1/2}+2\left(\int_0^t\left\|P^{\mathbb{V}}f-f\right\|_{\mathrm{H}^{-1}(\Omega)}^2\right)^{1/2}, \quad (3.44)$$

which implies the claim. □

3.4.1.4 Remark (short versus long integration times)**.** The bound for the pointwise in time $\mathrm{L}_2(\Omega)$ error, $\|e(t)\|$, appearing on the left-hand side of (3.41), is tight only for very short times. For example, it is well-known that on a uniform mesh of size $h\to 0$ on a convex domain Ω the energy term $\left(\int_0^t\|e\|_a^2\right)^{1/2}$is $\mathrm{O}(h^p)$, while $\|e(t)\|$ is $\mathrm{O}(h^{p+1})$.

3.4.1.5 Remark (dealing with the $\mathrm{H}^{-1}(\Omega)$ norm)**.** If we are lacking apriori information, the last term in (3.41) may be replaced using the Poincaré inequality

$$2\left\|P^{\mathbb{V}}f-f\right\|_{\mathrm{L}_2(0,T;\mathrm{H}^{-1}(\Omega))}\leq 2C_{\mathrm{P}}(\Omega)\left\|P^{\mathbb{V}}f-f\right\|_{\mathrm{L}_2(\Omega\times(0,T))}. \quad (3.45)$$

It is also possible to obtain bounds by using the Cauchy–Bunyakovskĭ–Schwarz inequality for $\mathrm{L}_2(\Omega)$ on the term $\left\langle P^{\mathbb{V}}f-f,e\right\rangle$—rather than the $(\mathrm{H}^{-1},\mathrm{H}_0^1)$ duality—and "absorb" the resulting $\sup_{[0,t]}\|e\|$ into the first term on the right hand side of (3.41). However, whenever possible, we shy away from this procedure as it incurs artificially higher constants and an $\mathrm{L}_1[0,T]$ accumulation on the right-hand side while the energy term on the left-hand side accumulates like $\mathrm{L}_2[0,T]$. This time-accumulation disparity

between the error and the estimator is likely to result in an error–estimator ratio bound that has the order of $\sqrt{T}$, that is, although having the right order of convergence, the estimator will overestimate the error over long integration times.

3.4.1.6 Remark (sharper versions of Theorem 3.4.1.3)**.** The error estimate (3.41) can be tightened further to

$$\left(\frac{1}{2}\|e(t)\|^2 + \int_0^t \|e\|_a^2\right)^{1/2} \leq \frac{1}{\sqrt{2}}\|e(0)\| + \left(\int_0^t \left\|P^{\mathbb{V}} f - f\right\|_{\mathrm{H}^{-1}(\Omega)} + C_0^2 \mathscr{E}[U]^2\right)^{1/2}. \quad (3.46)$$

But this estimate becomes noticeably better only when one of the terms $\|e(0)\|$ or $\left\|P^{\mathbb{V}} f - f\right\|_{\mathrm{H}^{-1}(\Omega)}$ dominates the $\mathscr{E}[U]$ term, which should not be allowed to happen. So there is no need to lengthen the discussion by insisting on such tight bounds, as long as it is possible to obtain the elliptic aposteriori estimate constant C_0 in the leading term on the right-hand side.

3.5 Fully discrete aposteriori estimate

The main result of this section is the aposteriori error bound, stated in Theorem 3.5.2.4, on the error between the approximate solution U of the fully discrete problem (3.12) and that of the exact problem (3.2).

The analysis in this section follows narrowly the one we performed in §3.4, albeit with the complications that the fully discrete scheme brings. We will first extend the discrete solution sequence to a continuous-time function. Then we derive an error identity on which we mimic the energy techniques of §3.4 to bound the error's energy norm in terms of some residual terms and the elliptic error's energy norm, which is finally controlled via gradient recovery estimators.

3.5.1 Time extension of the discrete solution

Recalling the fully discrete scheme (3.12), the fully discrete solution is the sequence of finite element functions $U^n \in \mathbb{V}^n$ defined at each discrete time t_n, $n = 0, \ldots, N$. Define the piecewise linear (affine) extension

$$U(t) := \sum_{n=0}^{N} U^n l_n(t), \quad (3.47)$$

where we use the one-dimensional piecewise linear continuous Lagrange basis functions, defined for $t \geq 0$, as

$$l_n(t) := \begin{cases} (t - t_{n-1})/\tau_n, & \text{for } t \in (t_{n-1}, t_n] \ \text{(and } n > 0), \\ (t_{n+1} - t)/\tau_{n+1}, & \text{for } t \in (t_n, t_{n+1}] \\ 0, & \text{otherwise.} \end{cases} \tag{3.48}$$

We warn the reader that we use the same symbol, U, to indicate the fully discrete solution's extension to $[0, T]$, as the one we used for its semidiscrete counterpart in §3.4.

3.5.2 Elliptic reconstruction and error splitting

Next we define the elliptic reconstruction, needed for the following analysis, similarly to that of the semidiscrete scheme (cf. (3.25)). For each $n \in [0 : N]$, with the discrete elliptic operator A^n as in §3.2.3, we define the corresponding elliptic reconstruction operator $\mathscr{R}^n : \mathbb{V}^n \to \mathrm{H}^1_0(\Omega)$, for each $V \in \mathbb{V}^n$, by solving for $\mathscr{R}^n V$ the elliptic problem

$$\mathscr{A}\mathscr{R}^n V = A^n V, \tag{3.49}$$

which can be read in the weak form as

$$a(\mathscr{R}^n V, \Phi) = \langle A^n V, \Phi \rangle \quad \forall\, \Phi \in \mathrm{H}^1_0(\Omega). \tag{3.50}$$

We denote

$$\omega^n := \mathscr{R}^n U^n, \ \text{for each } n = 0, \ldots, N, \tag{3.51}$$

and this sequence's piecewise linear extension by $\omega : [0, T] \to \mathrm{H}^1_0(\Omega)$, i.e.,

$$\omega(t) := \sum_{n=0}^{N} \omega^n l_n(t). \tag{3.52}$$

As in the semidiscrete analysis we introduce symbols for the *full error* $e := U - u$, the *elliptic error* $\epsilon := \omega - U$ and the *parabolic error* $\rho := \omega - u$, whereby

$$e = \rho - \epsilon, \tag{3.53}$$

and, based on the Assumption 3.4.0.5,

$$\begin{aligned} \|\epsilon(t)\|_a &\leq C_0 \mathscr{E}\big[U^n l_n(t) + U^{n-1} l_{n-1}(t)\big] \\ &\leq C_0 \big(\mathscr{E}[U^n]\, l_n(t) + \mathscr{E}\big[U^{n-1}\big]\, l_{n-1}(t)\big) \ \text{for } t \in [t_{n-1}, t_n]\,. \end{aligned} \tag{3.54}$$

The last step is guaranteed by the linearity of the operators G and ∇, hence the homogeneity $\mathscr{E}[\lambda V] = |\lambda| \, \|GV - \nabla V\|$, and by the triangle inequality.

3.5.2.1 Lemma (parabolic error identity)**.** *For each $n = 1, \dots, N$ and each $t \in (t_{n-1}, t_n)$ we have*

$$\partial_t e(t) + \mathscr{A}\rho(t) = (\Lambda^n U^{n-1} - U^{n-1})/\tau_n + \mathscr{A}[\omega(t) - \omega^n] + P^n f^n - f(t). \tag{3.55}$$

Proof By the definition of U, (3.47), for each $n = 1, \dots, N$ and $t \in (t_{n-1}, t_n)$ we have

$$\partial_t U(t) = U^n l_n'(t) + U^{n-1} l_{n-1}'(t) = (U^n - U^{n-1})/\tau_n \tag{3.56}$$

and using the fully discrete scheme (3.12), we have

$$\begin{aligned} \partial_t U(t) + \mathscr{A}\omega^n &= (\Lambda^n U^{n-1} - U^{n-1})/\tau_n + (U^n - \Lambda^n U^{n-1})/\tau_n + A^n U^n \\ &= (\Lambda^n U^{n-1} - U^{n-1})/\tau_n + P^n f^n. \end{aligned} \tag{3.57}$$

Hence

$$\partial_t U(t) + \mathscr{A}\omega(t) = (\Lambda^n U^{n-1} - U^{n-1})/\tau_n + \mathscr{A}[\omega(t) - \omega^n] + P^n f^n \tag{3.58}$$

and, using the exact PDE (3.2), we get

$$\begin{aligned} \partial_t e(t) + \mathscr{A}\rho(t) &= \partial_t U(t) + \mathscr{A}\omega(t) - \partial_t u(t) - \mathscr{A}u(t) \\ &= (\Lambda^n U^{n-1} - U^{n-1})/\tau_n + \mathscr{A}[\omega(t) - \omega^n] + P^n f^n - f(t), \end{aligned} \tag{3.59}$$

as stated. □

3.5.2.2 Definition (aposteriori error indicators)**.** The notation we introduce here will be valid for the rest of this section.

elliptic error indicator via recovery

$$\varepsilon_n := \mathscr{E}\big[U^n, \mathrm{H}_0^1(\Omega), \mathbb{V}^n\big] = C_0 \, \|\nabla U^n - G^n[U^n]\| \,, \tag{3.60}$$

with the functional $\mathscr{E}$ as defined in §2.4.2.6, and[1]

$$\tilde{\varepsilon}_n^2 := \frac{1}{3}\big(\varepsilon_n^2 + \varepsilon_{n-1}^2 + \varepsilon_n \varepsilon_{n-1}\big) \le \frac{1}{2}\big(\varepsilon_n^2 + \varepsilon_{n-1}^2\big) \,. \tag{3.61}$$

[1] In the numerical experiments we use $(\varepsilon_n^2 + \varepsilon_{n-1}^2)/2$ instead of $\tilde{\varepsilon}_n$.

time-discretisation error indicators

$$\theta_n := \frac{1}{\sqrt{3}} \begin{cases} \left\| P^n f^n - \Lambda^n \partial U^n - \left(P^{n-1} f^{n-1} - \Lambda^{n-1} \partial U^{n-1} \right) \right\|_{\mathrm{H}^{-1}(\Omega)} & \text{for } n \geq 2, \\ \left\| P^1 f^1 - \Lambda^1 \partial U^1 - A^0 U^0 \right\|_{\mathrm{H}^{-1}(\Omega)} & \text{for } n = 1, \end{cases} \tag{3.62}$$

where $\partial U^n := (U^n - U^{n-1})/\tau_n$, (cf. Lemma 4.7.0.15), also possible to use in its alternative (faster to compute but not as sharp) version

$$\tilde{\theta}_n := C_\mu \left\| U^{n-1} - U^n \right\|_a , \tag{3.63}$$

where C_μ is dependent on the shape regularity μ of the family of triangulations defined in (2.22).

mesh-change (coarsening) error indicators a main mesh-change indicator

$$\gamma_n := \tau_n^{-1} \left\| \Lambda^n U^{n-1} - U^{n-1} \right\|_{\mathrm{H}^{-1}(\Omega)} , \tag{3.64}$$

and a *higher order* mesh-change indicator

$$\tilde{\gamma}_n := {C_\mu}' \begin{cases} \left\| \hat{h}_n \left(P^n f^n - \Lambda^n \partial U^n - P^{n-1} f^{n-1} + \Lambda^{n-1} \partial U^{n-1} \right) \right\| , & n \geq 2, \\ \left\| \hat{h}_1 \left(P^1 f^1 - \Lambda^1 \partial U^1 - A^0 U^0 \right) \right\| , & n = 1, \end{cases} \tag{3.65}$$

where $\hat{h}_n(\boldsymbol{x}) = \max \{ h_{n-1}(\boldsymbol{x}), h_n(\boldsymbol{x}) \}$ for $\boldsymbol{x} \in \Omega$ and a constant C'_μ.

data approximation error indicator

$$\beta_n := \tau_n^{-1} \int_{t_{n-1}}^{t_n} \left\| P^n f^n - f(t) \right\|_{\mathrm{H}^{-1}(\Omega)} \, \mathrm{d}t. \tag{3.66}$$

3.5.2.3 Remark (computing $\mathrm{H}^{-1}(\Omega)$ norms)**.** Clearly the $\mathrm{H}^{-1}(\Omega)$ norms appearing in Definition 3.5.2.2 cannot be computed in practise. The corresponding indicators can be replaced by upper bounds using the (dual) Poincaré inequality

$$\left\| \phi \right\|_{\mathrm{H}^{-1}(\Omega)} \leq C_{\mathrm{P}} \left\| \phi \right\| . \tag{3.67}$$

Other alternatives will be described in Lemma 4.7.0.15 and are possible but will not be discussed further here.

3.5.2.4 Theorem (aposteriori estimate for fully discrete scheme)**.** *Let the sequence* $(U^n)_{n\in[0:N]}$, $U^n \in \mathbb{V}^n$, *be the solution of the fully discrete problem* (3.11) *and* U *its piecewise linear time-extension as in* (3.47). *Let* u *be the exact solution of the exact problem* (3.2) *then*

$$\left(\frac{\left\|U^N - u(T)\right\|^2}{2} + \int_0^T \|U(t) - u(t)\|_a^2 \,\mathrm{d}t\right)^{1/2} \leq \frac{\|U(0) - u(0)\|}{\sqrt{2}} + \eta_N \tag{3.68}$$

where the (global) error estimator *is given by the following discrete* $\mathrm{L}_2(0,T)$ *summation of the error indicators defined in §3.5.2.2:*

$$\eta_N^2 = \sum_{n=1}^{N} (\tilde{\varepsilon}_n + \gamma_n + \beta_n + \theta_n)^2 \,\tau_n. \tag{3.69}$$

Proof The proof shadows that of Lemma 3.4.1.1 and Theorem 3.4.1.3, but we must take into account the complications arising from the time discretisation. For the reader's convenience we divide it into steps.

Step 1. Using the notation from Lemma 3.5.2.1 and identity (3.55) therein we have that

$$\begin{aligned} \partial_t e(t) + \mathscr{A} e(t) = \mathscr{A}\epsilon(t) + (\Lambda^n U^{n-1} - U^{n-1})/\tau_n \\ + \mathscr{A}[\omega(t) - \omega^n] + P^n f^n - f(t). \end{aligned} \tag{3.70}$$

Testing this with e we obtain

$$\begin{aligned} \frac{1}{2}\,\mathrm{d}_t \|e(t)\|^2 + \|e(t)\|_a^2 = a(\epsilon(t), e(t)) + \left\langle (\Lambda^n U^{n-1} - U^{n-1})/\tau_n, e(t)\right\rangle \\ + \left\langle \mathscr{A}[\omega(t) - \omega^n], e(t)\right\rangle + \left\langle P^n f^n - f(t), e(t)\right\rangle, \end{aligned} \tag{3.71}$$

for all $t \in (t_{n-1}, t_n)$ and each $n = 1, \ldots, N$. Integrating over $[0,T]$ gives

$$\begin{aligned} \left\|e^N\right\|^2/2 + \int_0^T \|e(t)\|_a^2 \,\mathrm{d}t = \left\|e^0\right\|^2/2 + \int_0^T a(\epsilon(t), e(t)) \,\mathrm{d}t \\ + \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \Big(\left\langle (\Lambda^n U^{n-1} - U^{n-1})/\tau_n, e(t)\right\rangle \\ + a(\omega(t) - \omega^n, e(t)) + \left\langle P^n f^n - f(t), e(t)\right\rangle \Big) \,\mathrm{d}t \\ =: \mathcal{B}_1 + \mathcal{B}_2 + \mathcal{B}_3 + \mathcal{B}_4 + \left\|e^0\right\|^2/2. \end{aligned} \tag{3.72}$$

We proceed by bounding each of the terms $\mathcal{B}_j$, $j = 1, \ldots, 4$, appearing in the right-hand side of (3.72).

Step 2. The first term to be bounded in (3.72) yields the spatial discretisation error indicator as follows:

$$\begin{aligned}\mathcal{B}_1 &= \int_0^T a(\epsilon(t), e(t))\ \mathrm{d}t = \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} a(\epsilon(t), e(t))\ \mathrm{d}t \\ &\leq \sum_{n=1}^{N} \bigg(\int_{t_{n-1}}^{t_n} \|\epsilon\|_a^2 \bigg)^{1/2} \bigg(\int_{t_{n-1}}^{t_n} \|e\|_a^2 \bigg)^{1/2} \leq \sum_{n=1}^{N} \tilde{\varepsilon}_n \tau_n^{1/2} \bigg(\int_{t_{n-1}}^{t_n} \|e\|_a^2 \bigg)^{1/2}\end{aligned} \tag{3.73}$$

where we have used (3.61) and in view of (3.54) and (3.61), we may write

$$\int_{t_{n-1}}^{t_n} \|\epsilon\|_a^2 \leq \varepsilon_{n-1}^2 \int_{t_{n-1}}^{t_n} l_{n-1}^2 + 2\varepsilon_{n-1}\varepsilon_n \int_{t_{n-1}}^{t_n} l_{n-1} l_n + \varepsilon_n^2 \int_{t_{n-1}}^{t_n} l_n^2 = \tilde{\varepsilon}_n^2 \tau_n. \tag{3.74}$$

The second term in (3.72) contains mesh-change term which we bound as follows:

$$\begin{aligned}\mathcal{B}_2 &= \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \big\langle (\Lambda^n U^{n-1} - U^{n-1})/\tau_n, e(t) \big\rangle\ \mathrm{d}t \\ &\leq \sum_{n=1}^{N} \big\| \Lambda^n U^{n-1} - U^{n-1} \big\|_{\mathrm{H}^{-1}(\Omega)} \tau_n^{-1} \int_{t_{n-1}}^{t_n} \|e(t)\|_a\ \mathrm{d}t \\ &\leq \sum_{n=1}^{N} \gamma_n \tau_n^{1/2} \bigg(\int_{t_{n-1}}^{t_n} \|e\|_a^2 \bigg)^{1/2}\end{aligned} \tag{3.75}$$

where γ_n is defined by (3.64).

Similarly the data error term is bounded as follows

$$\mathcal{B}_4 = \int_0^T \langle P^n f^n - f(t), e(t) \rangle\ \mathrm{d}t \leq \sum_{n=1}^{N} \beta_n \tau_n^{1/2} \bigg(\int_{t_{n-1}}^{t_n} \|e\|_a^2 \bigg)^{1/2}, \tag{3.76}$$

where β_n is defined in (3.66).

Step 3. The third term in (3.72) yields a time discretisation term and is a bit more involved to estimate. Using the definition of ω^n, ω and $\mathscr{R}^n$, given in (3.49) and (3.52), we

observe that

$$\begin{aligned}
\mathcal{B}_3 &= \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} a(\omega - \omega^n, e(t))\ \mathrm{d}t \\
&= \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} a\big(l_{n-1}(t)\mathscr{R}^{n-1}U^{n-1} + l_n(t)\mathscr{R}^n U^n - \mathscr{R}^n U^n, e(t)\big)\ \mathrm{d}t \\
&= \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} l_{n-1}(t) a\big(\mathscr{R}^{n-1}U^{n-1} - \mathscr{R}^n U^n, e(t)\big)\ \mathrm{d}t \\
&= \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} l_{n-1}(t) \left\langle A^{n-1}U^{n-1} - A^n U^n, e(t)\right\rangle \mathrm{d}t \\
&\le \sum_{n=1}^{N} \left\|A^{n-1}U^{n-1} - A^n U^n\right\|_{\mathrm{H}^{-1}(\Omega)} \left(\int_{t_{n-1}}^{t_n} l_{n-1}^2\right)^{1/2} \left(\int_{t_{n-1}}^{t_n} \|e\|_a^2\right)^{1/2} \\
&\le \sum_{n=1}^{N} \theta_n \tau_n^{1/2} \left(\int_{t_{n-1}}^{t_n} \|e\|_a^2\right)^{1/2},
\end{aligned} \tag{3.77}$$

where in the last passage we use the discrete scheme (3.12) for the substitution

$$A^n U^n = (\Lambda^n U^{n-1} - U^n)/\tau_n + P^n f^n \text{ for } n \ge 1. \tag{3.78}$$

Step 4. Grouping together (3.72), (3.73), (3.75), (3.76) and (3.77), we obtain

$$\begin{aligned}
&\left\|e^N\right\|^2/2 + \int_0^T \|e(t)\|_a^2\ \mathrm{d}t \\
&\qquad\qquad \le \left\|e^0\right\|^2/2 + \sum_{n=1}^{N} (\tilde{\varepsilon}_n + \gamma_n + \beta_n + \theta_n)\, \tau_n^{1/2} \left(\int_{t_{n-1}}^{t_n} \|e\|_a^2\right)^{1/2}.
\end{aligned} \tag{3.79}$$

Using an $\mathrm{L}_2(\Omega)$ simplification (cf. §3.4.1.2), we conclude that

$$\left(\frac{\left\|e^N\right\|^2}{2} + \int_0^T \|e(t)\|_a^2\ \mathrm{d}t\right)^{1/2} \le \frac{\left\|e^0\right\|}{\sqrt{2}} + \left(\sum_{n=1}^{N} (\tilde{\varepsilon}_n + \gamma_n + \beta_n + \theta_n)^2\, \tau_n\right)^{1/2}. \tag{3.80}$$

Referring to the notation in (3.47) and Definition 3.5.2.2, we obtain the result. □

3.5.2.5 Remark (the alternative time indicator)**.** Assuming there is no mesh change from time t_{n-1} to time t_n, then the discrete Laplacians A^{n-1} and A^n, defined in (3.8), coincide. Thus the time discretisation error indicator θ_n, which is part of the estimator η_N in Theorem 3.5.2.4, can be written as

$$\theta_n = \frac{1}{\sqrt{2}} \left\|A^n U^n - A^{n-1}U^{n-1}\right\|_{\mathrm{H}^{-1}(\Omega)} = \frac{\tau_n}{\sqrt{2}} \left\|A^n \partial_t U\right\|_{\mathrm{H}^{-1}(\Omega)}. \tag{3.81}$$

Using the form given in (3.62) and using the dual Poincaré inequality (3.67), this indicator is easily bounded.

In the next result we show that the indicator θ_n is equivalent, up to higher order terms, to the alternative time indicator $\tilde{\theta}_n$, defined in (3.63), which requires only an energy norm computation. This alternative time indicator, which is more common in energy estimates [Pic98, e.g.], $\tilde{\theta}_n$ is also more "natural", as it measures the time derivative in the energy norm as opposed to the H^{-1} norm of the time derivative of AU. Due to mesh-change effects, this simpler indicator comes at the (affordable) price of having to add the higher order mesh change term $\tilde{\gamma}_n$ to the otherwise simpler γ_n.

3.5.2.6 Theorem (alternative time estimator)**.** *With the same assumptions and notation of Theorem 3.5.2.4 we have*

$$\left(\frac{\left\|U^N - u(T)\right\|^2}{2} + \int_0^T \|U(t) - u(t)\|_a^2 \,\mathrm{d}t\right)^{1/2} \leq \frac{\|U(0) - u(0)\|}{\sqrt{2}} + \tilde{\eta}_N \tag{3.82}$$

where the (alternative global) error estimator *is given by the following discrete* $\mathrm{L}_2(0,T)$ *summation of the error indicators defined in §3.5.2.2:*

$$\tilde{\eta}_N^2 := \sum_{n=1}^{N} \left(\tilde{\varepsilon}_n + \gamma_n + \tilde{\gamma}_n + \beta_n + \tilde{\theta}_n\right)^2 \tau_n. \tag{3.83}$$

Proof We proceed similarly to the proof of Theorem 3.5.2.4, in steps. The notation is the same and steps 1 and 2 are identical.

Step 3. This step starts similarly to its homologue in the proof of Theorem 3.5.2.4 by observing that

$$\mathcal{B}_3 = \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} l_{n-1}(t) \left\langle A^{n-1}U^{n-1} - A^nU^n, e(t)\right\rangle \,\mathrm{d}t. \tag{3.84}$$

The function $A^{n-1}U^{n-1} - A^nU^n$ belongs to $\mathbb{V}^n + \mathbb{V}^{n-1}$, but in general it is in neither of $\mathbb{V}^n$ nor $\mathbb{V}^{n-1}$). Thus, to proceed, we use the $\mathrm{L}_2(\Omega)$-projection and the Clément–Scott–Zhang interpolator denoted respectively by

$$\check{P}^n : \mathrm{L}_2(\Omega) \to \mathbb{V}^n + \mathbb{V}^{n-1} \text{ and } \hat{\Pi}^n : \mathrm{L}_2(\Omega) \to \mathbb{V}^n \cap \mathbb{V}^{n-1}. \tag{3.85}$$

We recall that the operators $\hat{\Pi}^n$ and $\check{P}^n$ are both known [SZ90, Car02, resp.] to enjoy

the following stability properties for all $n = 0, \dots, N$:

$$\left\| \hat{\Pi}^n \phi \right\|_a \leq C_{1,\mu} \|\phi\|_a \quad \forall\, \phi \in \mathrm{H}^1(\Omega), \tag{3.86}$$

$$\left\| \check{P}^n \phi \right\|_a \leq C_{2,\mu} \|\phi\|_a \quad \forall\, \phi \in \mathrm{H}^1(\Omega), \tag{3.87}$$

where μ is the shape-regularity of the triangulation family $\{\mathscr{T}^n\}_{n=0,\dots,N}$ defined in (2.22). Furthermore, the following interpolation inequality is valid [LM06, §B.3]

$$\left\| (\psi - \hat{\Pi}^n \psi)/\hat{h}_n \right\| \leq C_{3,\mu} \|\psi\|_a \quad \forall\, \psi \in \mathrm{H}^1_0(\Omega),\, n = 1, \dots, N, \tag{3.88}$$

where $\hat{h}_n := \max\{h_n, h_{n-1}\}$.

Step 4. Using these operators, we derive that

$$\begin{aligned}
\mathcal{B}_3 &= \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \left\langle A^{n-1}U^{n-1} - A^nU^n, \check{P}^n e(t) \right\rangle l_{n-1}(t)\,\mathrm{d}t \\
&= \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \Big(\left\langle A^{n-1}U^{n-1} - A^nU^n, \check{P}^n e(t) - \hat{\Pi}^n \check{P}^n e(t) \right\rangle \\
&\qquad + \left\langle A^{n-1}U^{n-1} - A^nU^n, \hat{\Pi}^n \check{P}^n e(t) \right\rangle \Big) l_{n-1}(t)\,\mathrm{d}t \\
&\leq \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \Big(\left\| \hat{h}_n \big(A^{n-1}U^{n-1} - A^nU^n\big) \right\| \left\| \hat{h}_n^{-1} \big(\check{P}^n e(t) - \hat{\Pi}^n \check{P}^n e(t)\big) \right\| \\
&\qquad + a\big(U^{n-1} - U^n, \hat{\Pi}^n \check{P}^n e(t)\big) \Big) l_{n-1}(t)\,\mathrm{d}t.
\end{aligned} \tag{3.89}$$

Using inequalities (3.86), (3.87) and (3.88), we get the bound

$$\begin{aligned}
\mathcal{B}_3 &\leq \sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \big(C_{3,\mu} \left\| \hat{h}_n (A^{n-1}U^{n-1} - A^nU^n) \right\| \left\| \check{P}^n e(t) \right\|_a \\
&\qquad + C_{1,\mu} \left\| U^{n-1} - U^n \right\|_a \left\| \check{P}^n e(t) \right\|_a \big) l_{n-1}(t)\,\mathrm{d}t \\
&\leq \sum_{n=1}^{N} \Big(C_{3,\mu} \left\| \hat{h}_n (A^{n-1}U^{n-1} - A^nU^n) \right\| + C_{1,\mu} \left\| U^{n-1} - U^n \right\|_a \Big) \\
&\qquad \times C_{2,\mu} \int_{t_{n-1}}^{t_n} \|e(t)\|_a\, l_{n-1}(t)\,\mathrm{d}t \\
&\leq \sum_{n=1}^{N} \Big(\tilde{\gamma}_n + \tilde{\theta}_n \Big) \tau_n^{1/2} \Big(\int_{t_{n-1}}^{t_n} \|e\|_a^2 \Big)^{1/2}
\end{aligned} \tag{3.90}$$

by taking $C_\mu := C_{1,\mu}C_{2,\mu}/3$, $C_\mu{}' := C_{3,\mu}C_{2,\mu}/3$ in (3.63) and (3.65) for the last step. We may now conclude exactly like the last step in the proof of Theorem 3.5.2.4, albeit with θ_n replaced by $\tilde{\gamma}_n + \tilde{\theta}_n$. □

3.6 Numerical experimentation: convergence rates

In this section and in §3.7 we study the numerical behaviour of the error indicators and estimators and compare this behaviour with the true error on three model problems. The C code that we used includes the adaptive FEM library ALBERTA [SS05]. The quadrature formal error is made negligible with respect to other error by using overkill quadrature formulas (exact on polynomials of degree 17 and less).

3.6.1 Benchmark problems

Consider three benchmark problems, the solution of which is known. Namely, take $d = 2$, each problem's data f, u_0 is then chosen such that the exact solution to 3.2 is given by:

$$u(\boldsymbol{x}, t) = \sin(\pi t) \exp\left(-10\,|\boldsymbol{x}|^2\right) \tag{3.91}$$

$$u(\boldsymbol{x}, t) = \sin(20\pi t) \exp\left(-10\,|\boldsymbol{x}|^2\right) \tag{3.92}$$

$$u(\boldsymbol{x}, t) = t \sin\left(\frac{2\arctan(x_2/x_1)}{3}\right) |\boldsymbol{x}|^{2/3} \exp\left(\frac{-1}{1-|\boldsymbol{x}|^2}\right). \tag{3.93}$$

The domain Ω for Problems (3.91) and (3.92) is the square $S := (-1,1)\times(-1,1)$. Problem (3.93), whose solution's gradient is singular at the origin, is considered on the L shaped domain $\Omega = S \setminus [0,1] \times [-1,0]$. The benchmark problems (3.91) and (3.92) have been chosen such that they can be compared with previous numerical studies [LM06].

For all Problems (3.91)–(3.93), we take zero initial condition, $u_0 = 0$ to avoid dealing with the initial adaptivity which is a side issue here.

The solution (3.92) has a time dominant discretisation error, while (3.93) was constructed to have a dominant spatial error. It is the product of a linear function in time, a well known solution to Laplace's equation producing the spatial singularity and a mollifier.

Problem (3.91) is used to test asymptotic behaviour of the indicators under uniform space-time refinements further in §3.6.3. Problems (3.93) and (3.92) will be used to test the adaptive strategies in §3.7.

3.6.2 Gradient recovery implementation

We take G^n to be the recovery operator defined by the ZZ local weighted averaging. It is built in the following way: fixing $V \in \mathbb{V}^n$, for each degree of freedom $\boldsymbol{x}$, we define

$$G^n[V](\boldsymbol{x}) := \frac{\sum_{K\in\mathscr{T}^n:\boldsymbol{x}\in K}|K|\ \nabla V|_K\,(\boldsymbol{x})}{\sum_{K\in\mathscr{T}^n:\boldsymbol{x}\in K}|K|}, \tag{3.94}$$

3.6.3 Indicator's numerical asymptotic behaviour

In the following convergence rate tests we discuss the practical realisation of Theorems 3.5.2.4 and 3.5.2.6, to which we refer for notation.

3.6.3.1 Definition (experimental order of convergence)**.** Given two sequences $a(i)$ and $h(i) \searrow 0$, $i = l, \ldots,$, we define experimental order of convergence (EOC) to be the local slope of the $\log a(i)$ vs. $\log h(i)$ curve, i.e.,

$$\mathrm{EOC}(a, h; i) := \frac{\log(a(i+1)/a(i))}{\log(h(i+1)/h(i))}. \tag{3.95}$$

3.6.3.2 Definition (effectivity index)**.** The main tool deciding the quality of an estimator is the effectivity index (EI) which is the ratio of the error and the estimator, i.e.,

$$\mathrm{EI}(t_n) := \eta_n / \|U - u\|_{\mathrm{L}_2(0,t_n;\mathrm{H}^1_0(\Omega))}. \tag{3.96}$$

If $\mathrm{EI}(t_n) \to 1$ as $\sup_{x,n} h_n(x) \to 0$ then we say the estimator is *asymptotically exact*.

We use a uniform timestep and uniform meshes that are fixed with respect to time. Hence for each test we have $\mathbb{V}^n = \mathbb{V}^0 = \mathbb{V}$ and $\tau_n = \tau(h)$ for all $n = 1, \ldots, N$. For each test we fix the polynomial degree p and two parameters k, c and then compute a sequence of solutions with $h = h(i) = 2^{-i/2}$, and $\tau = ch^k$ for a sequence of refinement levels $i = l, \ldots, L$.

Due to the finite element space invariance in time, the coarsening indicator γ_n vanishes and is thus not computed (this indicator will be discussed in §3.7).

The initial value being zero makes the initial error $U(0) - u(0)$ zero. Thus we do not need to calculate this term in the estimator.

For all solutions the boundary values are not exactly zero, but of a negligible value, hence little interpolation error is committed here (nonetheless some care is taken when

dealing with very small errors). Finally, the data approximation error term, β_n, though important for highly oscillatory data, will not be studied here given the regularity of our data.

What we compute on a space-time uniform mesh are the indicators ε_n and θ_n (or $\tilde{\theta}_n$,$\tilde{\gamma}_n$), defined in §3.5.2.2, and the corresponding *cumulative indicators* $(E_n)_{n=1,\dots,N}$ and $(\Theta_n)_{n=1,\dots,N}$ defined by:

$$\begin{aligned} E_m^2 &:= \sum_{n=1}^{m} \left(\varepsilon_n^2 + \varepsilon_{n-1}^2\right) \tau_n/2 \qquad \text{(for space)}\,, \\ \text{and } \Theta_m^2 &:= \sum_{n=1}^{m} \theta_n^2 \tau_n \text{ or } \sum_{n=1}^{m} \left(\tilde{\theta}_n^2 + \tilde{\gamma}_n^2\right) \tau_n \qquad \text{(for time)}\,. \end{aligned} \tag{3.97}$$

From the Theorems 3.5.2.4 and 3.5.2.6, we know that

$$\|U^m - u(t_m)\|^2 \leq E_m^2 + \Theta_m^2 + \sum_{n=1}^{m} \beta_n^2 \tau_n. \tag{3.98}$$

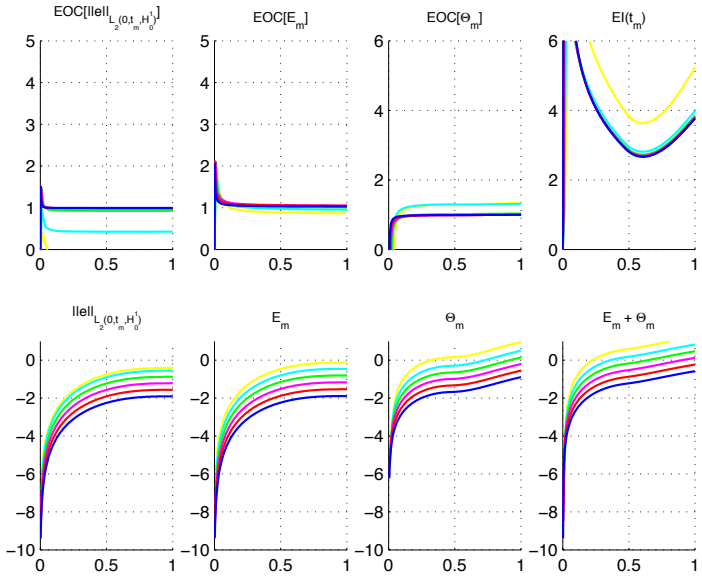
Our results and the comments are reported in the captions of figures.

In Figures 3.1–3.4 we visualise the results and comment on them, for Problem (3.91) for conforming finite elements of polynomial degree $p = 1, \dots, 4$, respectively. Having fixed p, k, c such that $\tau = ch^k$, for each level i, we plot Θ_m and E_m, $\|U - u\|_{\mathrm{L}_2(0,t_m;\mathrm{H}_0^1(\Omega))}$, their experimental order of convergence, EOC, and the effectivity index $\mathrm{EI}(t_m)$ versus (discrete) time $t_m = 0, \dots, T$. Each level i is realised as a curve in each of the Figures 3.1–3.4. These curves are coloured such that they darken (on the greytone scale) as i increases.

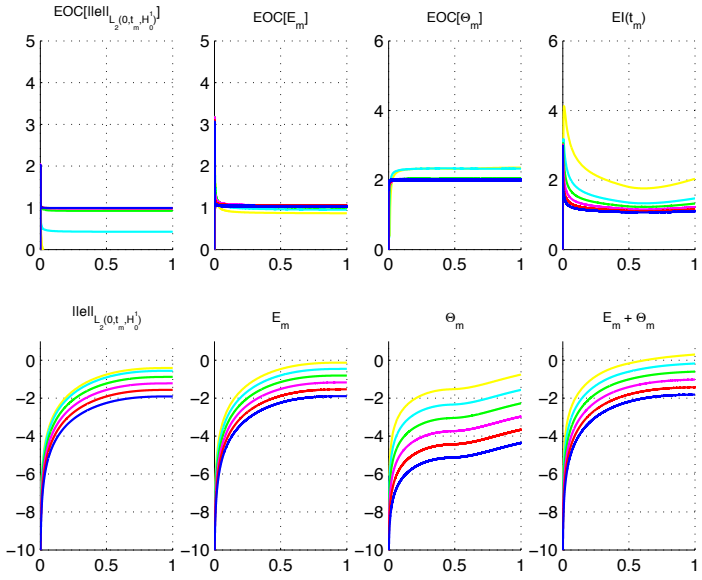
The conclusion is that the estimator is sharp and reliable, but to achieve asymptotic exactness (or close) the time indicator must be made smaller than the space indicator by taking $\tau \ll h^p$. In all these tests we used the first form for Θ_m appearing in (3.97).

In Figure 3.5 we summarise a comparison between the two time indicators θ_n and $\tilde{\theta}_n$, showing that the latter yields a much sharper bound, but with the added cost of having to compute the higher order term $\tilde{\gamma}_n$.

Figure 3.1: Numerical Results for Problem (3.91) with $\mathbb{P}^1$ elements and $h = h(i) = 2^{-i/2}$, $i = 4, \dots, 9$ (details in §3.6.3).

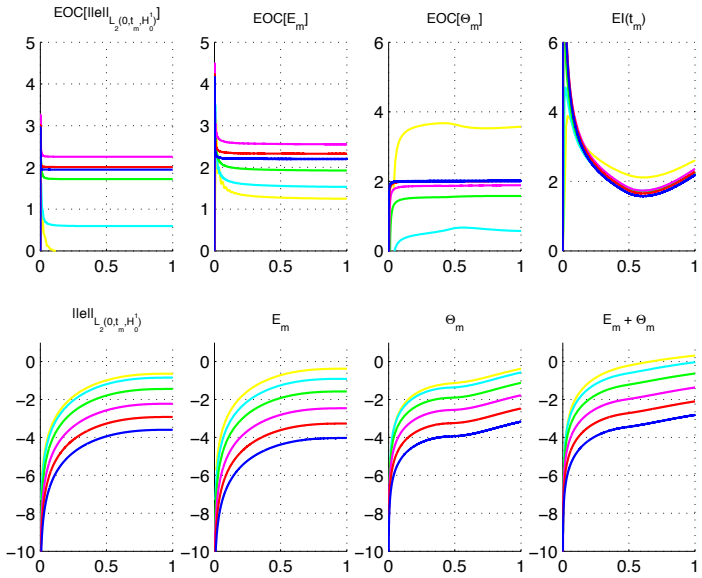


(a) Mesh-size is h and timestep $\tau = 0.1\,h$. On top we plot the EOCs of the single cumulative indicators E and Θ. Below we plot their logs. Both indicators have EOC $\to 1$, but the cumulative time error indicator Θ_m is dominant. The estimator is reliable and sharp, but not asymptotically exact and results in EI $\gg 1$.

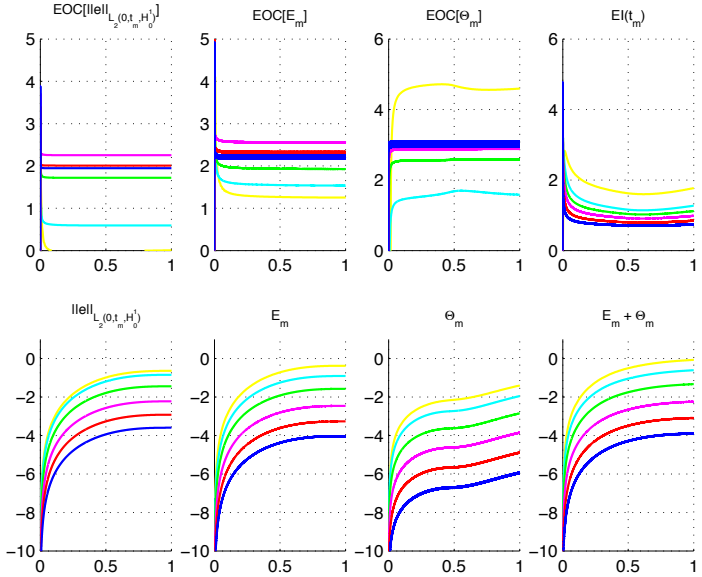


(b) Timestep is $\tau = 0.1\,h^2$. This choice leads to EOC$[\Theta_m] \to 2$ and EOC$[E_m] \approx 1$, i.e., the time indicator Θ_m is of higher order than the spatial indicator E_m which leads the estimator's order. Thus we obtain asymptotic exactness EI $\to 1$, as expected from ZZ estimators for $p = 1$.

Figure 3.2: Numerical Results for (3.91) with $\mathbb{P}^2$ elements and $h = h(i) = 2^{-i/2}$ with $i = 3, \ldots, 8$. We compute the same quantities as in Figure 3.1.

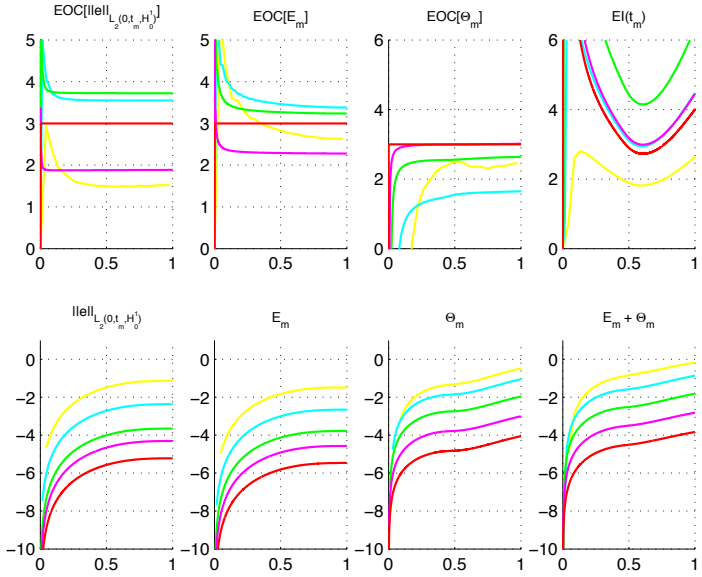


(a) Timestep $\tau = 0.1\,h^2$. The cumulative time error indicator Θ_m is dominant with $\mathrm{EOC}[\Theta_m] \to 2$, but $\mathrm{EI} \gg 1$.

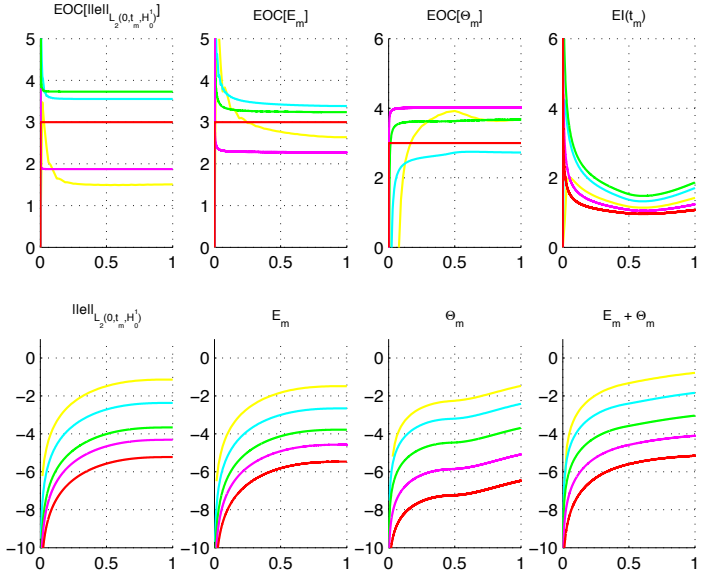


(b) Timestep is $\tau = 0.1\,h^3$. The spatial is dominant (EOC ≈ 2) showing the estimator is sharp and reliable for higher order polynomials as well, and close to asymptotically exact (EI just smaller than 1).

Figure 3.3: Numerical Results for (3.91) with $\mathbb{P}^3$ elements for mesh-sizes $h(i) = 2^{-i/2}$, $i = 2, \ldots, 6$.. We compute the same quantities as in Figures 3.1 and 3.2.

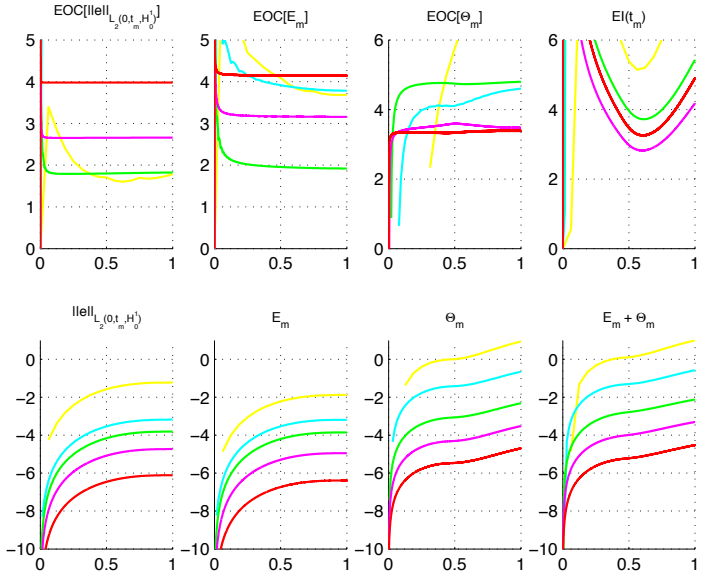


(a) Timestep is $\tau = 0.1\,h^3$. Again, the time indicator is dominant and $\mathrm{EOC}[\Theta_m] \to 3$, but $\mathrm{EI} \gg 1$.

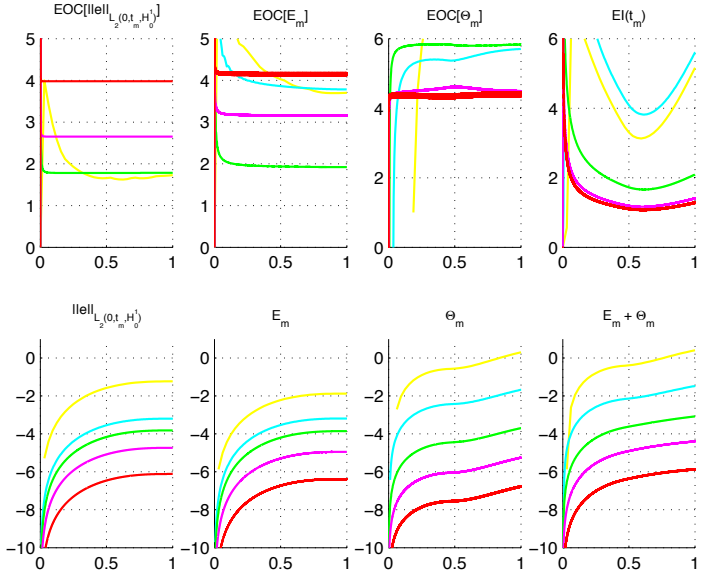


(b) Timestep is $\tau = 0.1\,h^4$. The elliptic error is dominant ($\mathrm{EOC}[E_m] \to 3$) and the estimator is sharp and reliable with very good EI.

Figure 3.4: Results for (3.91) with $\mathbb{P}^4$ elements and $h(i) = 2^{-i/2}$, $i = 2, \ldots, 6$. We compute the same time accumulation quantities as in Figures 3.1–3.3.

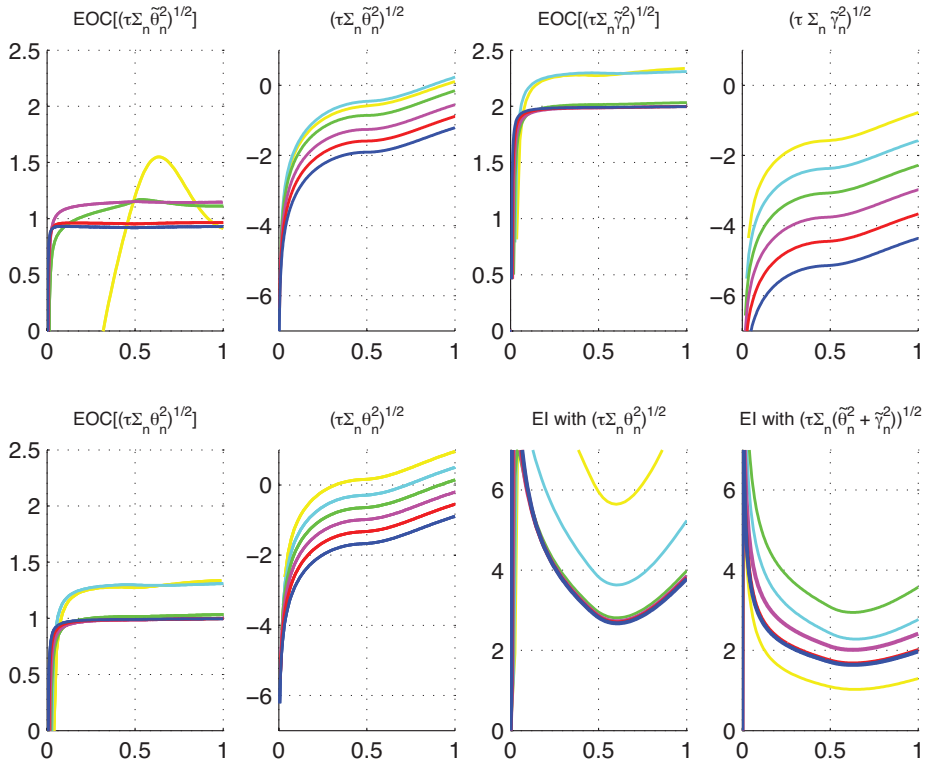


(a) Mesh-size is $\tau = 0.1\, h^4$. Again, the time indicator is dominant with order $\mathrm{EOC}[\Theta_m] \to 4$) and a quite good EI in this case.



(b) Mesh-size is $\tau = 0.1h^5$. The spatial error is dominant and $\mathrm{EOC}[E_m] \to 4$. Effectivity index improves slightly over previous case.

Figure 3.5: For each $m = 1, \dots, N$ we plot values and EOCs of two alternative time indicators $\left(\sum_{n=1}^{m} \tau_n \tilde{\theta}_n^2\right)^{1/2}$ (above) and $\left(\sum_{n=1}^{m} \tau_n \theta_n^2\right)^{1/2}$ (below) and the alternative mesh-change indicator $\sum_{n=1}^{m} \tau_n \tilde{\gamma}_n^2$ (above-right). All quantities are plotted against time. We took a uniform timestep $\tau = 0.1\, h$ and mesh-size $h = 2^{-i}$, $i = 4, \dots, 9$. The numerical results show (1) that the two time indicators are equivalent in order, as expected, and (2) that the term $\sum_{n=1}^{m} \tau_n \tilde{\gamma}_n^2$ is indeed a higher order term and can be safely ignored in most practical schemes. The indicators $\tilde{\theta}_n$ have a better effectivity index.



3.7 Numerical experimentation: adaptive schemes

We present now an adaptive algorithm based on the error indicators defined in §3.5.2.2. As with many adaptive methods for time-dependent problems [Pic98, SS05, CJ04], we perform space and time adaptivity separately. Adaptivity is controlled via the indicators η_n (or $\tilde{\eta}_n$)—see Theorems 3.5.2.4 and 3.5.2.6—which are kept under a given tolerance, tol.

Namely, at each timestep $t_{n-1} \to t_n$, we use adaptive schemes for elliptic problems as to minimise the indicators $\tilde{\varepsilon}_n$ and β_n. There are different strategies to perform the timestep adaptivity, all geared towards minimising θ_n (or $\tilde{\theta}_n$). Finally, the coarsening estimator γ_n is minimised by precomputing it and performing only one coarsening operation at the beginning of each timestep.

Note that it is not in the scope of this section to prove any rigorous result about the adaptive algorithm and, based on heuristic arguments only, we use it for illustration purposes.

3.7.1 Spatial adaptivity via maximum strategy

At each timestep an elliptic problem is solved. For linear elliptic problems, convergence of adaptive schemes is reasonably well understood [MNS02a, BDD04] so we follow the criteria given therein, namely the Maximum Strategy.

The algorithm we used can be pseudocoded as follows.

3.7.2 Space Adapt

Require: $(U^{\mathrm{old}}, \mathbb{V}^{\mathrm{old}}, \mathrm{tol}_\varepsilon, k_{\max}, t, \tau, \xi, \mathrm{tol}_\gamma)$

Ensure: $(U^{\mathrm{new}}, \mathbb{V}^{\mathrm{new}})$ solution of (3.11)

▷ Coarsening step:

$\boldsymbol{\gamma} = (\gamma^K)_{K\in\mathscr{T}} := \mathsf{Coarsening\ Preindicator}(U^{\mathrm{old}}, \mathbb{V}^{\mathrm{old}})$ (cf. [LP10d, §A]).

$\mathscr{T} := \mathsf{Mesh}(\mathbb{V}^{\mathrm{old}})$

find $\mathscr{C} \subset \mathscr{T}$ such that $\sum_{K\in\mathscr{C}} (\gamma^K)^2 \leq \mathrm{tol}_\gamma^2$

$\mathscr{T} := \mathsf{Coarsen}(\mathscr{T}, \mathscr{C})$ using [SS05, §1.1.2–1.1.3]

▷ Refinement loop using Maximum Strategy [SS05]:

$k := 0$

compute ε_n using (3.60)
$\mathscr{R} := \emptyset$ ▷ refinement set
while $\varepsilon_n > \mathrm{tol}_\varepsilon$ and $k \leq k_{\max}$ **do**
  for all $K \in \mathscr{T}^n$ **do**
    if $\varepsilon_{K,n}^2 \geq \xi \max_{L\in\mathscr{T}^n} \varepsilon_{L,n}^2$ **then**
      $\mathscr{R} := \{K\} \cup \mathscr{R}$ ▷ mark K for refinement
    end if
  end for
  $\mathscr{T} := \mathsf{Refine}(\mathscr{T}, \mathscr{R})$ using [SS05, §1.1.1] ▷ update $(U^{\mathrm{old}}, \mathbb{V})$
  set $\Lambda^n U^{n-1} := U^{\mathrm{old}}$, $\tau_n = \tau$, $t_n = t$ and solve for U^n in (3.12)
  $U := U^n$
  compute ε_n using (3.60)
  $k := k + 1$
end while
return $(U, \mathbb{V})$

3.7.3 Coarsening

In time-dependent problems mesh coarsening, which is not to be confused with the coarsening needed in proving optimal complexity for adaptive schemes [BDD04], is used to reduce the DOFs that become redundant in time.

Mesh coarsening is a delicate procedure and should be used sparingly as to avoid needless overhead computing time. In Algorithm 3.7.2, coarsening is performed only once, at the beginning, for each time-step.

The coarsening strategy we propose is based on *predicting the effect of a possible removal of degrees of freedom*. The reason for this is that in ALBERTA (and many other finite element codes) upon coarsening, all DOF-dependent vectors (encoding finite element function coefficients) are "coarsened" via interpolation. This makes it possible to compute the effect of coarsening, and the coarsening estimator γ_n defined in (3.64), *before* mesh-change occurs. The details of this procedure are discussed in § 3.8.

3.7.4 Timestep control

Timestep control can be achieved using two different strategies.

An *implicit timestep control* strategy is ready implemented in ALBERTA [SS05] using Algorithm 3.7.2 upon each timestep.

Here we propose an *explicit timestep control* strategy which we have implemented in ALBERTA. The reason for this is that the implicit strategy, though better in terms of timestep determination, is very time-consuming as it requires the repeated solution of the timestep. In contrast, the explicit strategy has a rougher—nonetheless still satisfactory—control over the timestep, but it is much faster. The conclusion is that the ideal control should be a smart implicit/explicit-switching algorithm.

The explicit strategy can be described as follows.

3.7.5 Explicit Timestep Adapt

Require: $(\tau_0, t_0, T, \mathscr{T}^0, u^0, \mathrm{tol}_\varepsilon, k_{\max}, \xi, \mathrm{tol}_\gamma, \mathrm{tol}_{\theta,\min}, \mathrm{tol}_\theta)$

Ensure: $(\tau_n, \mathbb{V}^n, U^n)_{n=1,\dots,N}$ satisfying (3.11) and possibly $\int_0^T \|U-u\|^2 \le \mathrm{tol}^2$

$(U^0, \mathbb{V}^0) = \mathsf{Initial\ Space\ Adapt}(\mathscr{T}^0, u^0, k_{\max}, \xi, \kappa)$ ▷ data interpolation

$n := 1$

$\tau_n := \tau_{n-1}$

$t_n := t_{n-1} + \tau_n$

while $t_n \le T$ **do**

  $(U^n, \mathbb{V}^n) := \mathsf{Space\ Adapt}(U^{n-1}, \mathbb{V}^{n-1}, \mathrm{tol}_\varepsilon, k_{\max}, \tau_n, t_n, \xi, \mathrm{tol}_\gamma)$

  compute θ_n

  if $\theta_n > \mathrm{tol}_\theta$ **then**

    $\tau_{n+1} := \tau_n/\sqrt{2}$

  else if $\theta_n \le \mathrm{tol}_{\theta,\min}$ **then**

    $\tau_{n+1} := \sqrt{2}\tau_n$

  end if

  $t_{n+1} := t_n + \tau_{n+1}$

  $n := n + 1$

end while

return $(U^n)_{n=1,\dots,N}$,

where the *global tolerance* tol is given by the relation

$$\mathrm{tol}^2 = T\big(\mathrm{tol}_\theta^2 + \mathrm{tol}_\varepsilon^2 + \mathrm{tol}_\gamma^2\big)\,. \tag{3.99}$$

Note that this algorithm does not guarantee reaching a tolerance, unlike more sophisticated ones found in the literature [CJ04, e.g.], but it guarantees termination in reasonable CPU times.

3.7.6 Numerical results

In Tables 3.1–3.3 we compare the implicit timestep control strategy described by algorithm 3.7.5 with a uniform timestep scheme. For the uniform strategy we take a stationary mesh in time and set $\tau = 0.04h^2$. We calculate the error for various numerical simulations using differing values of h using the uniform strategy and set those values as tolerances for the adaptive scheme varying ξ appropriately.

Each column displays results for either the uniform strategy or the adaptive strategy using various thresholds. These columns are further subdivided into two, the first containing $\sum_{n=1}^{N} \dim \mathbb{V}^n$ (i.e., the total number of degrees of freedom from all meshes over time) which we denote DOF and the second containing CPU time (seconds) for all model problems (3.91)–(3.93). An entry of OOM (out of memory) indicates a lack of memory to complete the simulation.

	Uniform		Adaptive					
			$\xi = 0.65$		$\xi = 0.70$		$\xi = 0.75$	
tol	DOF's	CPU	DOF's	CPU	DOF's	CPU	DOF's	CPU
0.573	232,290	3	24,080	4	22,792	5	22,240	4
0.295	3,489,090	49	42,042	8	39,414	8	38,630	6
0.149	54,097,020	598	82,172	15	77,932	15	76,452	16
0.0625	OOM	OOM	206,709	39	195,810	37	191,650	37

Table 3.1: Implicit timestep control with various spatial maximum strategy thresholds for Problem (3.91). The adaptive method clearly saves DOF and CPU time over the uniform method.

	Uniform		Adaptive					
			$\xi = 0.65$		$\xi = 0.7$		$\xi = 0.75$	
tol	DOF's	CPU	DOF's	CPU	DOF's	CPU	DOF's	CPU
0.296	3,489,090	47	12,092	5	11,430	5	11,498	5
0.21	13,940,289	196	17,038	7	16,140	8	16,201	7
0.104	54,097,020	602	106,188	32	100,058	29	22,597	10
0.03125	OOM	OOM	513,694	120	460,637	118	449,568	115

Table 3.2: Implicit timestep control with various spatial maximum strategy thresholds for spatial-error dominant Problem (3.93). Adaptivity saves DOF and CPU.

	Uniform		Adaptive			
			$\xi = 0.7$		$\xi = 0.75$	
tol	DOF's	CPU	DOF's	CPU	DOF's	CPU
1.000	925,809	12	159,070	43	127,610	58
0.569	3,489,090	49	237,960	142	204,376	180
0.295	54,097,020	605	471,733	755	471,542	920
0.149	OOM	OOM	940,618	1410	940,138	1850

Table 3.3: Implicit timestep control with various spatial maximum strategy thresholds for spatial-error dominant Problem (3.92). Adaptivity saves DOF (even better than explicit control) but the CPU time grows very quickly due to overhead.

3.7.6.1 Remark (implicit timestep control on fast oscillating solutions)**.** We take note of the CPU times from the results for Problem (3.92) given in Table 3.3. These show that implicit timestep control is undesirable for fast oscillating functions. This is because the timestep searching becomes computationally inefficient. Numerical simulations for an explicit timestep control strategy are given in Table 3.4. This algorithm is described in detail in the ALBERTA manual [SS05] section 1.5.4. The results show although for a method with low tolerance we use more degrees of freedom we make a substantial gain on the CPU time.

We then fix a value of ξ and compare an adaptive strategy with uniform for a single

	Uniform		Adaptive			
			$\xi = 0.7$		$\xi = 0.75$	
tol	DOF's	CPU	DOF's	CPU	DOF's	CPU
1.000	925,809	12	135,788	5	127,004	4
0.569	3,489,090	49	198,628	7	194,311	8
0.295	54,097,026	605	397,716	15	395,876	16
0.149	OOM	OOM	2,177,666	79	2,079,081	76

Table 3.4: Explicit timestep control with various spatial maximum strategy thresholds for time-error dominant Problem (3.92)

value of tol. This is to illustrate how the number of degrees of freedom of the mesh change over time, and how the implicit timestep control affects the timestep size for all test problems in Figures 3.6.

3.7.7 Incompatible data-singular solution

We close this section by testing the adaptive algorithm on an example with incompatible initial and boundary conditions, which is the type of situation where adaptivity is really needed in practise. Consider problem (3.2) with $\Omega = (0,1) \times (0,1)$, $f = 0$ and $u_0 = 1$. The initial conditions are thus incompatible with the homogeneous Dirichlet boundary conditions valid for all positive times. The exact solution u, though singular at all points of $\partial\Omega \times \{0\}$, can be readily evaluated "by hand" and may be represented in terms of Fourier series of the Laplacian's eigenvalues. Namely, we have

$$u(\boldsymbol{x}, t) = \sum_{m,n=1}^{\infty} C_{m,n} \exp(-\left(m^2 + n^2\right) \pi^2 t) \sin(m\pi x_1) \sin(n\pi x_2), \text{ for } t > 0, \tag{3.100}$$

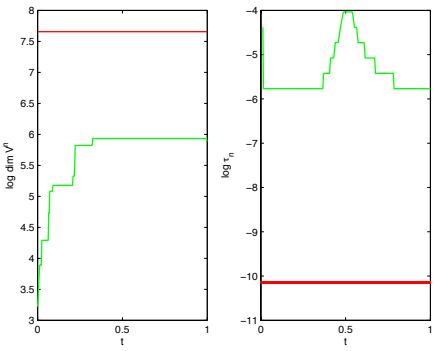
where the constant $C_{m,n}$ is given by

$$C_{m,n} = \frac{4}{nm\pi^2}(1 - \cos(m\pi) - \cos(n\pi) + \cos(n\pi)\cos(m\pi)). \tag{3.101}$$

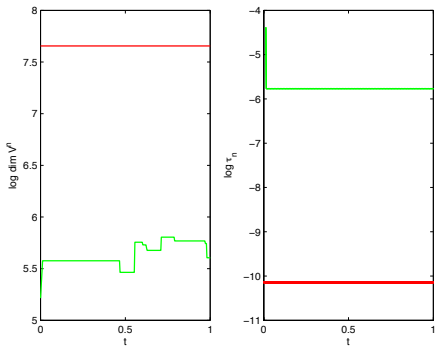
Since the solution (3.100) is an infinite Fourier series it cannot be computed exactly, but its rapid decay allows for an early truncation with machine-epsilon precision.

In order to generate a reference tolerance, which is common for the uniform and the adaptive scheme we couple $h = 0.05\tau$ and run the uniform refinement code. We use then

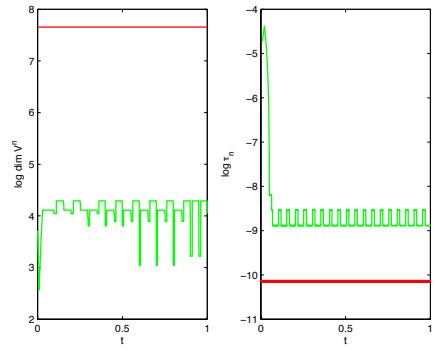
Figure 3.6: Adaptive (green) against uniform (red) degrees of freedom and timestep sizes. In each pair of graphs we plot the (log of) the DOF against time on the left, and the timestep against time on the right.



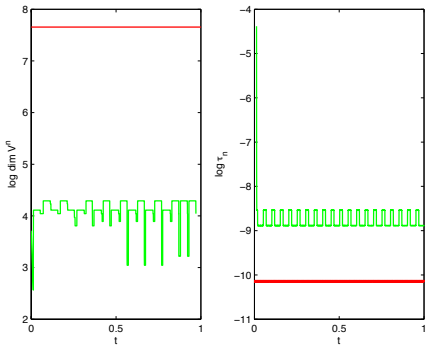
(a) Implicit timestep control for Problem (3.91). The explicit timestep control yields the same results (but is much more CPU efficient), thus it is not shown.



(b) Implicit timestep control for Problem (3.92), where the spatial error dominates. The explicit timestep control yields the same meshes and timesteps, thus not shown.

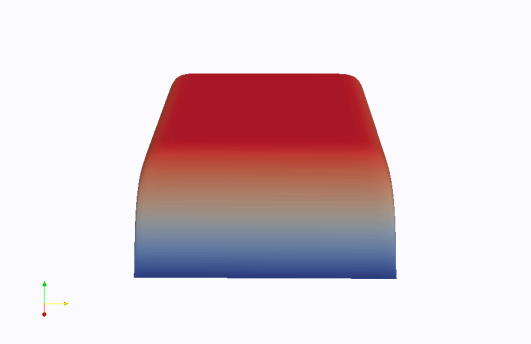


(c) Explicit timestep control for Problem (3.93), where the time discretisation error dominates. Interesting when compared with Figure 3.7(d).

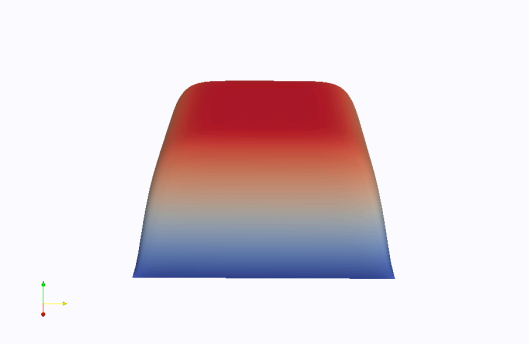


(d) Implicit timestep control for Problem (3.93). Comparing with Figure 3.7(c) shows that the implicit timestep control yields more efficient timestep and meshes, but at a much higher CPU cost (cf. Tables 3.3 and 3.4).

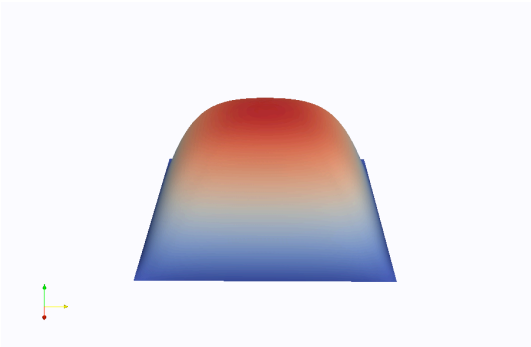
Figure 3.7: The adaptive scheme for (3.100) using implicit timestep control.



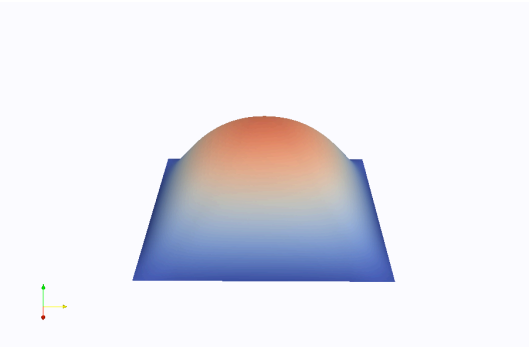
(a) Solution at time $t_n = 0.007544$ with $\dim(\mathbb{V}^n) = 894,677$



(b) Solution at time $t_n = 0.033302$ with $\dim(\mathbb{V}^n) = 98,773$



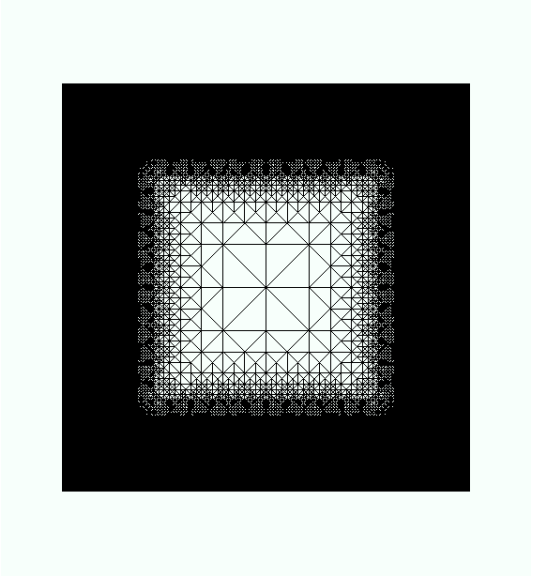
(c) Solution at time $t_n = 0.127492$ with $\dim(\mathbb{V}^n) = 18,613$



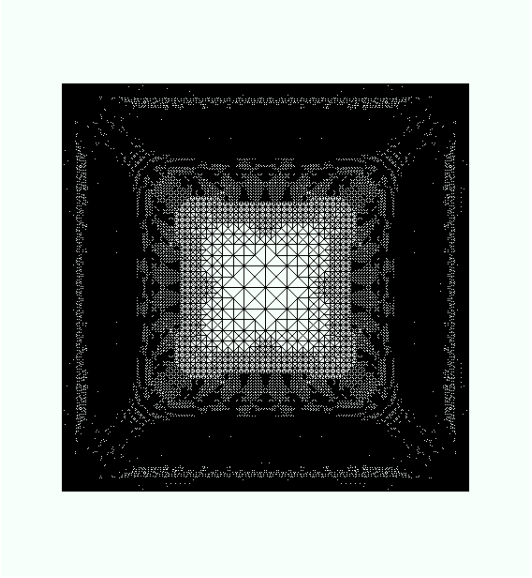
(d) Solution at time $t_n = 0.393893$ with $\dim(\mathbb{V}^n) = 3,525$

the error computed as a tolerance for the adaptive scheme, results of this are shown in Figure 3.7. In Figure 3.8 we visualise the adapted FE mesh for Problem (3.100) at various times.

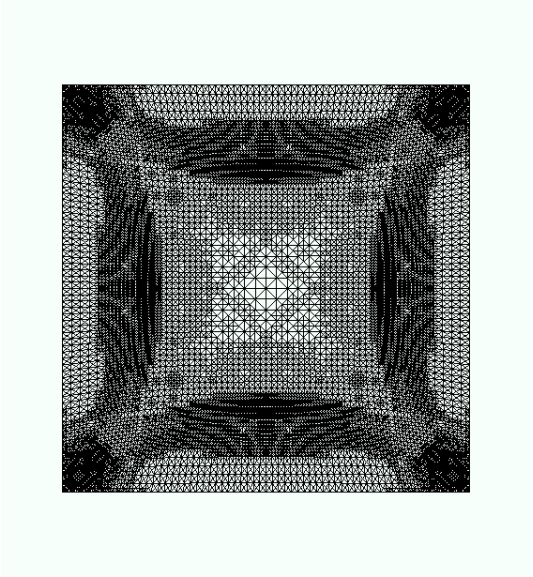
Figure 3.8: The adaptive scheme for (3.100) using implicit timestep control.



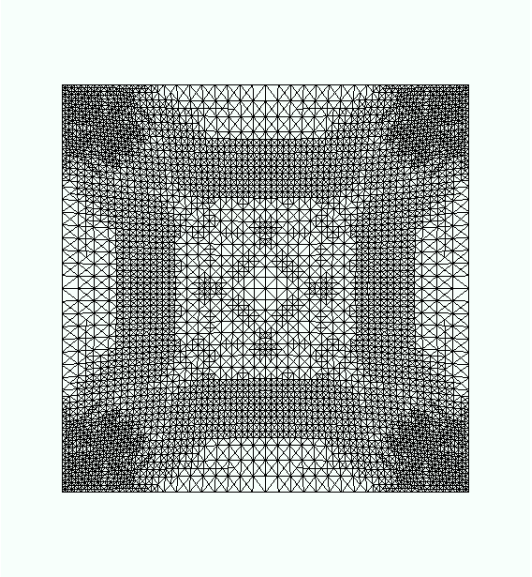
(a) Mesh at time $t_n = 0.007544$ with $\dim(\mathbb{V}^n) = 894,677$



(b) Mesh at time $t_n = 0.033302$ with $\dim(\mathbb{V}^n) = 98,773$



(c) Mesh at time $t_n = 0.127492$ with $\dim(\mathbb{V}^n) = 18,613$



(d) Mesh at time $t_n = 0.393893$ with $\dim(\mathbb{V}^n) = 3,525$

3.8 Building a coarsening estimator into ALBERTA

We describe next a practical implementation of the *coarsening error preindicator* (we use this term to emphasise the fact that this indicator can be computed apriori, as opposed to the other indicators involved in the adaptive strategy). Since we used ALBERTA for our computations, this section relies substantially on the principles described in the manual [SS05]. We briefly describe these principles in the next paragraph, in order to expose the main idea behind the coarsening preindicator.

3.8.1 Refinement, coarsening and interpolation in ALBERTA

Mathematically, a simplicial mesh (or partition, or triangulation) is a set of disjoint open simplexes, the union of the closure of which is $\overline{\Omega}$. A mesh into a new mesh is refined by *bisecting* a subset of its simplexes, following a special procedure which ensures mesh conformity (e.g., no hanging nodes) and does not deteriorate shape-regularity (on fully fitted polygonal domains). A mesh is thus represented as a binary tree, where each node represents a simplex. The children of each simplex are thus the 2 subsimplexes obtained by bisection. Hence, from a coding view-point, refinement means growing the binary tree.

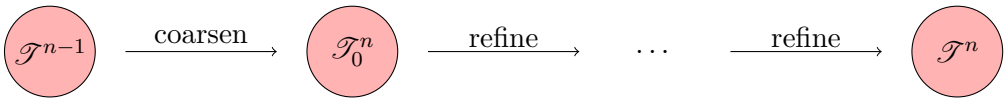
The inverse of refinement is coarsening. Thus coarsening a mesh in ALBERTA consists in removing pairs of sibling simplexes (both marked for coarsening) and produces the new—coarsened—mesh where the pairs of siblings are replaced by their parent.

The coarsening preindicator is a real number defined on each simplex, of the triangulation to be coarsened. This estimator can in fact be *precomputed* with respect to coarsening. This is in contrast with usual aposteriori error estimators which can be postcomputed only (i.e., after the discrete solution has been computed). To clarify this point, let us focus on the particular situation of interest. Let U^{n-1} be the solution from the previous timestep; $U^{n-1} \in \mathbb{V}^{n-1}$, the finite element space with respect to mesh $\mathscr{T}^{n-1}$. The error due to coarsening appears in the term

$$U^{n-1} - \Lambda^n U^{n-1}. \tag{3.102}$$

This term is nonzero only when simplexes are coarsened.

Furthermore, we assume that the new mesh $\mathscr{T}^n$ is a refinement of $\mathscr{T}_0^n$, which is a coarsening of the old mesh $\mathscr{T}^{n-1}$:



If Λ_0^n is the Lagrange interpolant onto the finite element space $\mathbb{V}_0^n$, relative to the new coarse mesh $\mathscr{T}_0^n$, it is not very difficult to predict $\Lambda_0^n U^{n-1}$ without actually computing it. Therefore this term can be predicted from (a) the simplexes of $\mathscr{T}^{n-1}$ marked for coarsening which leads to $\mathscr{T}_0^n$ and (b) the values of U^{n-1}.

Note that since $\mathscr{T}_0^n$ is subsequently *refined but not coarsened* to produce $\mathscr{T}^n$, as depicted in (3.8.1), then the additional coarsening error will be zero. Namely, if Λ^n denotes the Lagrange interpolant onto $\mathbb{V}^n$, the finite element space over $\mathscr{T}^n$, which is a refinement of $\mathscr{T}_0^n$, then $\Lambda^n U^{n-1} = \Lambda_0^n U^{n-1}$, and thus

$$U^{n-1} - \Lambda^n U^{n-1} = U^{n-1} - \Lambda_0^n U^{n-1}. \tag{3.103}$$

The coarsening strategy therefore consists in choosing a subset of simplexes of $\mathscr{T}^{n-1}$ which minimises term $\left\| U^{n-1} - \Lambda_0^n U^{n-1} \right\|$ *before* producing the new coarse mesh $\mathscr{T}_0^n$.

The rest of this section describes how $U^{n-1} - \Lambda_0^n U^{n-1}$ can be precomputed.

3.8.2 Notation

Let K be an element of the new coarse mesh $\mathscr{T}_0^n$ resulting from the coarsening of its two children which we denote by $K^{\pm}$. (Note that K^+ and K^- correspond to `child[0]` and `child[1]` of K in the ALBERTA manual [SS05].) Define the *fine space*

$$\mathbb{Y} := \left\{ \Phi|_K : \ \Phi \in \mathbb{V}^{n-1} \right\}. \tag{3.104}$$

Likewise define the *coarse space* $\mathbb{X}$ to be the local finite element space, i.e.,

$$\mathbb{X} := \left\{ \Phi|_K : \ \Phi \in \mathbb{V}_0^n \right\}; \tag{3.105}$$

simply put we just have $\mathbb{X} = \mathbb{P}^p$. We introduce also the *fine spaces* $\mathbb{Y}^{\pm}$, defined like $\mathbb{Y}$, but restricting functions over $K^{\pm}$, respectively (so functions in $\mathbb{Y}^{\pm}$ are in fact the same as $\mathbb{X} = \mathbb{P}^p$, albeit with different domains).

Denote by $\{\boldsymbol{x}_0, \ldots, \boldsymbol{x}_L\}$ and $\left\{\boldsymbol{x}_0^{\pm}, \ldots, \boldsymbol{x}_L^{\pm}\right\}$ the set of Lagrange degrees of freedom on the simplex K and its children $K^{\pm}$, respectively. We indicate with $\left\{\pi^0, \ldots, \pi^L\right\}$ and

$\{\pi_\pm^0, \dots, \pi_\pm^L\}$ the corresponding Lagrange polynomial bases of $\mathbb{X}$ and $\mathbb{Y}^\pm$, respectively, whereby

$$\pi^i(\boldsymbol{x}_j) = \pi_\pm^i(\boldsymbol{x}_j^\pm) = \delta_j^i. \tag{3.106}$$

For short we will write these bases as column vectors $\boldsymbol{\pi} = \left(\pi^0, \dots, \pi^L\right)^\intercal$, etc. We also define the (local) *coarse-on-fine matrices* by

$$\mathbf{A}^\pm := \left(\boldsymbol{\pi}(\boldsymbol{x}_0^\pm) \ \dots \ \boldsymbol{\pi}(\boldsymbol{x}_L^\pm)\right) = \left(\pi^i(\boldsymbol{x}_j^\pm)\right)_{i,j=0,\dots,L}. \tag{3.107}$$

These matrices are closely related to ALBERTA's *refine-interpolation* matrix [SS05, matrix A (1.5) in §1.4.4].

3.8.2.1 Proposition (coarse-on-fine matrix properties)**.** *The matrices* $\mathbf{A}^+$ *and* $\mathbf{A}^-$ *are independent of* K, K^+, K^- *and*

$$\boldsymbol{\pi}|_{K^\pm} = \mathbf{A}^\pm \boldsymbol{\pi}_\pm. \tag{3.108}$$

Proof Fix $i = 0, \dots, L$. Because π^i is a polynomial and $\{\pi_+^0, \dots, \pi_+^L\}$ is a polynomial basis, it follows that

$$\pi^i = \sum_{j=0}^{L} a_j^i \pi_+^j, \tag{3.109}$$

for some vector $\left(a_0^i, \dots, a_L^i\right)$. Applying π^i to $\boldsymbol{x}_j^+$, and recalling (3.106), we obtain

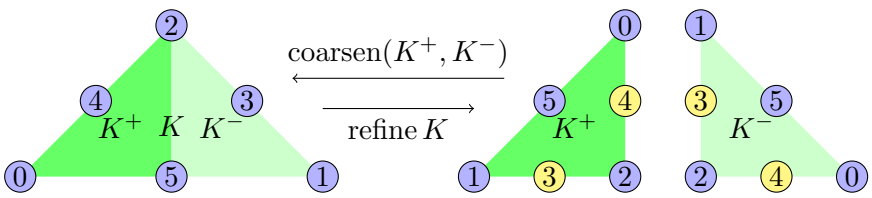
$$a_j^i = \pi^i(\boldsymbol{x}_j^+), \tag{3.110}$$

and hence

$$\pi^i = \left[\mathbf{A}^+ \boldsymbol{\pi}_+\right]^i. \tag{3.111}$$

□

3.8.2.2 Example (quadratic elements in 2 dimensions)**.** To make the discussion more accessible, we will illustrate it as we go with the concrete situation where $p = 2$ (quadratic elements) and $d = 2$. Following the ALBERTA conventions the relation between the coarse and fine triangles is given by the following diagram.



In this case, the coarse-on-fine matrices are computed as follows (omitting the zeroes for clarity):

$$\mathbf{A}^+ = \begin{bmatrix} 1 & & 3/8 & -1/8 \\ & & -1/8 & -1/8 \\ 1 & & & \\ & & & 1/2 \\ & & & 1/2 \quad 1 \\ & 1 & 3/4 & 1/4 \end{bmatrix}, \quad \mathbf{A}^- = \begin{bmatrix} & & -1/8 & -1/8 \\ 1 & & -1/8 & 3/8 \\ & 1 & & \\ & & 1/2 & \quad 1 \\ & & 1/2 & \\ & 1 & 1/4 & 3/4 \end{bmatrix} \tag{3.112}$$

3.8.3 Degrees of freedom and global–local relations

Denote by U the generic finite element function in the old space $\mathbb{V}^{n-1}$ and let $V := \Lambda_0^n U$. Then we have

$$U = \mathbf{u}^\intercal \boldsymbol{\Psi} \text{ and } V = \mathbf{v}^\intercal \boldsymbol{\Phi}, \tag{3.113}$$

where $\boldsymbol{\Psi} = \left(\Psi^0, \ldots, \Psi^N\right)^\intercal$ and $\boldsymbol{\Phi} = \left(\Phi^0, \ldots, \Phi^M\right)^\intercal$, are the columns of nodal Lagrange piecewise polynomial bases of $\mathbb{V}^{n-1}$ and $\mathbb{V}_0^n$, respectively, and $\mathbf{u}$ and $\mathbf{v}$ are the corresponding vectors of DOF values.

There are $L+1$ degrees of freedom (DOF) per simplex, e.g., $L = 5$ for $p = 2 = d$. The simplex K in $\mathscr{T}_0^n$ comes with a *local-to-global* index relation $g = g_K^{\mathscr{T}_0^n} : [0:L] \to [0:M]$ whereby

$$\Phi^{g(i)}\Big|_K = \pi^i \quad \forall\, j = 0, \ldots, L. \tag{3.114}$$

It follows that the finite element function V is locally represented on K by

$$Y := V|_K = \sum_{i=0}^{L} v_{g(i)} \pi^i =: \mathbf{y}^\intercal \boldsymbol{\pi}. \tag{3.115}$$

Similarly we have $g^\pm = g_{K^\pm}^{\mathscr{T}^{n-1}} : [0:L] \to [0:N]$ such that

$$Y^\pm := U|_{K^\pm} = \sum_{j=0}^{L} u_{g^\pm(j)} \pi_\pm^j =: \mathbf{y}^{\pm\intercal} \boldsymbol{\pi}_\pm. \tag{3.116}$$

The relation between the DOF coefficients $\mathbf{u}$ and $\mathbf{v}$ will be described next.

3.8.4 Local fine–coarse DOF relations

Some degrees of freedom—that is those depicted in yellow—are removed during coarsening. The others, which are kept, have their local index change. This information is fully encoded in the *fine-to-coarse* index maps $c^\pm : D^\pm \to C^\pm$ where

$$D^\pm := \left\{ j = 0, \dots, L : \ \boldsymbol{x}_j^\pm \in \{\boldsymbol{x}_0, \dots, \boldsymbol{x}_L\} \right\}. \tag{3.117}$$

and

$$C^\pm := c^\pm(D^\pm) \subseteq [0:L]\,. \tag{3.118}$$

A basic property of the fine-to-coarse maps is that

$$C^+ \cup C^- = [0:L]\,, \tag{3.119}$$

but C^+ and C^- need not be disjoint (in fact, for conforming methods these are never disjoint). The fine-to-coarse maps $c^\pm$ are injective and we denote their inverses, the *coarse-to-fine* maps, by $d^\pm : C^\pm \to D^\pm$.

In the example above, $p = 2 = d$, the fine-to-coarse maps $c^\pm : D^\pm \to [0:5]$, satisfy $D^+ = D^- = \{0,1,2,5\}$ (though D^+ and D^- do not generally coincide, as seen for $p = 3$, $d = 2$, e.g.) and evaluated by the schedule

$$\begin{aligned} j &= \ 0 \ \ 1 \ \ 2 \ \ 3 \ \ 4 \ \ 5, \\ c^+(j) &= \ 2 \ \ 0 \ \ 5 \ \ - \ \ - \ \ 4, \\ c^-(j) &= \ 1 \ \ 2 \ \ 5 \ \ - \ \ - \ \ 3. \end{aligned} \tag{3.120}$$

It follows that $C^+ = \{0,2,4,5\}$ and $C^- = \{1,2,3,5\}$ and

$$\begin{aligned} i &= \ 0 \ \ 1 \ \ 2 \ \ 3 \ \ 4 \ \ 5, \\ d^+(i) &= \ 1 \ \ - \ \ 0 \ \ - \ \ 5 \ \ 2, \\ d^-(i) &= \ - \ \ 0 \ \ 1 \ \ 5 \ \ - \ \ 2. \end{aligned} \tag{3.121}$$

3.8.4.1 Remark (redundancy of the coarse-to-fine maps)**.** The coarse-to-fine maps $c^\pm$ and their inverses $d^\pm$ are partially redundant with $\mathbf{A}^\pm$. Namely, if $j \in D^\pm$, then $j = d^\pm(i)$ and $i = c^\pm(j)$, for some $i = 0, \dots, L$. By definition of $c^\pm$ it follows that $\boldsymbol{x}_j^\pm = \boldsymbol{x}_i$. Therefore

$$[\mathbf{A}^\pm]_j^k = \pi^k(\boldsymbol{x}_j^\pm) = \pi^k(\boldsymbol{x}_i) = \delta_i^k. \tag{3.122}$$

We have thus proved the following result that will be used to compress $\mathbf{A}^\pm$ in the sequel.

3.8.4.2 Proposition (redundant coarse-on-fine columns)**.** *If $j \in D^{\pm}$, then $\mathbf{A}^{\pm}$'s j-th column is described by*

$$[\mathbf{A}^{\pm}]_j^k = \delta^k_{c^{\pm}(j)}. \tag{3.123}$$

3.8.5 Precomputing the coarsening error

The coarsening error is the difference between U, to which we have access via $\mathbf{u}$, and its interpolation on the locally coarser mesh V, to which we have no direct access. Working locally at the coarsening-marked element K^+ (and similarly for K^-), all we need is to compute $V|_{K^+}$ and subtract it from $U|_{K^+}$.

Recalling that in ALBERTA $V = \Lambda_0^n U$ is built by simply "dropping" the coefficients of the DOF removed by coarsening we have

$$\mathbf{y}^{\intercal}\boldsymbol{\pi} = Y = V|_K = \sum_{i\in C^+} u_{g_+(d^+(i))}\pi^i + \sum_{i\in C^-\setminus C^+} u_{g_-(d^-(i))}\pi^i, \tag{3.124}$$

that is, for $j = 0, \dots, L$, we set

$$v_{g(i)} := y_i := \begin{cases} u_{g_+(d^+(i))} = y^+_{d_+(i)} & \text{if } i \in C^+ \\ u_{g_-(d^-(i))} = y^-_{d_-(i)} & \text{otherwise .} \end{cases} \tag{3.125}$$

(Note that the vector $\mathbf{y}$ is the same for the two siblings $K^{\pm}$ and needs to be calculated only once.) Following the example with $p = 2 = d$, we see that

$$\begin{aligned} \mathbf{y} &= \left(y_1^+, y_0^-, y_0^+, y_5^-, y_5^+, y_2^+\right)^{\intercal} \\ &= \left(y_1^+, y_0^-, y_1^-, y_5^-, y_5^+, y_2^-\right)^{\intercal}. \end{aligned} \tag{3.126}$$

To conclude we rewrite the coarse basis, $\boldsymbol{\pi}$, in terms of the fine one, $\boldsymbol{\pi}_+$, using Proposition 3.8.2.1 as follows:

$$V|_{K^+} = Y|_{K^+} = \mathbf{y}^{\intercal}\,\boldsymbol{\pi}|_{K^+} = \mathbf{y}^{\intercal}\mathbf{A}^+\boldsymbol{\pi}_+. \tag{3.127}$$

Thus the coarsening error on K^+ is calculated as

$$[U - V]|_{K^+} = \mathbf{y}^{+\intercal}\boldsymbol{\pi}_+ - \mathbf{y}^{\intercal}\mathbf{A}^+\boldsymbol{\pi}_+ = \boldsymbol{\pi}_+{}^{\intercal}\left(\mathbf{y}^+ - \mathbf{A}^{+\intercal}\mathbf{y}\right) = \sum_{j=0}^{L}\left(y_j^+ - \mathbf{y}^{\intercal}[\mathbf{A}^+]_j\right)\pi_j^+. \tag{3.128}$$

Recalling Proposition 3.8.4.2, if $j \in D^+$ we have

$$\mathbf{y}^{\intercal}[\mathbf{A}^+]_j = \sum_{i=0}^{L} y_k\delta^k_{c^+(j)} = y_{c^+(j)} = y_j^+, \tag{3.129}$$

and thus the coefficient for π_j^+ is 0, and it needs not be calculated. Proceeding similarly on K^- we may summarise the findings as follows.

3.8.5.1 Theorem (coarsening error calculation)**.** *Let $U \in \mathbb{V}^{n-1}$ with the notation of §3.8.3, to calculate the coarsening error that would result from coarsening the elements $K^+, K^- \in \mathscr{T}^{n-1}$ into $K \in \mathscr{T}^n$*

(1) calculate $\mathbf{y}$ *following (3.125) using the coarse-to-fine map d^+ defined in §3.8.4,*

(2) obtain the error using

$$\begin{aligned} [U - \Lambda_0^n U]|_{K^+} &= \sum_{j \in [0:L] \setminus D^+} \left(y_j^+ - \mathbf{y}^\intercal [\mathbf{A}^+]_j \right) \pi_+^j, \\ [U - \Lambda_0^n U]|_{K^-} &= \sum_{j \in [0:L] \setminus D^-} \left(y_j^- - \mathbf{y}^\intercal [\mathbf{A}^-]_j \right) \pi_-^j. \end{aligned} \tag{3.130}$$

3.8.5.2 Remark. Note that the j-th coefficient of the coarsening error's local DOF vector is zero when $j \in D^\pm$, respectively. So the calculation needs to be carried out only for those $j \notin D^\pm$.

Also, the coefficients for the DOF that are common to K^+ and K^- must be equal, so they can be in fact computed once.

For example in the case of quadratic elements in $d = 2$ we have

$$\begin{aligned} Y^+ - Y|_{K^+} &= \pi_+^3 \left(y_3^+ - \frac{3}{8} y_1^+ + \frac{1}{8} y_0^- - \frac{3}{4} y_2^+ \right) \\ &\quad + \pi_+^4 \left(y_4^+ + \frac{1}{8} y_1^+ + \frac{1}{8} y_0^- - \frac{1}{4} y_2^+ - \frac{1}{2} y_5^+ - \frac{1}{2} y_5^- \right), \\ Y^- - Y|_{K^-} &= \pi_-^3 \left(y_3^- + \frac{1}{8} y_1^+ + \frac{1}{8} y_0^- - \frac{1}{2} y_5^- - \frac{1}{2} y_5^+ - \frac{1}{4} y_2^+ \right) \\ &\quad + \pi_-^4 \left(y_4^- + \frac{1}{8} y_1^+ - \frac{3}{8} y_0^- - \frac{3}{4} y_2^+ \right). \end{aligned} \tag{3.131}$$

3.8.6 Coarsening error algorithm

As seen in §3.8.5, the information needed for the coarsening error computation for Lagrange finite elements of degree p in dimension d, is contained in the coarse-on-fine matrices $\mathbf{A}^\pm$ defined by (3.107) and the fine-to-coarse maps, $d^\pm$, and their domains $C^\pm$ defined in §3.8.4. This information is independent of the particular pair of simplex siblings $K^\pm$

and their parent K and can be included in the code via given index permutations and efficient matrix-vector multiplication.

With this information at hand and the notation previously introduced in this section, we formulate an ALBERTA-implementable algorithm to precompute the coarsening error on all simplexes.

Coarsening Preindicator

Require: $(U = \mathbf{u}^\intercal \mathbf{\Phi}, \mathbb{V}, \mathscr{T})$

Ensure: $\boldsymbol{\gamma} = (\gamma_K : K \in \mathscr{T})$

for all $K \in \mathscr{T}$ **do**

if [2]childorder$(K) = 0$ **then**

$D := D^+$, $D' := D^-$, $c := c^+$, $c' := c^-$, $\mathbf{A} := \mathbf{A}^+$

else

$D := D^-$, $D' := D^+$, $c := c^-$, $c' := c^+$, $\mathbf{A} := \mathbf{A}^-$

end if

$K' :=$ sibling K

initialise two local DOF vectors $\mathbf{y}$ and $\mathbf{r}$

for all $j \in D$ **do**

$y_{c(j)} = u_{g_K(j)}$

end for

for all $j \in D'$ **do**

$y_{c'(j)} = u_{g_{K'}(j)}$

end for

for all $j \notin D \cup D'$ **do**

$r_j = u_{g_K(j)} - \mathbf{y}^\intercal [\mathbf{A}]_j$

end for

$\gamma_K = 0$

[2]The element information in ALBERTA is quite local and to determine whether an element is left or right child is not trivial. In ALBERTA 1.2 this can be done utilising `EL->index` which provides a global indexing of elements. Testing the `EL_INFO->parent->child[0]->index` against `EL->index` gives the correct child order of K. In ALBERTA 2.0 `EL->index` is unavailable so we check the global index of DOF for both parent and children.

for all $i \notin D \cup D'$ **do**
for all $j \notin D \cup D'$ **do**
$\gamma_K = \gamma_K + r_i r_j \langle \Phi_i, \Phi_j \rangle_K$
end for
end for
end for

3.8.7 Coarsening preindicator matrices

To close, we provide here the information needed to implement Algorithm 3.8.6 for Lagrange piecewise $\mathbb{P}^p$ finite elements in dimension $d = 2$. (For dimension 3 the situation is complicated by the "types" of tetrahedrons, whereby the matrices $A^{\pm}$ and the maps $c^{\pm}$ may depend on the type and is not covered in this section.)

3.8.8 $\mathbb{P}^1$ elements

The coarse-on-fine matrices (omitting 0 entries for clarity) are given by

$$\mathbf{A}^+ = \begin{bmatrix} & 1 & 1/2 \\ & & 1/2 \\ 1 & & \end{bmatrix}, \mathbf{A}^- = \begin{bmatrix} & & 1/2 \\ 1 & & 1/2 \\ & 1 & \end{bmatrix}, \tag{3.132}$$

the fine-to-coarse maps and the coarse-to-fine maps are respectively given by

$$\begin{aligned} i &= 0 \quad 1 \quad 2, & & & i &= 0 \quad 1 \quad 2, \\ c^+(i) &= 2 \quad 0 \quad -, & \text{and} & & d^+(i) &= 1 \quad - \quad 0, \\ c^-(i) &= 1 \quad 2 \quad -, & & & d^-(i) &= - \quad 0 \quad 1. \end{aligned} \tag{3.133}$$

3.8.9 $\mathbb{P}^2$ elements

See the worked example in §3.8.

3.8.10 $\mathbb{P}^3$ elements

The coarse-on-fine matrices are given by

$$\mathbf{A}^+ = \begin{bmatrix}
1 & -1/16 & 5/16 & & 1/16 & & -1/16 \\
 & -1/16 & 1/16 & & 1/16 & & 1/16 \\
1 & & & & & & \\
 & & & & -1/4 & & -1/8 \\
 & & & & 1/2 & & \\
 & & & & 1/2 & 1 & \\
 & & & & -1/4 & & 1 & 3/8 \\
 & 9/16 & 15/16 & 1 & -1/16 & & & 3/16 \\
 & 9/16 & -5/16 & & -1/16 & & & -3/16 \\
 & & & 1 & 1/2 & & & 3/4
\end{bmatrix} \tag{3.134}$$

and

$$\mathbf{A}^- = \begin{bmatrix}
 & -1/16 & 1/16 & & 1/16 & & & -1/16 \\
1 & -1/16 & 1/16 & & 5/16 & & & \\
 1 & & & & & & & 1/16 \\
 & & -1/4 & & & & 1 & \\
 & & 1/2 & & & & & -1/8 \\
 & & 1/2 & & & & & -3/16 \\
 & & -1/4 & & & & & 3/16 \\
 & 9/16 & -1/16 & & -5/16 & & & 3/8 \\
 & 9/16 & -1/16 & & 15/16 & 1 & & \\
 & & 1/2 & 1 & & & & 3/4
\end{bmatrix} \tag{3.135}$$

the fine-to-coarse maps

$$\begin{aligned}
i &= 0\ 1\ 2\ 3\ 4\ 5\ 6\ 7\ 8\ 9, \\
c^+(i) &= 2\ 0\ -\ -\ 7\ 9\ -\ 5\ 6\ -, \\
c^-(i) &= 1\ 2\ -\ -\ 9\ 8\ -\ 3\ 4\ -.
\end{aligned} \tag{3.136}$$

and the coarse-to-fine maps

$$\begin{aligned} i &= 0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 6 \quad 7 \quad 8 \quad 9, \\ d^+(i) &= 1 \quad - \quad 0 \quad - \quad - \quad 7 \quad 8 \quad 4 \quad - \quad 5, \\ d^-(i) &= - \quad 0 \quad 1 \quad 7 \quad 8 \quad - \quad - \quad - \quad 5 \quad 4. \end{aligned} \tag{3.137}$$

3.8.11 $\mathbb{P}^4$ elements

The coarse-on-fine matrices are given by

$$
\mathbf{A}^{+} =
\begin{bmatrix}
1 & 35/128 & -5/128 & 3/128 & -5/128 & & 3/128 & -5/128 & \\
 & -5/128 & 3/128 & 3/128 & -5/128 & & -5/128 & -5/128 & \\
 & & & & & & & & \\
1 & & & & & & & & \\
 & & & -1/16 & 3/16 & & 1/8 & 1/16 & \\
 & & & & -3/8 & & -1/8 & & \\
 & & & & 1/2 & & & & \\
 & & & & 1/2 & 1 & & & \\
 & & & & -3/8 & 1 & 3/8 & & \\
 & & & -1/16 & 3/16 & & 1 & -1/8 & 5/16 \\
 & & & & 1/32 & & & -1/32 & 5/32 \\
 & 35/32 & 1 & 15/32 & -3/32 & & & 3/64 & 15/64 \\
1 & -35/64 & & 45/64 & 9/64 & & & & \\
 & 7/32 & & -5/32 & -3/32 & & & 3/32 & 5/32 \\
 & & & & 9/16 & & & 3/8 & 15/16 \quad 1 \\
 & & & & 9/16 & & & -3/8 & -5/16 \\
 & & & & & 1 & & 3/4 & \\
\end{bmatrix}
$$

$$
\mathbf{A}^{-} = \begin{bmatrix}
 & & & -5/128 & & 3/128 & 3/128 & & -5/128 & & & & -5/128 & -5/128 & \\
1 & & & -5/128 & & 3/128 & -5/128 & & 35/128 & & & & -5/128 & 3/128 & \\
 & 1 & & & & & & & & & & & & & \\
 & & & 3/16 & & -1/16 & & & & 1 & & & 5/16 & -1/8 & \\
 & & & -3/8 & & & & & & & 1 & & & 3/8 & \\
 & & & 1/2 & & & & & & & & 1 & & & \\
 & & & 1/2 & & & & & & & & & & & \\
 & & & -3/8 & & & & & & & & & & -1/8 & \\
 & & & 3/16 & & -1/16 & & & & & & & 1/16 & 1/8 & \\
 & & & 1/32 & & -3/32 & -5/32 & & 7/32 & & & & 5/32 & 3/32 & \\
 & & 1 & 1/64 & & 9/64 & 45/64 & & -35/64 & & & & -15/64 & -3/64 & \\
 & & & 1/32 & & -3/32 & 15/32 & 1 & 35/32 & & & & 5/32 & -1/32 & \\
 & & & -3/16 & & 9/16 & & & & & & & -5/16 & -3/8 & \\
 & & & -3/16 & & 9/16 & & & & & & & 15/16 & 3/8 & 1 \\
 & & & 3/4 & 1 & & & & & & & & & 3/4 &
\end{bmatrix}.
$$

The fine-to-coarse maps are given by

$$\begin{array}{rcccccccccccccc} i = & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14, \\ c^+(i) = & 2 & 0 & 10 & - & 9 & - & - & 14 & - & 6 & 7 & 8 & - & - & 12, \\ c^-(i) = & 1 & 2 & 10 & - & 14 & - & - & 11 & - & 3 & 4 & 5 & - & - & 13. \end{array}$$

and the coarse-to-fine maps by

$$\begin{array}{rcccccccccccccc} i = & 0 & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 & 11 & 12 & 13 & 14, \\ d^+(i) = & 1 & - & 0 & - & - & - & 9 & 10 & 11 & 4 & 2 & - & 14 & - & 7, \\ d^-(i) = & - & 0 & 1 & 9 & 10 & 11 & - & - & - & - & 2 & 7 & - & 14 & 4. \end{array}$$

Chapter 4

A finite element method for linear elliptic problems in nonvariational form

In this chapter we move from recovery techniques for first derivatives of a finite element function and begin to study applications of second derivative recovery techniques. In this work we make use of the *Hessian recovery* to derive a finite element method for nonvariational form elliptic operators.

Note that the concept of Hessian recovery we use is slightly different to that which is generally studied in the literature. The Hessian recovery we will be making use of is a *representation* of the Hessian of a piecewise smooth object defined in a distributional sense (see Definition 4.1.3.3). The more commonly used are double applications of recovery operators, a global $\mathrm{L}_2(\Omega)$ projection [BX03b, Ova07, cf.].

This chapter is set out as follows. In §4.1 we introduce some notation and set out the model problem. We then present a discretisation scheme for the model problem using standard conforming finite elements in $\mathrm{C}^0(\Omega)$. In §4.2 we present a linear algebra technique, inspired by the *Schur complement* idea, for solving the linear system arising from the discretisation. In §4.3 we show the system proposed in §4.1.3 is well posed. We show for the case of the Laplacian (and constant perturbations of it) the nonvariational finite element method coincides with that of the standard finite element method (cf.

Theorem 4.3.0.5). In §4.4 we propose two methods for the implementation of inhomogeneous Dirichlet boundary conditions, the first is a direct enforcement of the boundary conditions into the problem matrix. The second alters the method proposed in §4.1.3 to incorporate additional information contained in the boundary degrees of freedom. In §4.5 we summarise extensive numerical experiments on model linear boundary value problems in nonvariational form. In §4.6 we deal with the analysis of the method deriving both apriori and residual aposteriori bounds. In §4.7 we numerically demonstrate that these bounds are of the correct order and apply a heuristic adaptive algorithm based on the aforementioned residual bounds. Finally, in §4.8 we apply this method to approximate the solution of general quasilinear problems in nonvariational form.

4.1 Set up

4.1.1 Notation

We recall the notation of §2, in addition we denote by $\langle f \rangle_\omega$ the integral of a function f over the domain ω and drop the subscript if $\omega = \Omega$.

We assume for simplicity that Ω is a polygonal domain and consider the following problem: Find $u \in \mathrm{H}^2(\Omega) \cap \mathrm{H}_0^1(\Omega)$ such that

$$\begin{aligned} \mathscr{L}u &= f \text{ in } \Omega, \\ u &= 0 \text{ on } \partial\Omega, \end{aligned} \tag{4.1}$$

where the data $f : \overline{\Omega} \to \mathbb{R}$ is prescribed and $\mathscr{L}$ is a general linear, second order, uniformly elliptic partial differential operator. Let $\boldsymbol{A} \in \mathrm{L}_\infty(\Omega)^{d\times d}$ and for each $\boldsymbol{x} \in \Omega$ let $\boldsymbol{A}(\boldsymbol{x}) \in \mathrm{Sym}^+(\mathbb{R}^{d\times d})$, the space of symmetric, $d \times d$ matrices such that the operator

$$\begin{aligned} \mathscr{L} : \quad \mathrm{H}^2(\Omega) \cap \mathrm{H}_0^1(\Omega) &\to \mathrm{L}_2(\Omega) \\ u &\mapsto \mathscr{L}u := \boldsymbol{A}{:}\mathrm{D}^2 u, \end{aligned} \tag{4.2}$$

is uniformly elliptic. We use $\boldsymbol{X}{:}\boldsymbol{Y} := \mathrm{trace}\,(\boldsymbol{X}^\intercal \boldsymbol{Y})$ to denote the Frobenius inner product between two matrices.

4.1.2 Classical and strong solutions of nonvariational problems (4.1)

In this section we give a brief review of known results for problems of the form

$$\begin{aligned} \boldsymbol{A}{:}\mathrm{D}^2 u &= f \text{ in } \Omega \\ u &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{4.3}$$

4.1.2.1 Definition (Hölder continuity)**.** A function $v : \Omega \to \mathbb{R}$ is *uniformly α–Hölder continuous* if for any $\boldsymbol{x}, \boldsymbol{y} \in \Omega$

$$\sup_{\boldsymbol{x},\boldsymbol{y}\in\Omega} \frac{|v(\boldsymbol{x}) - v(\boldsymbol{y})|}{|\boldsymbol{x} - \boldsymbol{y}|^{\alpha}} < \infty, \tag{4.4}$$

where $\alpha \in (0, 1]$.

We may view Hölder continuity as a "fractional derivative" of sorts (§A.2). We define the *Hölder spaces*, $\mathrm{C}^{k,\alpha}(\Omega) \subset \mathrm{C}^k(\Omega)$ to be the space consisting of functions whose $k-th$ partial derivatives are uniformly α–Hölder continuous. The *Hölder norms* are defined as

$$\|v\|_{\mathrm{C}^{k,\beta}(\Omega)} = \sum_{|\boldsymbol{\alpha}|<k} \|D^{\boldsymbol{\alpha}} v\|_{\mathrm{L}_\infty(\Omega)} + \sum_{|\boldsymbol{\alpha}|=k} \sup_{\boldsymbol{x},\boldsymbol{y}\in\Omega} \frac{|D^{\boldsymbol{\alpha}} v(\boldsymbol{x}) - D^{\boldsymbol{\alpha}} v(\boldsymbol{y})|}{|\boldsymbol{x} - \boldsymbol{y}|^{\beta}}. \tag{4.5}$$

4.1.2.2 Definition (Hölder domains)**.** A domain $\Omega \subset \mathbb{R}^d$ is said to be a Hölder domain of class $\mathrm{C}^{k,\alpha}$ if at any point, $\boldsymbol{x} \in \partial\Omega$, under an appropriate change of coordinates, the boundary $\partial\Omega$ can be represented as a function, in $\mathrm{C}^{k,\alpha}$.

4.1.2.3 Definition (classical solution)**.** A *classical solution* of (4.3) is a function $u \in \mathrm{C}^2(\overline{\Omega})$ which satisfies the problem (and its boundary conditions) everywhere.

We look at conditions under which the problem (4.3) admits a classical solution via the Schauder theory [GT83, §6]. The fundamental idea behind such an approach is that if the model problem (4.3) has Hölder continuous coefficients, then the problem can be treated (at least in a local sense) as a perturbation of a problem with constant coefficients.

4.1.2.4 Theorem (existence of a classical solution of (4.3) [GT83, Thm 6.14])**.** *Let $\Omega \subset \mathbb{R}^d$ be a $\mathrm{C}^{2,\alpha}$ domain. Suppose that $\boldsymbol{A} \in \mathrm{C}^{0,\alpha}(\overline{\Omega})^{d\times d}$ and $f \in \mathrm{C}^{0,\alpha}(\overline{\Omega})$ are given functions such that the problem*

$$\begin{aligned} \boldsymbol{A}{:}\mathrm{D}^2 u &= f \text{ in } \Omega \\ u &= 0 \text{ on } \partial\Omega, \end{aligned} \tag{4.6}$$

is uniformly elliptic. Then (4.6) admits a unique solution $u \in \mathrm{C}^{2,\alpha}(\overline{\Omega})$*. There also exists a constant independent of* u *such that*

$$\|u\|_{\mathrm{C}^{2,\alpha}(\Omega)} \leq C \|f\|_{\mathrm{C}^{0,\alpha}(\Omega)} . \tag{4.7}$$

We now state a result giving the conditions under which (4.3) admits a strong solution.

4.1.2.5 Definition (strong solution)**.** A *strong solution* of (4.3) is a function $u \in \mathrm{H}^2(\Omega) \cap \mathrm{H}^1_0(\Omega)$, that is a twice weakly differentiable function, which satisfies the problem almost everywhere.

4.1.2.6 Theorem (existence of a strong solution of (4.3) [GT83, Thm 9.15])**.** *Let* $\Omega \subset \mathbb{R}^d$ *be a* $\mathrm{C}^{1,1}$ *domain. Suppose also that* $\boldsymbol{A} \in \mathrm{C}^0(\Omega)^{d\times d}$ *and* $f \in \mathrm{L}_2(\Omega)$ *such that the problem*

$$\begin{aligned} \boldsymbol{A}{:}\mathrm{D}^2 u &= f \text{ in } \Omega \\ u &= 0 \text{ on } \partial\Omega \end{aligned} \tag{4.8}$$

is uniformly elliptic. Then (4.8) has a unique solution $u \in \mathrm{H}^2(\Omega) \cap \mathrm{H}^1_0(\Omega)$*. There also exists a constant independent of* u *such that*

$$\|u\|_2 \leq C \|f\| . \tag{4.9}$$

4.1.2.7 Remark (less regular solutions)**.** Note that the theory of viscosity solutions has been developed for non classical solutions of (4.1) if $\boldsymbol{A}$ does not satisfy the regularity assumed above. A brief description of viscosity solutions is given in §A.5.

4.1.2.8 Assumption (regularity of $\boldsymbol{A}$)**.** From hereonin we will assume that the coefficient matrix $\boldsymbol{A}$ is sufficiently smooth on $\overline{\Omega}$ such that solutions exist and belong to at least $\mathrm{H}^2(\Omega) \cap \mathrm{H}^1_0(\Omega)$.

4.1.2.9 Assumption (regularity of Ω)**.** Theorem 4.1.2.6 specifies that Ω must be a $\mathrm{C}^{1,1}$ domain. We must approximate any such domain with one which is only $\mathrm{C}^{0,1}$. We thus assume that the model problem admits a unique strong solution even when Ω is only $\mathrm{C}^{0,1}$.

4.1.3 Discretisation

In this chapter we use the following notation for finite element spaces, denoting

$$\mathbb{V} := \left\{ \Phi \in \mathrm{H}^1(\Omega) : \ \Phi|_K \in \mathbb{P}^p \, \forall \, K \in \mathscr{T} \right\}, \tag{4.10}$$

$$\mathring{\mathbb{V}} := \mathbb{V} \cap \mathrm{H}^1_0(\Omega), \tag{4.11}$$

where $\mathbb{P}^k$ denotes the linear space of polynomials in d variables of degree no higher than a positive integer k. We consider p to be a fixed integer and denote by

$$\mathring{N} := \dim \mathring{\mathbb{V}} \tag{4.12}$$

$$N := \dim \mathbb{V} \tag{4.13}$$

$$\dot{N} := N - \mathring{N}. \tag{4.14}$$

4.1.3.1 Remark (black and white notation)**.** The notation we are using in this chapter to try to keep things as clear as possible is that a function $\mathring{\phi}$ has a support only on the "interior of Ω", i.e., $\mathring{\phi}\Big|_{\partial\Omega} = 0$. A function $\dot{\phi}$ has support only "close to $\partial\Omega$". In practice we only use this notation for finite element basis functions. If a finite element basis function has no such accent then it does not have the restrictions described.

In view of Remark 4.1.3.1 we define

$$\mathring{\boldsymbol{\Phi}} := (\mathring{\Phi}_1, ..., \mathring{\Phi}_{\mathring{N}})^{\mathsf{T}}, \tag{4.15}$$

$$\boldsymbol{\Phi} := (\mathring{\Phi}_1, ..., \mathring{\Phi}_{\mathring{N}}, \dot{\Phi}_1, ..., \dot{\Phi}_{\dot{N}})^{\mathsf{T}} \text{ and} \tag{4.16}$$

$$\dot{\boldsymbol{\Phi}} := (\dot{\Phi}_1, ..., \dot{\Phi}_{\dot{N}})^{\mathsf{T}} \tag{4.17}$$

where $\{\mathring{\Phi}_1, ..., \mathring{\Phi}_{\mathring{N}}\}$ and $\{\mathring{\Phi}_1, ..., \mathring{\Phi}_{\mathring{N}}, \dot{\Phi}_1, ..., \dot{\Phi}_{\dot{N}}\}$ form a basis of $\mathring{\mathbb{V}}$ and $\mathbb{V}$ respectively. We see that $\mathring{\boldsymbol{\Phi}}, \boldsymbol{\Phi}$ and $\dot{\boldsymbol{\Phi}}$ are all vectors of basis functions.

4.1.3.2 Remark (algebraic notation)**.** We will use a similar notation convention for scalars, vectors and matrices. In this case these objects will be associated to functions satisfying the restrictions above. For example $\dot{N}$ denotes the number of degrees of freedom of $\mathbb{V}$ lying on $\partial\Omega$.

We also use a separate notation for "geometric" and "numerical" matrices. We denote a matrix $\boldsymbol{X} \in \mathbb{R}^{d\times d}$ to be a geometric matrix, where $d = 1, 2, 3$, notice the slanted notation.

For example the matrix $\boldsymbol{A}$ is a geometric matrix. We denote a matrix $\mathsf{X} \in \mathbb{R}^{N\times N}$ to be a numerical matrix, where $N = \dim \mathbb{V}$, notice the upright notation. For example a mass matrix M is a numerical matrix.

In the case of numerical matrices we use the "black and white" notation as a method of quickly assessing the dimension of the matrix, for example the matrix $\mathring{\mathsf{B}}^{\bullet} \in \mathbb{R}^{\mathring{N}\times\bullet N}$.

This notation becomes especially useful in clarifying the linear algebra arguments in §4.2.

Testing the model problem (4.1) with $\phi \in \mathrm{C}^\infty(\overline{\Omega})$ gives

$$\begin{aligned} \langle \mathscr{L}u, \phi\rangle = \big\langle \boldsymbol{A}{:}\mathrm{D}^2 u, \phi\big\rangle = \langle f, \phi\rangle \qquad & \text{in } \Omega \\ u = 0 \qquad & \text{on } \partial\Omega. \end{aligned} \tag{4.18}$$

In order to discretise (4.18) with $\mathring{\mathbb{V}}$ we shall use an appropriate definition of a Hessian of a finite element function. Such a function may not admit a Hessian in the classical sense, so we consider it as a distribution (or generalised function) which we recall the definition, next, and then Riesz–represent it in the FE space itself.

4.1.3.3 Definition (generalised Hessian)**.** Let $\boldsymbol{n} : \partial\Omega \to \mathbb{R}^d$ be the outward pointing normal of Ω. Given a piecewise smooth and continuous function v its *generalised Hessian*, defined in the distributional sense, is given by

$$\big\langle \mathrm{D}^2 v \,|\, \phi\big\rangle = -\langle \nabla v \otimes \nabla\phi\rangle + \langle \nabla v \otimes \boldsymbol{n}\phi\rangle_{\partial\Omega} \quad \forall\, \phi \in \mathrm{C}^\infty(\overline{\Omega}), \tag{4.19}$$

where we are using $\boldsymbol{x}\otimes\boldsymbol{y} := \boldsymbol{x}\boldsymbol{y}^{\intercal}$ to denote the tensor product between two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, and $\langle f\rangle = \int_\Omega f$ as indicated in §4.1.1.

4.1.3.4 Definition (finite element Hessian)**.** We define the *finite element Hessian* as follows. Let $V \in \mathring{\mathbb{V}}$ then

$$\langle \boldsymbol{H}[V], \Phi\rangle := -\langle \nabla V \otimes \nabla\Phi\rangle + \langle \nabla V \otimes \boldsymbol{n}\ \Phi\rangle_{\partial\Omega} \quad \forall\, \Phi \in \mathbb{V}. \tag{4.20}$$

It follows that $\boldsymbol{H}$ is a linear operator on $\mathring{\mathbb{V}}$.

Taking the model problem (4.18) we substitute the finite element Hessian directly, reducing the space of test functions to $\mathring{\mathbb{V}}$, we wish to find $U \in \mathring{\mathbb{V}}$ such that

$$\Big\langle \boldsymbol{A}{:}\boldsymbol{H}[U], \mathring{\Phi}\Big\rangle = \Big\langle f, \mathring{\Phi}\Big\rangle \quad \forall\, \mathring{\Phi} \in \mathring{\mathbb{V}}. \tag{4.21}$$

4.1.3.5 Theorem (nonvariational finite element method (NVFEM))**.** *The* nonvariational finite element solution *for the model problem's discretisation (4.21) is given as* $U = \mathring{\mathbf{\Phi}}^{\mathsf{T}}\mathbf{u}$, *where* $\mathbf{u} \in \mathbb{R}^{\mathring{N}}$ *is the solution to the following linear system*

$$\mathring{\mathring{\mathbf{D}}}\mathring{\mathbf{u}} := \sum_{\alpha=1}^{d}\sum_{\beta=1}^{d} \mathring{\mathbf{B}}^{\alpha\beta}\mathbf{M}^{-1}\mathring{\mathbf{C}}_{\alpha\beta}\mathring{\mathbf{u}} = \mathring{\mathbf{f}}. \tag{4.22}$$

Let $\boldsymbol{A} = [a^{\alpha\beta}]_{\alpha\beta=1}^{d}$, *the components of (4.22) are then given by*

$$\mathring{\mathbf{B}}^{\alpha\beta} := \left\langle \mathring{\mathbf{\Phi}}, a^{\alpha\beta}\mathbf{\Phi}^{\mathsf{T}} \right\rangle \in \mathbb{R}^{\mathring{N}\times N}, \tag{4.23}$$

$$\mathbf{M} := \langle \mathbf{\Phi}, \mathbf{\Phi}^{\mathsf{T}} \rangle \in \mathbb{R}^{N\times N}, \tag{4.24}$$

$$\mathring{\mathbf{C}}_{\alpha\beta} := -\left\langle \partial_\beta \mathbf{\Phi}, \partial_\alpha \mathring{\mathbf{\Phi}}^{\mathsf{T}} \right\rangle + \left\langle \mathbf{\Phi} n_\beta, \partial_\alpha \mathring{\mathbf{\Phi}}^{\mathsf{T}} \right\rangle_{\partial\Omega} \in \mathbb{R}^{N\times\mathring{N}}, \tag{4.25}$$

$$\mathring{\mathbf{f}} := \left\langle f, \mathring{\mathbf{\Phi}} \right\rangle \in \mathbb{R}^{\mathring{N}}. \tag{4.26}$$

Proof Since $\boldsymbol{H}[U] \in \mathbb{V}^{d\times d}$ we will denote $\boldsymbol{H}[U] = [\boldsymbol{H}_{\alpha\beta}[U]]_{\alpha\beta=1}^{d}$ for each $\alpha, \beta = 1, \ldots, d$, $\boldsymbol{H}_{\alpha\beta}[U] = \mathbf{\Phi}^{\mathsf{T}}\mathbf{h}_{\alpha\beta}$. Then, testing (4.21) with $\mathring{\mathbf{\Phi}}$,

$$\begin{aligned}
\left\langle f, \mathring{\mathbf{\Phi}} \right\rangle &= \sum_{\alpha=1}^{d}\sum_{\beta=1}^{d} \left\langle a^{\alpha\beta}\boldsymbol{H}_{\alpha\beta}[U], \mathring{\mathbf{\Phi}} \right\rangle \\
&= \sum_{\alpha=1}^{d}\sum_{\beta=1}^{d} \left\langle \mathring{\mathbf{\Phi}}, a^{\alpha\beta}\mathbf{\Phi}^{\mathsf{T}}\mathbf{h}_{\alpha\beta} \right\rangle \\
&= \sum_{\alpha=1}^{d}\sum_{\beta=1}^{d} \left\langle \mathring{\mathbf{\Phi}}, a^{\alpha\beta}\mathbf{\Phi}^{\mathsf{T}} \right\rangle \mathbf{h}_{\alpha\beta}. \\
&= \sum_{\alpha=1}^{d}\sum_{\beta=1}^{d} \mathring{\mathbf{B}}^{\alpha\beta}\mathbf{h}_{\alpha\beta}
\end{aligned} \tag{4.27}$$

Utilising Definition 4.1.3.4 for each $\alpha, \beta = 1 \ldots d$ we can compute $\mathbf{h}_{\alpha\beta} \in \mathbb{R}^N$, noting $U = \mathring{\mathbf{\Phi}}^{\mathsf{T}}\mathbf{u}$,

$$\begin{aligned}
\langle \mathbf{\Phi}, \mathbf{\Phi}^{\mathsf{T}} \rangle \mathbf{h}_{\alpha\beta} &= \langle \mathbf{\Phi}, \boldsymbol{H}_{\alpha\beta}[U] \rangle \\
&= -\langle \partial_\beta \mathbf{\Phi}, \partial_\alpha U \rangle + \langle \mathbf{\Phi}\boldsymbol{n}_\beta, \partial_\alpha U \rangle_{\partial\Omega} \\
&= \left(-\left\langle \partial_\beta \mathbf{\Phi}, \partial_\alpha \mathring{\mathbf{\Phi}}^{\mathsf{T}} \right\rangle + \left\langle \mathbf{\Phi}\boldsymbol{n}_\beta, \partial_\alpha \mathring{\mathbf{\Phi}}^{\mathsf{T}} \right\rangle_{\partial\Omega} \right) \mathring{\mathbf{u}}.
\end{aligned} \tag{4.28}$$

Using the definition of $\mathbf{M}$ (4.24) and $\mathring{\mathbf{C}}_{\alpha\beta}$ (4.25) we see that for each $\alpha, \beta = 1 \ldots d$, we have

$$\mathbf{M}\mathbf{h}_{\alpha\beta} = \mathring{\mathbf{C}}_{\alpha\beta}\mathring{\mathbf{u}}, \tag{4.29}$$

i.e.,

$$\mathbf{h}_{\alpha\beta} = \mathbf{M}^{-1}\mathring{\mathbf{C}}_{\alpha\beta}\mathring{\mathbf{u}}. \tag{4.30}$$

Substituting $\mathbf{h}_{\alpha\beta}$ from (4.29) into (4.27) we obtain the desired result. □

To further illustrate the method we present the discretisation of some simple examples.

4.1.3.6 Example (for $d = 1$ and general $\boldsymbol{A}$)**.** In this example we consider the problem

$$\begin{aligned} a(x)\Delta u &= f \text{ in } \Omega \\ u &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{4.31}$$

Here we are looking at the case $d = 1$ to clarify the situation for general operators. In this instance we discretise the problem by seeking $U \in \mathring{\mathbb{V}}$ such that

$$\left\langle f, \mathring{\Phi} \right\rangle = \left\langle a(x)\boldsymbol{H}[U], \mathring{\Phi} \right\rangle \quad \forall\, \mathring{\Phi} \in \mathring{\mathbb{V}}. \tag{4.32}$$

We proceed by discretising this problem in a similar fashion to the Proof of Theorem 4.1.3.5.

$$\begin{aligned} \mathring{\mathbf{f}} := \left\langle f, \mathring{\boldsymbol{\Phi}} \right\rangle &= \left\langle \mathring{\boldsymbol{\Phi}}, a(x)\boldsymbol{\Phi}^{\mathsf{T}}\mathbf{h} \right\rangle \\ &= \left\langle \mathring{\boldsymbol{\Phi}}, a(x)\boldsymbol{\Phi}^{\mathsf{T}} \right\rangle \mathbf{h}. \end{aligned} \tag{4.33}$$

Setting

$$\mathring{\mathbf{B}} := \left\langle \mathring{\boldsymbol{\Phi}}, a(x)\boldsymbol{\Phi}^{\mathsf{T}} \right\rangle \tag{4.34}$$

the finite element coefficient vector is given as the solution of the following linear system: Find $\mathring{\mathbf{u}}$ such that

$$\mathring{\mathbf{B}}\mathbf{M}^{-1}\mathring{\mathbf{C}}\mathring{\mathbf{u}} = \mathring{\mathbf{f}}. \tag{4.35}$$

4.1.3.7 Example (for $d = 1$ and general $\boldsymbol{A}$)**.** To further clarify the notation we use, we will follow Example (4.1.3.6) using a more standard notation.

Recall that $\boldsymbol{H}[U] \in \mathbb{V}$ so we may write

$$\boldsymbol{H}[U] = \sum_{j=1}^{N} \mathbf{h}_j \boldsymbol{\Phi}_j. \tag{4.36}$$

Also $U \in \mathring{\mathbb{V}}$ so it follows

$$U = \sum_{j=1}^{\mathring{N}} \mathbf{u}_j \mathring{\boldsymbol{\Phi}}_j. \tag{4.37}$$

As in the previous example we note that

$$\begin{aligned}\mathsf{f}_i := \left\langle f, \mathring{\boldsymbol{\Phi}}_i \right\rangle &= \left\langle \mathring{\boldsymbol{\Phi}}_i, a(x) \sum_{j=1}^{N} \mathsf{h}_j \boldsymbol{\Phi}_j \right\rangle \\ &= \sum_{j=1}^{N} \left\langle \mathring{\boldsymbol{\Phi}}_i, a(x) \boldsymbol{\Phi}_j \right\rangle \mathsf{h}_j \\ &=: \sum_{j=1}^{N} \mathring{\mathsf{B}}_{i,j} \mathsf{h}_j \quad \forall\, i = [1 : \mathring{N}],\end{aligned} \tag{4.38}$$

where now we use $\mathring{\mathsf{B}}_{i,j}$ to denote the i,j-th component of $\mathring{\mathsf{B}}$. Now from Definition 4.1.3.4 of the finite element Hessian it is clear

$$\left\langle \boldsymbol{\Phi}_i, \sum_{j=1}^{N} \mathsf{h}_j \boldsymbol{\Phi}_j \right\rangle = -\left\langle \sum_{j=1}^{\mathring{N}} \mathring{\mathsf{u}}_j \nabla \mathring{\boldsymbol{\Phi}}_j, \nabla \boldsymbol{\Phi}_i \right\rangle + \left\langle \sum_{j=1}^{\mathring{N}} \mathring{\mathsf{u}}_j \nabla \mathring{\boldsymbol{\Phi}}_j, \boldsymbol{\Phi}_i \boldsymbol{n} \right\rangle_{\partial\Omega} \qquad \forall i = [1 : N]. \tag{4.39}$$

Hence we see

$$\sum_{j=1}^{N} \langle \boldsymbol{\Phi}_i, \boldsymbol{\Phi}_j \rangle \, \mathsf{h}_j = \sum_{j=1}^{\mathring{N}} \left(-\left\langle \nabla \boldsymbol{\Phi}_i, \nabla \mathring{\boldsymbol{\Phi}}_j \right\rangle + \left\langle \nabla \mathring{\boldsymbol{\Phi}}_j, \boldsymbol{\Phi}_i \boldsymbol{n} \right\rangle_{\partial\Omega} \right) \mathring{\mathsf{u}}_j \quad \forall\, i = [1 : N]. \tag{4.40}$$

Which gives us that

$$\sum_{j=1}^{N} \mathsf{M}_{i,j} \mathsf{h}_j = \sum_{j=1}^{\mathring{N}} \mathring{\mathsf{C}}_{i,j} \mathring{\mathsf{u}}_j \quad \forall\, i = [1 : N]. \tag{4.41}$$

Combining (4.38) and (4.41) gives the desired result from Example (4.1.3.6).

We refrain from using this notation as it becomes extremely cumbersome especially for $d > 1$.

4.1.3.8 Example (for $d = 2$ and $\boldsymbol{A} = -\boldsymbol{I}$)**.** In this example we consider the problem

$$\begin{aligned} -\Delta u &= f \text{ in } \Omega \\ u &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{4.42}$$

The discrete formulation of this problem (in view of (4.21)) is: Find $U \in \mathring{\mathbb{V}}$ such that

$$\left\langle -\operatorname{trace} \boldsymbol{H}[U], \mathring{\Phi} \right\rangle = \left\langle f, \mathring{\Phi} \right\rangle \quad \forall\, \mathring{\Phi} \in \mathring{\mathbb{V}}. \tag{4.43}$$

We proceed along the same lines as the previous examples.

$$\begin{aligned}\mathring{\mathsf{f}} := \left\langle f, \mathring{\boldsymbol{\Phi}} \right\rangle &= \left\langle -\operatorname{trace} \boldsymbol{H}[U], \mathring{\boldsymbol{\Phi}} \right\rangle \\ &= \sum_{\alpha=1}^{d} \left\langle \mathring{\boldsymbol{\Phi}}, \boldsymbol{\Phi}^{\intercal} \mathsf{h}_{\alpha,\alpha} \right\rangle \\ &= \sum_{\alpha=1}^{d} \left\langle \mathring{\boldsymbol{\Phi}}, \boldsymbol{\Phi}^{\intercal} \right\rangle \mathsf{h}_{\alpha,\alpha}.\end{aligned} \tag{4.44}$$

Recall from (4.30) we have the coefficient vector of the finite element Hessian $\mathbf{h}_{\alpha,\alpha}$ explicitly given in terms of that of the finite element coefficient vector $\mathring{\mathbf{u}}$, thus if we define

$$\mathring{\mathbf{B}} := \left\langle \mathring{\mathbf{\Phi}}, \mathbf{\Phi}^{\mathsf{T}} \right\rangle \tag{4.45}$$

we see the finite element coefficient vector is given as the solution of the following linear system: Find $\mathring{\mathbf{u}}$ such that

$$\sum_{\alpha=1}^{d} \mathring{\mathbf{B}}\mathbf{M}^{-1}\mathring{\mathbf{C}}_{\alpha\alpha}\mathring{\mathbf{u}} = \mathring{\mathbf{f}}. \tag{4.46}$$

Note that although the matrix $\mathring{\mathbf{B}}$ resembles the mass matrix $\mathbf{M}$ they are not equal since $\mathring{\mathbf{B}} \in \mathbb{R}^{\mathring{N}\times N}$ but $\mathbf{M} \in \mathbb{R}^{N\times N}$. Although later in this Chapter, specifically in Theorem 4.3.0.5, it will be shown that for simple problems, like the Laplacian, this system is equivalent to a simpler one, that of the standard finite element stiffness matrix.

4.1.3.9 Example (for $d = 2$ and general $\boldsymbol{A}$)**.** For a general elliptic operator of the form

$$\begin{aligned} \boldsymbol{A}{:}\mathrm{D}^2 u &= f \text{ in } \Omega \\ u &= 0 \text{ on } \partial\Omega \end{aligned} \tag{4.47}$$

the formulation (4.22) takes the form

$$\left(\mathring{\mathbf{B}}^{11}\mathbf{M}^{-1}\mathring{\mathbf{C}}_{11} + \mathring{\mathbf{B}}^{22}\mathbf{M}^{-1}\mathring{\mathbf{C}}_{22} + \mathring{\mathbf{B}}^{12}\mathbf{M}^{-1}\mathring{\mathbf{C}}_{12} + \mathring{\mathbf{B}}^{21}\mathbf{M}^{-1}\mathring{\mathbf{C}}_{21}\right)\mathring{\mathbf{u}} = \mathring{\mathbf{f}}. \tag{4.48}$$

4.2 Solution of the linear system

4.2.0.10 Remark (solving (4.22) is computationally intense)**.** The system matrix $\mathring{\mathring{\mathbf{D}}} = \sum\sum \mathring{\mathbf{B}}^{\alpha\beta}\mathbf{M}^{-1}\mathring{\mathbf{C}}_{\alpha\beta}$ (4.22) is generally not sparse, ruling out the use of efficient iterative solvers.

In this section we will present a method to solve formulation (4.22) in a general setting. This method makes use of the sparsity of the component matrices $\mathring{\mathbf{B}}^{\alpha\beta}, \mathring{\mathbf{C}}_{\alpha\beta}$ and $\mathbf{M}$.

4.2.0.11 Remark (diagonalising $\mathring{\mathring{\mathbf{D}}}$)**.** An interesting point of note is that if the mass matrix $\mathbf{M}$ were diagonalised, by mass lumping, then for each α and β the matrix $\mathring{\mathbf{B}}^{\alpha\beta}\mathbf{M}^{-1}\mathring{\mathbf{C}}_{\alpha\beta}$ would still be sparse (albeit less so than the individual matrices $\mathring{\mathbf{B}}^{\alpha\beta}$ and $\mathring{\mathbf{C}}_{\alpha\beta}$). Hence

the system can be easily solved using existing sparse methods. However mass lumping is only applicable to $\mathbb{P}^1$ finite elements and since it is a quadrature approximation we loose resolution for more complicated problems. For higher order finite elements it would be desirable to exploit the sparse structure of the component matrices that make up the system.

4.2.1 A generalised Schur complement

We observe the matrix $\mathring{\mathring{\mathbf{D}}}$ in the system (4.22) is a sum of Schur complements $\mathring{\mathbf{B}}^{\alpha\beta}\mathbf{M}^{-1}\mathring{\mathbf{C}}_{\alpha\beta}$. With that in mind we introduce the $(d^2+1)^2$ block matrix

$$\mathbf{E} = \begin{bmatrix} \mathbf{M} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & -\mathring{\mathbf{C}}_{11} \\ \mathbf{0} & \mathbf{M} & \cdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & -\mathring{\mathbf{C}}_{12} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{M} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & -\mathring{\mathbf{C}}_{1d} \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{M} & \cdots & \mathbf{0} & \mathbf{0} & -\mathring{\mathbf{C}}_{21} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{M} & \mathbf{0} & -\mathring{\mathbf{C}}_{dd-1} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{M} & -\mathring{\mathbf{C}}_{dd} \\ \mathring{\mathbf{B}}^{11} & \mathring{\mathbf{B}}^{12} & \dots & \mathring{\mathbf{B}}^{1d} & \mathring{\mathbf{B}}^{21} & \dots & \mathring{\mathbf{B}}^{dd-1} & \mathring{\mathbf{B}}^{dd} & \mathbf{0} \end{bmatrix}. \tag{4.49}$$

4.2.1.1 Lemma (generalised Schur complement)**.** *Given*

$$\mathbf{v} = \left[\mathbf{h}_{1,1}, \mathbf{h}_{1,2}, \dots, \mathbf{h}_{d,d-1}, \mathbf{h}_{d,d}, \mathring{\mathbf{u}}\right]^{\mathsf{T}}, \tag{4.50}$$

$$\mathbf{b} = \left[\mathbf{0}, \mathbf{0} \dots, \mathbf{0}, \mathbf{0}, \mathring{\mathbf{f}}\right]^{\mathsf{T}}, \tag{4.51}$$

solving the system

$$\mathring{\mathring{\mathbf{D}}}\mathring{\mathbf{u}} = \sum_{\alpha=1}^{d}\sum_{\beta=1}^{d} \mathring{\mathbf{B}}^{\alpha\beta}\mathbf{M}^{-1}\mathring{\mathbf{C}}_{\alpha\beta}\mathring{\mathbf{u}} = \mathring{\mathbf{f}}, \tag{4.52}$$

is equivalent to solving

$$\mathbf{E}\mathbf{v} = \mathbf{b} \tag{4.53}$$

for $\mathring{\mathbf{u}}$.

Proof The proof is just block Gaussian elimination on $\mathbf{E}$. Left-multiplying the first d^2 rows by $\mathbf{M}^{-1}$ yields

$$\begin{bmatrix} \mathbf{I} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & -\mathbf{M}^{-1}\mathring{\mathbf{C}}_{11} \\ \mathbf{0} & \mathbf{I} & \cdots & \mathbf{0} & \mathbf{0} & -\mathbf{M}^{-1}\mathring{\mathbf{C}}_{12} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I} & \mathbf{0} & -\mathbf{M}^{-1}\mathring{\mathbf{C}}_{dd-1} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{I} & -\mathbf{M}^{-1}\mathring{\mathbf{C}}_{dd} \\ \mathring{\mathbf{B}}^{11} & \mathring{\mathbf{B}}^{12} & \dots & \mathring{\mathbf{B}}^{dd-1} & \mathring{\mathbf{B}}^{dd} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{h}_{1,1} \\ \mathbf{h}_{1,2} \\ \vdots \\ \mathbf{h}_{d,d-1} \\ \mathbf{h}_{d,d} \\ \mathring{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{0} \\ \mathring{\mathbf{f}} \end{bmatrix}. \tag{4.54}$$

Multiplying the i-th row by the i-th entry of the (d^2+1)-th row for $i = 1, \dots, d^2$

$$\begin{bmatrix} \mathring{\mathbf{B}}^{11} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & -\mathring{\mathbf{B}}^{11}\mathbf{M}^{-1}\mathring{\mathbf{C}}_{11} \\ \mathbf{0} & \mathring{\mathbf{B}}^{12} & \cdots & \mathbf{0} & \mathbf{0} & -\mathring{\mathbf{B}}^{12}\mathbf{M}^{-1}\mathring{\mathbf{C}}_{12} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathring{\mathbf{B}}^{dd-1} & \mathbf{0} & -\mathring{\mathbf{B}}^{dd-1}\mathbf{M}^{-1}\mathring{\mathbf{C}}_{dd-1} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathring{\mathbf{B}}^{dd} & -\mathring{\mathbf{B}}^{dd}\mathbf{M}^{-1}\mathring{\mathbf{C}}_{dd} \\ \mathring{\mathbf{B}}^{11} & \mathring{\mathbf{B}}^{12} & \dots & \mathring{\mathbf{B}}^{dd-1} & \mathring{\mathbf{B}}^{dd} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{h}_{1,1} \\ \mathbf{h}_{1,2} \\ \vdots \\ \mathbf{h}_{d,d-1} \\ \mathbf{h}_{d,d} \\ \mathring{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{0} \\ \mathring{\mathbf{f}} \end{bmatrix}. \tag{4.55}$$

Subtracting each of the first d^2 rows from the (d^2+1)-th row reduces the system into row echelon form

$$\begin{bmatrix} \mathring{\mathbf{B}}^{11} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & -\mathring{\mathbf{B}}^{11}\mathbf{M}^{-1}\mathring{\mathbf{C}}_{11} \\ \mathbf{0} & \mathring{\mathbf{B}}^{12} & \cdots & \mathbf{0} & \mathbf{0} & -\mathring{\mathbf{B}}^{12}\mathbf{M}^{-1}\mathring{\mathbf{C}}_{12} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathring{\mathbf{B}}^{dd-1} & \mathbf{0} & -\mathring{\mathbf{B}}^{dd-1}\mathbf{M}^{-1}\mathring{\mathbf{C}}_{dd-1} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathring{\mathbf{B}}^{dd} & -\mathring{\mathbf{B}}^{dd}\mathbf{M}^{-1}\mathring{\mathbf{C}}_{dd} \\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} & \mathring{\mathring{\mathbf{D}}} \end{bmatrix} \begin{bmatrix} \mathbf{h}_{1,1} \\ \mathbf{h}_{1,2} \\ \vdots \\ \mathbf{h}_{d,d-1} \\ \mathbf{h}_{d,d} \\ \mathring{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{0} \\ \mathring{\mathbf{f}} \end{bmatrix}. \tag{4.56}$$

□

4.2.1.2 Remark (structure of the block matrix)**.** In fact this method for the solution of the system $\mathring{\mathring{\mathbf{D}}}\mathring{\mathbf{u}} = \mathring{\mathbf{f}}$ is not surprising given the discretisation presented in the proof of Theorem 4.1.3.5 is equivalent to the following system:

Find $U \in \mathring{\mathbb{V}}$ and $\boldsymbol{H}[U] \in \mathbb{V}^{d\times d}$ such that

$$\begin{cases} \langle \boldsymbol{H}[U], \Phi \rangle = -\langle \nabla U \otimes \nabla \Phi \rangle + \langle \nabla U \otimes \boldsymbol{n}\ \Phi \rangle_{\partial\Omega} & \forall\, \Phi \in \mathbb{V} \\ \left\langle \boldsymbol{A} : \boldsymbol{H}[U], \mathring{\Phi} \right\rangle = \left\langle f, \mathring{\Phi} \right\rangle & \forall\, \mathring{\Phi} \in \mathring{\mathbb{V}}. \end{cases} \tag{4.57}$$

4.2.1.3 Remark (complexity of the block matrix E)**.** Observe E is a $(d^2+1)^2$ block matrix. We can use the symmetry of the problem to reduce the dimension of E. Recall that $\boldsymbol{A} \in \mathrm{Sym}^+(\mathbb{R}^{d\times d})$, this implies for each $\alpha,\beta = 1,\dots,d$ that $\mathring{\mathsf{B}}^{\alpha\beta} = \mathring{\mathsf{B}}^{\beta\alpha}$. From Definition 4.1.3.4, $\boldsymbol{H}[U] \in \mathrm{Sym}^+(\mathbb{R}^{d\times d})$ giving for each $\alpha,\beta = 1,\dots,d$ that $\mathring{\mathsf{C}}_{\alpha\beta} = \mathring{\mathsf{C}}_{\beta\alpha}$. Hence the system matrix can be simplified to

$$\mathring{\mathring{\mathsf{D}}}\mathring{\mathsf{u}} = \sum_{\alpha=1}^{d} \mathring{\mathsf{B}}^{\alpha\alpha}\mathsf{M}^{-1}\mathring{\mathsf{C}}_{\alpha\alpha}\mathring{\mathsf{u}} + 2\sum_{\alpha=1}^{d}\sum_{\beta>\alpha}^{d} \mathring{\mathsf{B}}^{\alpha\beta}\mathsf{M}^{-1}\mathring{\mathsf{C}}_{\alpha\beta}\mathring{\mathsf{u}}. \tag{4.58}$$

Upon applying the same Schur complement argument given in the proof of Lemma 4.2.1.1 the size of E is reduced to $((d^2+d)/2+1)^2$.

4.2.1.4 Remark (storage issues)**.** We will be using the generalised minimal residual method (GMRES) to solve this system. The GMRES, as with any iterative solver, requires a subroutine to compute a matrix-vector multiplication. Hence we need to store the component matrices $\mathring{\mathsf{B}}^{\alpha\beta}, \mathring{\mathsf{C}}_{\alpha\beta}$ and M.

4.3 Invertibility of the system

In this section we will show the system (4.22) and by Lemma 4.2.1.1 the equivalent block system (5.41) are both well posed.

4.3.0.5 Theorem (equivalence to the standard FEM)**.** *In the case that the problem coefficients in (4.18) are (piecewise) constant then the problem*

$$\boldsymbol{A}{:}\mathrm{D}^2 u = \operatorname{div} \boldsymbol{A}\nabla u \tag{4.59}$$

and the nonvariational finite element solution coincides with that of the standard finite element method. That is $\mathring{\mathsf{u}}$ *solves both*

$$\mathring{\mathring{\mathsf{D}}}\mathring{\mathsf{u}} = \mathring{\mathsf{f}} \tag{4.60}$$

and

$$\mathring{\mathring{\mathsf{S}}}\mathring{\mathsf{u}} = \mathring{\mathsf{f}}. \tag{4.61}$$

Where

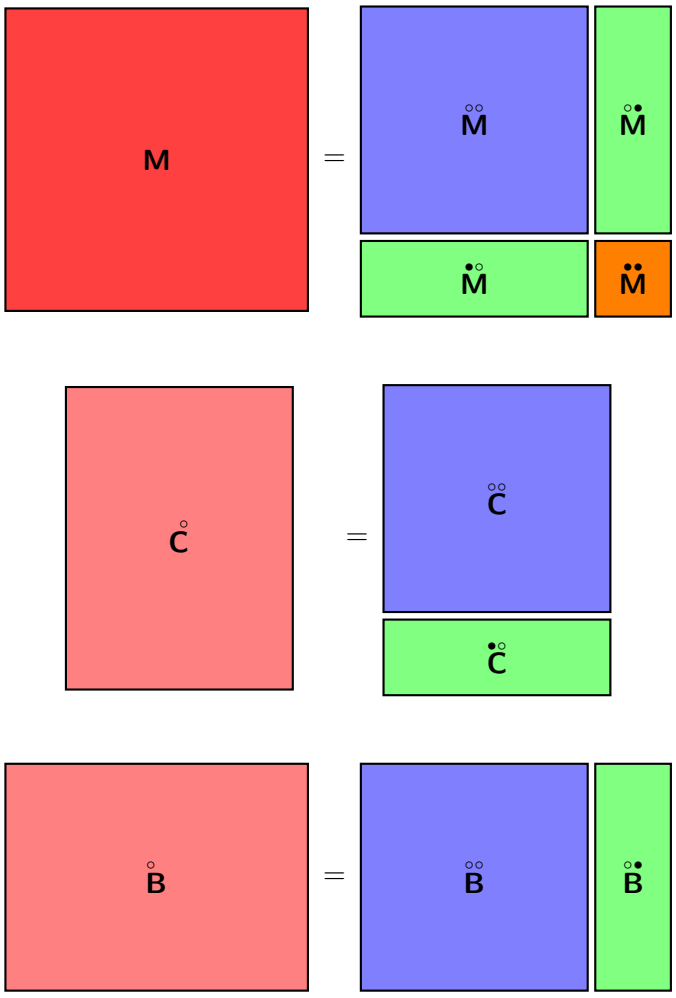
$$\mathring{\mathring{\mathsf{S}}} = \sum_{\alpha,\beta=1}^{d} \left\langle \partial_\beta \mathring{\boldsymbol{\Phi}}, a^{\alpha,\beta}\partial_\alpha \mathring{\boldsymbol{\Phi}}^{\mathsf{T}} \right\rangle \tag{4.62}$$

is the standard finite element stiffness matrix.

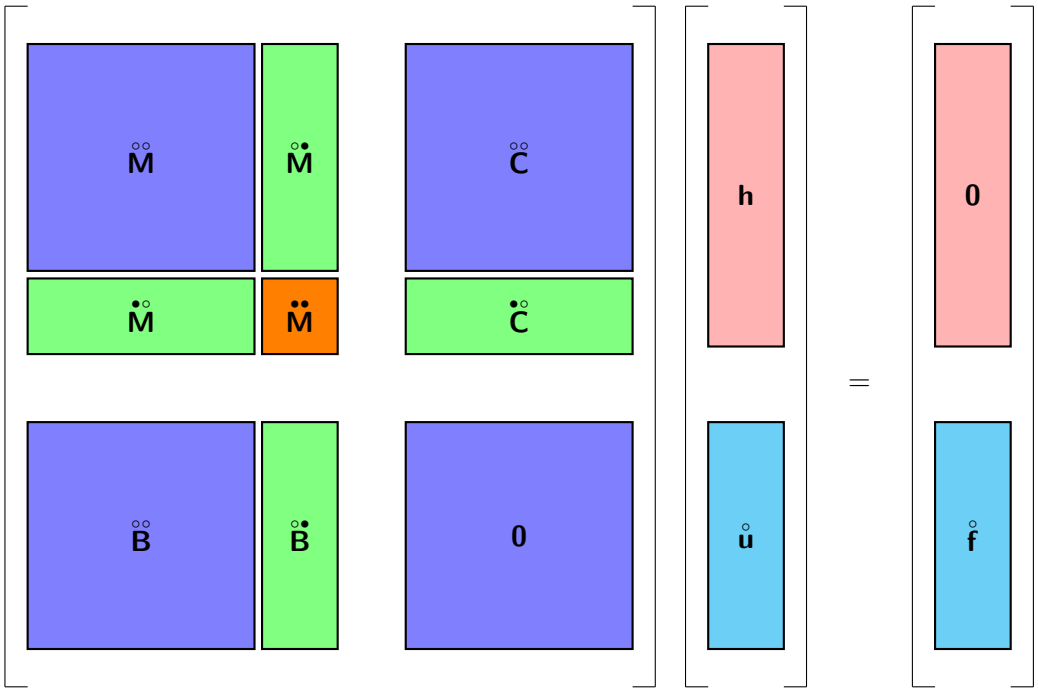
Proof For clarity we will first present the case when $d = 1$ and $\boldsymbol{A} = -\boldsymbol{I}$. We work on the equivalent block system

$$\begin{bmatrix} \mathsf{M} & -\mathring{\mathsf{C}} \\ \mathring{\mathsf{B}} & \mathsf{0} \end{bmatrix} \begin{bmatrix} \mathsf{h} \\ \mathring{\mathsf{u}} \end{bmatrix} = \begin{bmatrix} \mathsf{0} \\ \mathring{\mathsf{f}} \end{bmatrix}. \tag{4.63}$$

We are going to assume the matrix is ordered such that we may split the components of the block system as follows



Hence the system we will study is given as



where our algebraic notation comes from Remark 4.1.3.2 so

$$\overset{\circ\circ}{\mathbf{B}}, \overset{\circ\circ}{\mathbf{M}}, \overset{\circ\circ}{\mathbf{C}} \in \mathbb{R}^{\overset{\circ}{N}\times\overset{\circ}{N}} \tag{4.64}$$

$$\overset{\circ\bullet}{\mathbf{B}}, \overset{\circ\bullet}{\mathbf{M}} \in \mathbb{R}^{\overset{\circ}{N}\times\overset{\bullet}{N}} \tag{4.65}$$

$$\overset{\bullet\circ}{\mathbf{M}}, \overset{\bullet\circ}{\mathbf{C}} \in \mathbb{R}^{\overset{\bullet}{N}\times\overset{\circ}{N}} \tag{4.66}$$

$$\overset{\bullet\bullet}{\mathbf{M}} \in \mathbb{R}^{\overset{\bullet}{N}\times\overset{\bullet}{N}}. \tag{4.67}$$

We begin by reducing the block system

$$\begin{bmatrix} \overset{\circ\circ}{\mathbf{M}} & \overset{\circ\bullet}{\mathbf{M}} & -\overset{\circ\circ}{\mathbf{C}} \\ \overset{\bullet\circ}{\mathbf{M}} & \overset{\bullet\bullet}{\mathbf{M}} & -\overset{\bullet\circ}{\mathbf{C}} \\ \overset{\circ\circ}{\mathbf{B}} & \overset{\circ\bullet}{\mathbf{B}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \overset{\circ}{\mathbf{h}} \\ \overset{\bullet}{\mathbf{h}} \\ \overset{\circ}{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \overset{\circ}{\mathbf{f}} \end{bmatrix} \tag{4.68}$$

into echelon form. Multiplying the first row by $\overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}$ and subtracting the result from the second row gives

$$\begin{bmatrix} \overset{\circ\circ}{\mathbf{M}} & \overset{\circ\bullet}{\mathbf{M}} & -\overset{\circ\circ}{\mathbf{C}} \\ \mathbf{0} & \overset{\bullet\bullet}{\mathbf{M}} - \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}} & -\overset{\bullet\circ}{\mathbf{C}} + \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\circ}{\mathbf{C}} \\ \overset{\circ\circ}{\mathbf{B}} & \overset{\circ\bullet}{\mathbf{B}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \overset{\circ}{\mathbf{h}} \\ \overset{\bullet}{\mathbf{h}} \\ \overset{\circ}{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \overset{\circ}{\mathbf{f}} \end{bmatrix}. \tag{4.69}$$

Multiplying the first row by $\overset{\circ\circ}{\mathbf{B}}\overset{\circ\circ}{\mathbf{M}}^{-1}$ and subtracting the result from the third row gives

$$\begin{bmatrix} \overset{\circ\circ}{\mathbf{M}} & \overset{\circ\bullet}{\mathbf{M}} & -\overset{\circ\circ}{\mathbf{C}} \\ \mathbf{0} & \overset{\bullet\bullet}{\mathbf{M}} - \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}} & -\overset{\bullet\circ}{\mathbf{C}} + \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\circ}{\mathbf{C}} \\ \mathbf{0} & \overset{\circ\bullet}{\mathbf{B}} - \overset{\circ\circ}{\mathbf{B}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}} & \overset{\circ\circ}{\mathbf{B}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\circ}{\mathbf{C}} \end{bmatrix} \begin{bmatrix} \overset{\circ}{\mathbf{h}} \\ \overset{\bullet}{\mathbf{h}} \\ \overset{\circ}{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \overset{\circ}{\mathbf{f}} \end{bmatrix}. \tag{4.70}$$

Finally multiplying the second row by $\left(\overset{\circ\bullet}{\mathbf{B}} - \overset{\circ\circ}{\mathbf{B}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}}\right)\left(\overset{\bullet\bullet}{\mathbf{M}} - \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}}\right)^{-1}$ and taking the result from the third row results in

$$\begin{bmatrix} \overset{\circ\circ}{\mathbf{M}} & \overset{\circ\bullet}{\mathbf{M}} & -\overset{\circ\circ}{\mathbf{C}} \\ \mathbf{0} & \overset{\bullet\bullet}{\mathbf{M}} - \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}} & -\overset{\bullet\circ}{\mathbf{C}} + \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\circ}{\mathbf{C}} \\ \mathbf{0} & \mathbf{0} & \begin{array}{c}\overset{\circ\circ}{\mathbf{B}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\circ}{\mathbf{C}} + \\ \left(\overset{\circ\bullet}{\mathbf{B}} - \overset{\circ\circ}{\mathbf{B}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}}\right)\left(\overset{\bullet\bullet}{\mathbf{M}} - \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}}\right)^{-1}\left(\overset{\circ\circ}{\mathbf{C}} - \overset{\circ\bullet}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\circ}{\mathbf{C}}\right)\end{array} \end{bmatrix} \begin{bmatrix} \overset{\circ}{\mathbf{h}} \\ \overset{\bullet}{\mathbf{h}} \\ \overset{\circ}{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \overset{\circ}{\mathbf{f}} \end{bmatrix}. \tag{4.71}$$

Since $\boldsymbol{A} = -\boldsymbol{I}$, it follows that

$$\begin{aligned} \overset{\circ\circ}{\mathbf{B}} &= \left\langle \overset{\circ}{\boldsymbol{\Phi}}, \boldsymbol{A}\overset{\circ}{\boldsymbol{\Phi}}^{\mathsf{T}} \right\rangle = -\left\langle \overset{\circ}{\boldsymbol{\Phi}}, \overset{\circ}{\boldsymbol{\Phi}}^{\mathsf{T}} \right\rangle = -\overset{\circ\circ}{\mathbf{M}} \text{ and} \\ \overset{\circ\bullet}{\mathbf{B}} &= \left\langle \overset{\circ}{\boldsymbol{\Phi}}, \boldsymbol{A}\overset{\bullet}{\boldsymbol{\Phi}}^{\mathsf{T}} \right\rangle = -\left\langle \overset{\circ}{\boldsymbol{\Phi}}, \overset{\bullet}{\boldsymbol{\Phi}}^{\mathsf{T}} \right\rangle = -\overset{\circ\bullet}{\mathbf{M}} \end{aligned} \tag{4.72}$$

from this we deduce that

$$\overset{\circ\bullet}{\mathbf{B}} - \overset{\circ\circ}{\mathbf{B}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}} = \mathbf{0} \tag{4.73}$$

and thus

$$\left(\overset{\circ\bullet}{\mathbf{B}} - \overset{\circ\circ}{\mathbf{B}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}}\right)\left(\overset{\bullet\bullet}{\mathbf{M}} - \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}}\right)^{-1}\left(\overset{\bullet\circ}{\mathbf{C}} - \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\circ}{\mathbf{C}}\right) = \mathbf{0}. \tag{4.74}$$

Now the block system is given as follows

$$\begin{bmatrix} \overset{\circ\circ}{\mathbf{M}} & \overset{\circ\bullet}{\mathbf{M}} & -\overset{\circ\circ}{\mathbf{C}} \\ \mathbf{0} & \overset{\bullet\bullet}{\mathbf{M}} - \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}} & -\overset{\bullet\circ}{\mathbf{C}} + \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\circ}{\mathbf{C}} \\ \mathbf{0} & \mathbf{0} & -\overset{\circ\circ}{\mathbf{C}} \end{bmatrix} \begin{bmatrix} \overset{\circ}{\mathbf{h}} \\ \overset{\bullet}{\mathbf{h}} \\ \overset{\circ}{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \overset{\circ}{\mathbf{f}} \end{bmatrix}. \tag{4.75}$$

The result follows from noticing that since no contribution of $\overset{\circ\circ}{\mathbf{C}}$ is assembled on $\partial\Omega$ and hence

$$-\overset{\circ\circ}{\mathbf{C}} = \sum_{\alpha,\beta=1}^{d} \left\langle \partial_\beta \overset{\circ}{\boldsymbol{\Phi}}, \partial_\alpha \overset{\circ}{\boldsymbol{\Phi}}^{\mathsf{T}} \right\rangle = \overset{\circ\circ}{\mathbf{S}}. \tag{4.76}$$

The argument for $d > 1$ and general constant operator $\boldsymbol{A}$ follows the same lines. We consider the block system

$$\begin{bmatrix} \mathring{\mathring{\mathbf{M}}} & \mathring{\dot{\mathbf{M}}} & \dots & \mathbf{0} & \mathbf{0} & -\mathring{\mathring{\mathbf{C}}}_{1,1} \\ \dot{\mathring{\mathbf{M}}} & \dot{\dot{\mathbf{M}}} & \dots & \mathbf{0} & \mathbf{0} & -\dot{\mathring{\mathbf{C}}}_{1,1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & \mathring{\mathring{\mathbf{M}}} & \mathring{\dot{\mathbf{M}}} & -\mathring{\mathring{\mathbf{C}}}_{d,d} \\ \mathbf{0} & \mathbf{0} & \dots & \dot{\mathring{\mathbf{M}}} & \dot{\dot{\mathbf{M}}} & -\dot{\mathring{\mathbf{C}}}_{d,d} \\ \mathring{\mathring{\mathbf{B}}}^{1,1} & \mathring{\dot{\mathbf{B}}}^{1,1} & \dots & \mathring{\mathring{\mathbf{B}}}^{d,d} & \mathring{\dot{\mathbf{B}}}^{d,d} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathring{\mathbf{h}}_{1,1} \\ \dot{\mathbf{h}}_{1,1} \\ \vdots \\ \mathring{\mathbf{h}}_{d,d} \\ \dot{\mathbf{h}}_{d,d} \\ \mathring{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{0} \\ \mathring{\mathbf{f}} \end{bmatrix} \tag{4.77}$$

and proceed by conducting block Gaussian elimination. Let us define

$$\mathring{\mathring{\mathbf{F}}}_{\alpha\beta} := \left(\mathring{\dot{\mathbf{B}}}^{\alpha,\beta} - \mathring{\mathring{\mathbf{B}}}^{\alpha,\beta}\mathring{\mathring{\mathbf{M}}}^{-1}\mathring{\dot{\mathbf{M}}}\right)\left(\dot{\dot{\mathbf{M}}} - \dot{\mathring{\mathbf{M}}}\mathring{\mathring{\mathbf{M}}}^{-1}\mathring{\dot{\mathbf{M}}}\right)^{-1}\left(\dot{\mathring{\mathbf{C}}}_{\alpha,\beta} - \dot{\mathring{\mathbf{M}}}\mathring{\mathring{\mathbf{M}}}^{-1}\mathring{\mathring{\mathbf{C}}}_{\alpha,\beta}\right), \tag{4.78}$$

then the resultant system is

$$
\begin{bmatrix}
\mathring{\mathring{\mathbf{M}}} & \mathring{\dot{\mathbf{M}}} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
\mathbf{0} & \ddot{\mathbf{M}} - \dot{\mathring{\mathbf{M}}}\mathring{\mathbf{M}}^{-1}\mathring{\dot{\mathbf{M}}} & \cdots & \mathbf{0} & \mathbf{0} & \mathbf{0} \\
\cdots & \cdots & \ddots & \cdots & \cdots & \cdots \\
\mathbf{0} & \mathbf{0} & \cdots & \mathring{\mathring{\mathbf{M}}} & \mathring{\dot{\mathbf{M}}} & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \ddot{\mathbf{M}} - \dot{\mathring{\mathbf{M}}}\mathring{\mathbf{M}}^{-1}\mathring{\dot{\mathbf{M}}} & \mathbf{0} \\
\mathbf{0} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & \sum_{\alpha=1}^{d}\sum_{\beta=1}^{d}\left[\mathring{\mathring{\mathbf{B}}}^{\alpha,\beta}\mathring{\mathring{\mathbf{M}}}^{-1}\mathring{\mathring{\mathbf{C}}}_{\alpha,\beta} + \mathring{\mathring{\mathbf{F}}}_{\alpha\beta}\right]
\end{bmatrix}
\begin{bmatrix}
-\mathring{\mathring{\mathbf{C}}}_{1,1} & & & \\
-\ddot{\mathring{\mathbf{C}}}_{1,1} + \dot{\mathring{\mathbf{M}}}\mathring{\mathring{\mathbf{M}}}^{-1}\mathring{\mathring{\mathbf{C}}}_{1,1} & & & \\
\vdots & & & \\
-\mathring{\mathring{\mathbf{C}}}_{d,d} & & & \\
-\ddot{\mathring{\mathbf{C}}}_{d,d} + \dot{\mathring{\mathbf{M}}}\mathring{\mathring{\mathbf{M}}}^{-1}\mathring{\mathring{\mathbf{C}}}_{d,d} & & & \\
\end{bmatrix}
\begin{bmatrix}
\mathring{\mathbf{h}}_{1,1} & \mathbf{0} \\
\dot{\mathbf{h}}_{1,1} & \mathbf{0} \\
\vdots & \cdots \\
\mathring{\mathbf{h}}_{d,d} & \mathbf{0} \\
\dot{\mathbf{h}}_{d,d} & \mathbf{0} \\
\mathring{\mathbf{u}} & \mathring{\mathbf{f}}
\end{bmatrix}
=
\begin{bmatrix}
\mathbf{0} & \mathbf{0} \\
\mathbf{0} & \mathbf{0} \\
\cdots & \cdots \\
\mathbf{0} & \mathbf{0} \\
\mathbf{0} & \mathbf{0} \\
\mathbf{0} & \mathring{\mathbf{f}}
\end{bmatrix}
\tag{4.79}
$$

Recall

$$\boldsymbol{A} = \left[a^{\alpha,\beta}\right]_{\alpha,\beta=1}^{d} \tag{4.80}$$

then for each $\alpha, \beta = 1, \ldots, d$ the matrices

$$\mathring{\overset{\bullet}{\mathbf{B}}}^{\alpha,\beta} = a^{\alpha,\beta}\mathring{\overset{\bullet}{\mathbf{M}}} \text{ and} \tag{4.81}$$

$$\mathring{\mathbf{B}}^{\alpha,\beta} = a^{\alpha,\beta}\mathring{\mathbf{M}}. \tag{4.82}$$

Thus for each $\alpha, \beta = 1, \ldots, d$ the matrices

$$\left(\mathring{\overset{\bullet}{\mathbf{B}}}^{\alpha,\beta} - \mathring{\mathbf{B}}^{\alpha,\beta}\mathring{\mathbf{M}}^{-1}\mathring{\overset{\bullet}{\mathbf{M}}}\right) = \mathbf{0} \tag{4.83}$$

and we see for each $\alpha, \beta = 1, \ldots, d$

$$\mathring{\mathbf{F}}_{\alpha\beta} = \mathbf{0}. \tag{4.84}$$

Now note

$$\begin{aligned}\sum_{\alpha=1}^{d}\sum_{\beta=1}^{d}\mathring{\mathbf{B}}^{\alpha,\beta}\mathring{\mathbf{M}}^{-1}\mathring{\mathbf{C}}_{\alpha,\beta} &= \sum_{\alpha=1}^{d}\sum_{\beta=1}^{d}a^{\alpha,\beta}\mathring{\mathbf{M}}\mathring{\mathbf{M}}^{-1}\mathring{\mathbf{C}}_{\alpha,\beta} \\ &= \sum_{\alpha=1}^{d}\sum_{\beta=1}^{d}a^{\alpha,\beta}\mathring{\mathbf{C}}_{\alpha,\beta} \\ &= \mathring{\mathbf{S}},\end{aligned} \tag{4.85}$$

which concludes the proof. □

We will now state some fundamental results which we will use to prove the invertibility of $\mathbf{E}$ for general elliptic problems.

4.3.0.6 Lemma (invertibility of block matrices and their Schur complements)**.** *Given a matrix* $\mathbf{P}$ *with the block structure*

$$\mathbf{P} = \begin{bmatrix}\mathbf{Q} & \mathbf{R} \\ \mathbf{S} & \mathbf{T}\end{bmatrix} \tag{4.86}$$

where $\mathbf{Q}$ *is nonsingular. Then* $\mathbf{P}$ *is invertible if and only if its Schur complement*

$$\mathbf{T} - \mathbf{S}\mathbf{Q}^{-1}\mathbf{R} \tag{4.87}$$

is invertible.

Proof Due to the nonsingular nature of $\mathbf{Q}$ it is sufficient to show that

$$\det \mathbf{P} = \det \left(\mathbf{T} - \mathbf{S}\mathbf{Q}^{-1}\mathbf{R} \right) \det \mathbf{Q}. \tag{4.88}$$

To show this we use block Gaussian elimination on $\mathbf{P}$. Left–multiplying the first block row by $\mathbf{Q}^{-1}$ gives

$$\begin{bmatrix} \mathbf{I} & \mathbf{Q}^{-1}\mathbf{R} \\ \mathbf{S} & \mathbf{T} \end{bmatrix}. \tag{4.89}$$

Left–multiplying the first block row by $\mathbf{S}$ and subtracting the result from the second block row gives

$$\begin{bmatrix} \mathbf{I} & \mathbf{Q}^{-1}\mathbf{R} \\ \mathbf{0} & \mathbf{T} - \mathbf{S}\mathbf{Q}^{-1}\mathbf{R} \end{bmatrix}. \tag{4.90}$$

Thus we see

$$\begin{bmatrix} \mathbf{Q}^{-1} & \mathbf{0} \\ -\mathbf{S}\mathbf{Q}^{-1} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{S} & \mathbf{T} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{Q}^{-1}\mathbf{R} \\ \mathbf{0} & \mathbf{T} - \mathbf{S}\mathbf{Q}^{-1}\mathbf{R} \end{bmatrix}. \tag{4.91}$$

Taking determinants it follows

$$\det \mathbf{Q}^{-1} \det \mathbf{P} = \det \left(\mathbf{T} - \mathbf{S}\mathbf{Q}^{-1}\mathbf{R} \right), \tag{4.92}$$

concluding the proof. □

4.3.0.7 Corollary (invertibility of the principal minor)**.** *Given a matrix* $\mathbf{P}$ *with the block structure*

$$\mathbf{P} = \begin{bmatrix} \mathbf{Q} & \mathbf{R} \\ \mathbf{S} & \mathbf{T} \end{bmatrix}. \tag{4.93}$$

If $\mathbf{P}$ *is invertible and the Schur complement of* $\mathbf{P}$ *,* $\mathbf{T} - \mathbf{S}\mathbf{Q}^{-1}\mathbf{R}$*, is also invertible, then* $\mathbf{Q}$ *is invertible.*

Proof The result follows from the proof of Lemma 4.3.0.6. □

4.3.0.8 Lemma (invertibility of $\ddot{\mathbf{M}} - \dot{\mathring{\mathbf{M}}}\ddot{\mathring{\mathbf{M}}}^{-1}\dot{\mathring{\mathbf{M}}}$)**.** *The term*

$$\ddot{\mathbf{M}} - \dot{\mathring{\mathbf{M}}}\ddot{\mathring{\mathbf{M}}}^{-1}\dot{\mathring{\mathbf{M}}} \tag{4.94}$$

appearing on the diagonal of (4.79) is positive definite and hence invertible.

Proof Note that

$$\overset{\bullet\bullet}{\mathbf{M}} - \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}} \tag{4.95}$$

is the Schur complement of the block matrix

$$\mathbf{M} = \begin{bmatrix} \overset{\circ\circ}{\mathbf{M}} & \overset{\circ\bullet}{\mathbf{M}} \\ \overset{\bullet\circ}{\mathbf{M}} & \overset{\bullet\bullet}{\mathbf{M}} \end{bmatrix}. \tag{4.96}$$

Recall the mass matrices $\mathbf{M}$ and $\overset{\circ\circ}{\mathbf{M}}$ are both Gram matrices (Definition A.1.0.10) and thus positive definite hence invertible. The result now follows immediately from Lemma 4.3.0.6. □

4.3.0.9 Corollary (invertibility of $\mathbf{E}$ for simple problems)**.** *If the problem coefficients in (4.18) are piecewise constant, the problem*

$$\boldsymbol{A}{:}\mathrm{D}^2 u = \operatorname{div} \boldsymbol{A}\nabla u \tag{4.97}$$

and the block matrix arising from discretising this problem, $\mathbf{E}$*, is invertible.*

Proof Upon reducing $\mathbf{E}$ is block echelon form, it is sufficient to show each component on the diagonal is invertible.

From (4.79) in the proof of Theorem 4.3.0.5 we see the only terms on the diagonal of the reduced matrix echelon($\mathbf{E}$) are

$$\overset{\circ\circ}{\mathbf{M}} \tag{4.98}$$

$$\overset{\bullet\bullet}{\mathbf{M}} - \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}} \text{ and} \tag{4.99}$$

$$\sum_{\alpha=1}^{d}\sum_{\beta=1}^{d}\left[\overset{\circ\circ}{\mathbf{B}}^{\alpha,\beta}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\circ}{\mathbf{C}}_{\alpha,\beta} + \overset{\circ\circ}{\mathbf{F}}_{\alpha\beta}\right]. \tag{4.100}$$

Recall from Theorem 4.3.0.5 for the class of problem we consider in this Corollary for each $\alpha, \beta = 1, \dots d$, $\overset{\circ\circ}{\mathbf{F}}_{\alpha\beta} = \mathbf{0}$ and $\overset{\circ\circ}{\mathbf{B}}^{\alpha,\beta}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\circ}{\mathbf{C}}_{\alpha,\beta} = \overset{\circ\circ}{\mathbf{S}}$. Both $\overset{\circ\circ}{\mathbf{M}}$ and $\overset{\circ\circ}{\mathbf{S}}$ are Grammian matrices and guaranteed invertible. The term $\overset{\bullet\bullet}{\mathbf{M}} - \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}}$ is also invertible from Lemma 4.3.0.8, concluding the proof. □

4.3.0.10 Remark (towards showing invertibility of $\mathbf{E}$)**.** Showing invertibility of $\mathbf{E}$ directly for general problems is not a trivial task. This Remark is aimed at demonstrating the difficulties that arise in the simple case for $d = 1$.

Recall from the Proof of Theorem 4.3.0.5 we split the block matrix

$$\mathbf{E} = \begin{bmatrix} \overset{\circ\circ}{\mathbf{M}} & \overset{\circ\bullet}{\mathbf{M}} & -\overset{\circ\circ}{\mathbf{C}} \\ \overset{\bullet\circ}{\mathbf{M}} & \overset{\bullet\bullet}{\mathbf{M}} & -\overset{\bullet\circ}{\mathbf{C}} \\ \overset{\circ\circ}{\mathbf{B}} & \overset{\circ\bullet}{\mathbf{B}} & \mathbf{0} \end{bmatrix}. \tag{4.101}$$

Applying block Gaussian elimination results in

$$\begin{bmatrix} \overset{\circ\circ}{\mathbf{M}} & \overset{\circ\bullet}{\mathbf{M}} & -\overset{\circ\circ}{\mathbf{C}} \\ \mathbf{0} & \overset{\bullet\bullet}{\mathbf{M}} - \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}} & -\overset{\bullet\circ}{\mathbf{C}} + \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\circ}{\mathbf{C}} \\ \mathbf{0} & \mathbf{0} & \overset{\circ\circ}{\mathbf{B}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\circ}{\mathbf{C}} + \\ & & \left(\overset{\circ\bullet}{\mathbf{B}} - \overset{\circ\circ}{\mathbf{B}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}}\right)\left(\overset{\bullet\bullet}{\mathbf{M}} - \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}}\right)^{-1}\left(\overset{\bullet\circ}{\mathbf{C}} - \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\circ}{\mathbf{C}}\right) \end{bmatrix}. \tag{4.102}$$

To show $\mathbf{E}$ is positive definite is equivalent to showing that each element on the diagonal of (4.102) is positive definite.

We have already shown that $\overset{\circ\circ}{\mathbf{M}}$ and $\overset{\bullet\bullet}{\mathbf{M}} - \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}}$ are positive definite in Lemma 4.3.0.8. It remains to show the term

$$\overset{\circ\circ}{\mathbf{B}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\circ}{\mathbf{C}} + \left(\overset{\circ\bullet}{\mathbf{B}} - \overset{\circ\circ}{\mathbf{B}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}}\right)\left(\overset{\bullet\bullet}{\mathbf{M}} - \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}}\right)^{-1}\left(\overset{\bullet\circ}{\mathbf{C}} - \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\circ}{\mathbf{C}}\right) \tag{4.103}$$

is positive definite. First we will study the major component $\overset{\circ\circ}{\mathbf{B}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\circ}{\mathbf{C}}$ and show $\overset{\circ\circ}{\mathbf{B}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\circ}{\mathbf{C}}$ is the product of positive definite matrices.

Due to the ellipticity of $\boldsymbol{A}$, there exists an $\alpha_0 > 0$ such that

$$\boldsymbol{\xi}^{\intercal}\boldsymbol{A}\boldsymbol{\xi} > \alpha_0 |\boldsymbol{\xi}|^2 \qquad \forall\, \boldsymbol{\xi} \in \mathbb{R}^d/\mathbf{0}. \tag{4.104}$$

From this the matrix $\overset{\circ\circ}{\mathbf{B}}$ is nothing but a perturbation of the mass matrix $\overset{\circ\circ}{\mathbf{M}}$. Recall that

$$\begin{aligned} \overset{\circ\circ}{\mathbf{B}}_{i,j} &= \int_\Omega \overset{\circ}{\Phi}_i(x)\boldsymbol{A}(x)\overset{\circ}{\Phi}_j(x)\,\mathrm{d}x \\ &> \int_\Omega \overset{\circ}{\Phi}_i(x)\alpha_0\overset{\circ}{\Phi}_j(x)\,\mathrm{d}x \\ &> \alpha_0 \int_\Omega \overset{\circ}{\Phi}_i(x)\overset{\circ}{\Phi}_j(x)\,\mathrm{d}x \\ &> \alpha_0 \overset{\circ\circ}{\mathbf{M}}_{i,j} \end{aligned} \tag{4.105}$$

and hence $\overset{\circ\circ}{\mathbf{B}}$ must be positive definite. The term $\overset{\circ\circ}{\mathbf{C}}$ coincides with the standard finite element stiffness matrix for the case $d = 1$ and hence is a Grammian and thus positive definite.

At the time of writing we are unable to prove any kind of invertibility on the term

$$\left(\overset{\circ\bullet}{\mathbf{B}} - \overset{\circ\circ}{\mathbf{B}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}}\right)\left(\overset{\bullet\bullet}{\mathbf{M}} - \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}}\right)^{-1}\left(\overset{\bullet\circ}{\mathbf{C}} - \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\circ}{\mathbf{C}}\right). \tag{4.106}$$

The difficulty comes from the fact both

$$\overset{\circ\bullet}{\mathbf{B}} - \overset{\circ\circ}{\mathbf{B}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}} \text{ and} \tag{4.107}$$

$$\overset{\bullet\circ}{\mathbf{C}} - \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\circ}{\mathbf{C}} \tag{4.108}$$

are rectangular. Although the matrix (4.106) does have some nice properties. Notice (4.106) is itself the Schur complement of the block matrix

$$\begin{bmatrix} \overset{\bullet\bullet}{\mathbf{M}} - \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}} & -\overset{\bullet\circ}{\mathbf{C}} + \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\circ}{\mathbf{C}} \\ \overset{\circ\bullet}{\mathbf{B}} - \overset{\circ\circ}{\mathbf{B}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}} & \mathbf{0} \end{bmatrix}. \tag{4.109}$$

Also note that the components are themselves Schur complements.

$$\overset{\circ\bullet}{\mathbf{B}} - \overset{\circ\circ}{\mathbf{B}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}} \text{ is the Schur complement of } \begin{bmatrix} \overset{\circ\circ}{\mathbf{M}} & \overset{\circ\bullet}{\mathbf{M}} \\ \overset{\circ\circ}{\mathbf{B}} & \overset{\circ\bullet}{\mathbf{B}} \end{bmatrix}, \tag{4.110}$$

$$\overset{\bullet\bullet}{\mathbf{M}} - \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\bullet}{\mathbf{M}} \text{ is the Schur complement of } \begin{bmatrix} \overset{\circ\circ}{\mathbf{M}} & \overset{\circ\bullet}{\mathbf{M}} \\ \overset{\bullet\circ}{\mathbf{M}} & \overset{\bullet\bullet}{\mathbf{M}} \end{bmatrix} \text{ and} \tag{4.111}$$

$$\overset{\bullet\circ}{\mathbf{C}} - \overset{\bullet\circ}{\mathbf{M}}\overset{\circ\circ}{\mathbf{M}}^{-1}\overset{\circ\circ}{\mathbf{C}} \text{ is the Schur complement of } \begin{bmatrix} \overset{\circ\circ}{\mathbf{M}} & \overset{\circ\circ}{\mathbf{C}} \\ \overset{\bullet\circ}{\mathbf{M}} & \overset{\bullet\circ}{\mathbf{C}} \end{bmatrix}. \tag{4.112}$$

4.3.0.11 Remark ("indirectly" showing invertibility of $\mathbf{E}$)**.** The rectangular nature of the components discussed in Remark 4.3.0.10 make showing $\mathbf{E}$ is invertible at best extremely difficult. We propose to circumvent this difficulty by making use of Corollary 4.3.0.7, expanding the matrix $\mathbf{E}$ such that each component becomes square and showing that this enlarged matrix is invertible.

4.3.0.12 Theorem (invertibility of $\tilde{\mathbf{D}}$)**.** *The matrix*

$$\tilde{\mathbf{D}} := \sum_{\alpha=1}^{d}\sum_{\beta=1}^{d} \mathbf{B}^{\alpha\beta}\mathbf{M}^{-1}\mathbf{C}_{\alpha\beta} \tag{4.113}$$

is positive definite and hence invertible.

Proof The case for $d = 1$ follows the same lines as that of Remark 4.3.0.10.

For $d > 1$ the proof is more delicate. We introduce three block matrices which take the form

$$\mathfrak{B} := \begin{bmatrix} \mathsf{B}^{11} & \dots & \mathsf{B}^{1d} \\ \vdots & \ddots & \vdots \\ \mathsf{B}^{d1} & \dots & \mathsf{B}^{dd} \end{bmatrix} \qquad \in \mathbb{R}^{dN \times dN} \tag{4.114}$$

$$\mathfrak{M} := \mathsf{M}^{-1} \boldsymbol{I} \qquad \in \mathbb{R}^{dN \times dN} \tag{4.115}$$

$$\mathfrak{C} := \begin{bmatrix} \mathsf{C}_{11} & \dots & \mathsf{C}_{d1} \\ \vdots & \ddots & \vdots \\ \mathsf{C}_{1d} & \dots & \mathsf{C}_{dd} \end{bmatrix} \qquad \in \mathbb{R}^{dN \times dN.} \tag{4.116}$$

Observe

$$\operatorname{trace}(\mathfrak{B}\mathfrak{M}\mathfrak{C}) = \sum_{\alpha=1}^{d} \sum_{\beta=1}^{d} \mathsf{B}^{\alpha\beta} \mathsf{M}^{-1} \mathsf{C}_{\alpha\beta}. \tag{4.117}$$

We will then proceed with the proof under the observation that if we can show $\mathfrak{B}$, $\mathfrak{M}$, $\mathfrak{C}$ are positive definite matrices then the $\operatorname{trace}(\mathfrak{B}\mathfrak{M}\mathfrak{C})$ must also be positive definite.

From Lemma 4.3.0.8 it is clear that M is a positive definite matrix. From this we know M^{-1} is also positive definite and hence $\mathfrak{M}$ is positive definite.

The matrix $\mathfrak{C}$ is positive definite since it forms a stiffness matrix.

The matrix $\mathfrak{B}$ is positive definite due to the ellipticity of $\boldsymbol{A}$. We will use i, j to denote the "numerical" components of $\mathfrak{B}$, that is, we define

$$\mathfrak{B}_{ij} = \begin{bmatrix} \mathsf{B}_{ij}^{11} & \dots & \mathsf{B}_{ij}^{1d} \\ \vdots & \ddots & \vdots \\ \mathsf{B}_{ij}^{d1} & \dots & \mathsf{B}_{ij}^{dd} \end{bmatrix}. \tag{4.118}$$

In this case we want to show there exists an $\alpha_0 > 0$ such that

$$\boldsymbol{\xi}^{\mathsf{T}} \mathfrak{B}_{ij} \boldsymbol{\xi} > \alpha_0 |\xi|^2 \quad \forall\, \boldsymbol{\xi} \in \mathbb{R}^d / \{\mathbf{0}\}, \quad \forall\, i, j = 1, \dots N. \tag{4.119}$$

Writing $\mathfrak{B}$ componentwise (geometrically)

$$\begin{aligned}\boldsymbol{\xi}^{\intercal}\mathfrak{B}_{ij}\boldsymbol{\xi} &= \begin{bmatrix}\xi_1 & \dots & \xi_d\end{bmatrix}\begin{bmatrix}\mathbf{B}_{ij}^{11} & \dots & \mathbf{B}_{ij}^{1d}\\ \vdots & \ddots & \vdots\\ \mathbf{B}_{ij}^{d1} & \dots & \mathbf{B}_{ij}^{dd}\end{bmatrix}\begin{bmatrix}\xi_1\\ \vdots\\ \xi_d\end{bmatrix}\\ &= \begin{bmatrix}\xi_1 & \dots & \xi_d\end{bmatrix}\begin{bmatrix}\int_\Omega \Phi_j(\boldsymbol{x})a^{11}(\boldsymbol{x})\Phi_i(\boldsymbol{x})\,\mathrm{d}x & \dots & \int_\Omega \Phi_j(\boldsymbol{x})a^{1d}(\boldsymbol{x})\Phi_i(\boldsymbol{x})\,\mathrm{d}x\\ \vdots & \ddots & \vdots\\ \int_\Omega \Phi_j(\boldsymbol{x})a^{d1}(\boldsymbol{x})\Phi_i(\boldsymbol{x})\,\mathrm{d}x & \dots & \int_\Omega \Phi_j(\boldsymbol{x})a^{dd}(\boldsymbol{x})\Phi_i(\boldsymbol{x})\,\mathrm{d}x\end{bmatrix}\begin{bmatrix}\xi_1\\ \vdots\\ \xi_d\end{bmatrix},\end{aligned} \tag{4.120}$$

by the definition of the component matrices $\mathbf{B}^{\alpha\beta}$. Now

$$\begin{aligned}\boldsymbol{\xi}^{\intercal}\mathfrak{B}_{ij}\boldsymbol{\xi} &= \begin{bmatrix}\xi_1 & \dots & \xi_d\end{bmatrix}\int_\Omega \Phi_j(\boldsymbol{x})\begin{bmatrix}a^{11} & \dots & a^{1d}\\ \vdots & \ddots & \vdots\\ a^{d1} & \dots & a^{dd}\end{bmatrix}\Phi_i(\boldsymbol{x})\,\mathrm{d}x\begin{bmatrix}\xi_1\\ \vdots\\ \xi_d\end{bmatrix}\\ &= \int_\Omega \Phi_j(\boldsymbol{x})\begin{bmatrix}\xi_1 & \dots & \xi_d\end{bmatrix}\begin{bmatrix}a^{11} & \dots & a^{1d}\\ \vdots & \ddots & \vdots\\ a^{d1} & \dots & a^{dd}\end{bmatrix}\begin{bmatrix}\xi_1\\ \vdots\\ \xi_d\end{bmatrix}\Phi_i(\boldsymbol{x})\,\mathrm{d}x\\ &> \int_\Omega \Phi_j(\boldsymbol{x})\alpha_0\left|\xi\right|^2\Phi_i(\boldsymbol{x})\,\mathrm{d}x\\ &> \alpha_0\left|\xi\right|^2\mathbf{M}_{i,j},\end{aligned} \tag{4.121}$$

by the ellipticity of $\boldsymbol{A}$ and definition of the mass matrix $\mathbf{M}$. Positivity is guaranteed by the positivity of the mass matrix. □

4.3.0.13 Corollary (invertibility of $\tilde{\mathbf{E}}$)**.** *The block matrix*

$$\tilde{\mathbf{E}} = \begin{bmatrix}\mathbf{M} & \mathbf{0} & \dots & \mathbf{0} & -\mathbf{C}_{11}\\ \mathbf{0} & \mathbf{M} & \dots & \mathbf{0} & -\mathbf{C}_{12}\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ \mathbf{0} & \mathbf{0} & \dots & \mathbf{M} & -\mathbf{C}_{dd}\\ \mathbf{B}^{11} & \mathbf{B}^{12} & \dots & \mathbf{B}^{dd} & \mathbf{0}\end{bmatrix} \tag{4.122}$$

is invertible.

Proof The proof is the same calculations as in Lemma 4.2.1.1 to show the system

$$\tilde{\mathbf{D}}\mathbf{u} = \mathbf{f} \tag{4.123}$$

is equivalent to the block system

$$\tilde{\mathbf{E}}\tilde{\mathbf{v}} = \tilde{\mathbf{b}} \tag{4.124}$$

and hence if $\tilde{\mathbf{D}}$ is invertible, $\tilde{\mathbf{E}}$, must also be. □

4.3.0.14 Corollary (invertibility of $\mathbf{E}$)**.** *Combining the results of Corollary 4.3.0.13 and Corollary 4.3.0.7 we see the matrix* $\mathbf{E}$ *is always invertible assuring the NVFE solution is well posed.*

Proof The extended block matrix is defined as

$$\tilde{\mathbf{E}} = \begin{bmatrix} \mathbf{E} & \dot{\mathbf{E}} \\ \dot{\mathbf{E}} & \mathbf{0} \end{bmatrix} \tag{4.125}$$

where

$$\dot{\mathbf{E}} = \left[\dot{\mathbf{B}}^{11}, \ldots, \dot{\mathbf{B}}^{dd}, \mathbf{0} \right]^{\mathsf{T}} \tag{4.126}$$

$$\dot{\mathbf{E}} = \left[-\dot{\mathbf{C}}_{11}, \ldots, -\dot{\mathbf{C}}_{dd}, \mathbf{0} \right]^{\mathsf{T}}. \tag{4.127}$$

Hence we may apply Corollary 4.3.0.7 noticing $\tilde{\mathbf{E}}$ is invertible from Corollary 4.3.0.13 and $\tilde{\mathbf{D}}$, the Schur complement of $\tilde{\mathbf{E}}$ is also invertible. □

4.3.0.15 Remark (condition number)**.** The convergence rate of an iterative solver applied to a linear system $\mathbf{N}\mathbf{v} = \mathbf{g}$ will depend on the condition number $\kappa(\mathbf{N})$, defined as the ratio of the maximum and minimum eigenvalues of $\mathbf{N}$:

$$\kappa(\mathbf{N}) := \frac{\lambda_{\max}}{\lambda_{\min}}. \tag{4.128}$$

Numerically we observe the condition number of the block matrix $\kappa(\mathbf{E}) \leq Ch^{-2}$ (see Table 4.1).

4.4 Inhomogeneous Dirichlet boundary values

Given additional problem data $g \in \mathrm{C}^0(\Omega)$, to solve

$$\begin{aligned} \mathscr{L}u &= f \text{ in } \Omega, \\ u &= g \text{ on } \partial\Omega, \end{aligned} \tag{4.129}$$

it is not immediate how to enforce the boundary conditions.

We have derived two approaches to tackle the implementation of inhomogeneous boundary values.

The first is a direct extension of the method presented in §4.2 for homogeneous boundary conditions. In this case we directly enforce the boundary values into the system matrix. This method is very practical, in fact many finite element codes enforce boundary data in an analogous way [SS05].

The second approach is a more natural method, it involves adding singular blocks onto the system matrix to make each component square analogously to the method used to prove invertibility in the previous section. In this case boundary conditions are enforced in a much weaker sense. When the mesh is under resolved we see very mild "oscillations" over the boundary, as illustrated in Figure 4.1. As the mesh is refined though note these "oscillations" dissipate. This is very reminiscent of Nitsche's method.

4.4.1 Method 1 - directly enforcing boundary conditions into the system matrix

As before we will illustrate the method with an example for $d = 1$ and after present the case for general dimension. We split the matrices as follows

$$\mathbf{E}\mathbf{v} = \mathbf{b}$$

$$\downarrow$$

$$\begin{bmatrix} \mathring{\mathring{\mathbf{M}}} & \mathring{\dot{\mathbf{M}}} & -\mathring{\mathring{\mathbf{C}}} \\ \dot{\mathring{\mathbf{M}}} & \dot{\dot{\mathbf{M}}} & -\dot{\mathring{\mathbf{C}}} \\ \mathring{\mathring{\mathbf{B}}} & \mathring{\dot{\mathbf{B}}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathring{\mathbf{h}} \\ \dot{\mathbf{h}} \\ \mathring{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathring{\mathbf{f}} \end{bmatrix}.$$

To enforce the boundary conditions we must add one extra row and column into the system to account for the values of $\dot{\mathbf{u}}$ as follows

$$\begin{bmatrix} \mathring{\mathring{\mathbf{M}}} & \mathring{\dot{\mathbf{M}}} & -\mathring{\mathring{\mathbf{C}}} & -\mathring{\dot{\mathbf{C}}} \\ \dot{\mathring{\mathbf{M}}} & \dot{\dot{\mathbf{M}}} & -\dot{\mathring{\mathbf{C}}} & -\dot{\dot{\mathbf{C}}} \\ \mathring{\mathring{\mathbf{B}}} & \mathring{\dot{\mathbf{B}}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathring{\mathbf{h}} \\ \dot{\mathbf{h}} \\ \mathring{\mathbf{u}} \\ \dot{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathring{\mathbf{f}} \\ \dot{\mathbf{g}} \end{bmatrix}. \tag{4.130}$$

We may then eliminate the bottom row and reduce the system back to one of the same complexity as the original $\mathbf{E}$ as follows

$$\begin{bmatrix} \mathring{\mathring{\mathbf{M}}} & \mathring{\dot{\mathbf{M}}} & -\mathring{\mathring{\mathbf{C}}} \\ \dot{\mathring{\mathbf{M}}} & \dot{\dot{\mathbf{M}}} & -\dot{\mathring{\mathbf{C}}} \\ \mathring{\mathring{\mathbf{B}}} & \mathring{\dot{\mathbf{B}}} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathring{\mathbf{h}} \\ \dot{\mathbf{h}} \\ \mathring{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathring{\dot{\mathbf{C}}}\dot{\mathbf{g}} \\ \dot{\dot{\mathbf{C}}}\dot{\mathbf{g}} \\ \mathring{\mathbf{f}} \end{bmatrix}, \tag{4.131}$$

setting $\dot{\mathbf{u}} = \dot{\mathbf{g}}$ upon solution of the system.

In general dimension the linear system resulting from discretising the inhomogeneous Dirichlet problem (4.129) under this methodology of boundary enforcement would be

$$\begin{bmatrix} \mathbf{E} & \dot{\mathbf{E}} \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{v} \\ \dot{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \dot{\mathbf{g}} \end{bmatrix} \tag{4.132}$$

where $\mathbf{E}, \mathbf{v}$ and $\mathbf{b}$ are defined in Theorem 4.1.3.5. Let $\dot{\mathbf{\Phi}} = \left(\dot{\Phi}_1, \ldots, \dot{\Phi}_{\dot{N}}\right)^{\mathsf{T}}$ then $\dot{\mathbf{g}}, \dot{\mathbf{E}}$ and its components are defined as follows

$$\dot{\mathbf{E}} = \left[-\dot{\mathbf{C}}_{1,1}, \quad -\dot{\mathbf{C}}_{1,2}, \quad \ldots, \quad -\dot{\mathbf{C}}_{d,d-1}, \quad -\dot{\mathbf{C}}_{d,d}, \quad \mathbf{0}\right]^{\mathsf{T}},$$

$$\dot{\mathbf{C}}_{\alpha,\beta} = -\left\langle \partial_\beta \mathbf{\Phi}, \partial_\alpha \dot{\mathbf{\Phi}}^{\mathsf{T}} \right\rangle + \left\langle \mathbf{\Phi} \boldsymbol{n}_\beta, \partial_\alpha \dot{\mathbf{\Phi}}^{\mathsf{T}} \right\rangle_{\partial\Omega} \in \mathbb{R}^{N \times \dot{N}},$$
$$\dot{\mathbf{g}}_j = g(x_j)\dot{\Phi}_j \in \mathbb{R}^{\dot{N}},$$

where x_j is the Lagrange node associated with $\dot{\Phi}_j$.

Note the block matrix (4.132) can then be solved for $\mathbf{v}$ as follows

$$\mathbf{E}\mathbf{v} = \mathbf{b} - \dot{\mathbf{E}}\dot{\mathbf{g}}, \tag{4.133}$$

and then setting $\dot{\mathbf{u}} = \dot{\mathbf{g}}$.

4.4.2 Method 2 - natural enforcement of boundary conditions

In this method we are essentially adding additional data into the problem. Recall from Remark 4.2.1.2 that the discrete system is:

Find $U \in \mathring{\mathbb{V}}$ and $\boldsymbol{H}[U] \in \mathbb{V}^{d\times d}$ such that

$$\begin{cases} \langle \boldsymbol{H}[U], \Phi \rangle = -\langle \nabla U \otimes \nabla \Phi \rangle + \langle \nabla U \otimes \boldsymbol{n}\ \Phi \rangle_{\partial\Omega} & \forall\, \Phi \in \mathbb{V} \\ \left\langle \boldsymbol{A} : \boldsymbol{H}[U], \mathring{\Phi} \right\rangle = \left\langle f, \mathring{\Phi} \right\rangle & \forall\, \mathring{\Phi} \in \mathring{\mathbb{V}}. \end{cases} \tag{4.134}$$

In this method we enlarge the space of test functions to $\mathbb{V}$, i.e., Find $U \in \mathbb{V}$ and $\boldsymbol{H}[U] \in \mathbb{V}^{d\times d}$ such that

$$\begin{cases} \langle \boldsymbol{H}[U], \Phi\rangle = -\langle \nabla U \otimes \nabla\Phi\rangle + \langle \nabla U \otimes \boldsymbol{n}\ \Phi\rangle_{\partial\Omega} \quad \forall\, \Phi \in \mathbb{V} \\ \\ \langle \boldsymbol{A}{:}\boldsymbol{H}[U], \Phi\rangle = \langle f, \Phi\rangle \quad \forall\, \Phi \in \mathbb{V}. \end{cases} \tag{4.135}$$

We again demonstrate the method with $d = 1$ for clarity. If we proceed to discretise (4.135) using the methodology set out in §4.1.3 the result is a linear algebra problem of the following form: find $\mathbf{u} \in \mathbb{R}^N$ such that

$$\tilde{\mathbf{D}}\mathbf{u} := \mathbf{B}\mathbf{M}^{-1}\mathbf{C}\mathbf{u} = \mathbf{f}. \tag{4.136}$$

Or in its now familiar block structure

$$\tilde{\mathbf{E}}\tilde{\mathbf{v}} = \tilde{\mathbf{b}} \tag{4.137}$$

$$\begin{bmatrix} \mathring{\mathring{\mathbf{M}}} & \mathring{\dot{\mathbf{M}}} & -\mathring{\mathring{\mathbf{C}}} & -\mathring{\dot{\mathbf{C}}} \\ \dot{\mathring{\mathbf{M}}} & \dot{\dot{\mathbf{M}}} & -\dot{\mathring{\mathbf{C}}} & -\dot{\dot{\mathbf{C}}} \\ \mathring{\mathring{\mathbf{B}}} & \mathring{\dot{\mathbf{B}}} & \mathbf{0} & \mathbf{0} \\ \dot{\mathring{\mathbf{B}}} & \dot{\dot{\mathbf{B}}} & \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathring{\mathbf{h}} \\ \dot{\mathbf{h}} \\ \mathring{\mathbf{u}} \\ \dot{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathring{\mathbf{f}} \\ \dot{\mathbf{f}} + \dot{\mathbf{g}} \end{bmatrix}. \tag{4.138}$$

In general dimension the linear system resulting from discretising the inhomogeneous Dirichlet problem (4.129) under this methodology would be

$$\begin{bmatrix} \mathbf{E} & \dot{\mathbf{E}} \\ \dot{\mathbf{E}} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{v} \\ \dot{\mathbf{u}} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \dot{\mathbf{f}} + \dot{\mathbf{g}} \end{bmatrix}, \tag{4.139}$$

where

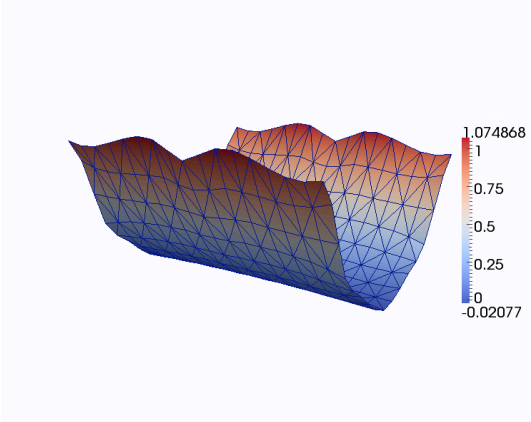
$$\dot{\mathbf{E}} = \left[\dot{\mathbf{B}}^{11}, \ldots, \dot{\mathbf{B}}^{dd}, \mathbf{0}\right]^{\mathsf{T}} \tag{4.140}$$

$$\dot{\mathbf{E}} = \left[-\dot{\mathbf{C}}_{11}, \ldots, -\dot{\mathbf{C}}_{dd}, \mathbf{0}\right]^{\mathsf{T}}. \tag{4.141}$$

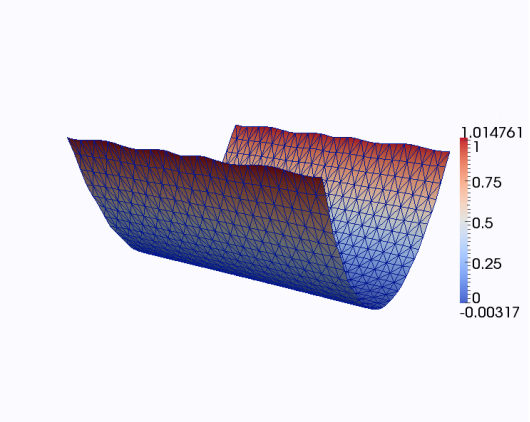
4.4.2.1 Remark (comparison of the two methods of boundary enforcement)**.** It can be shown that Method 1 and Method 2 coincide in the case that $\boldsymbol{A}$ is piecewise constant [1] although this is not true in general.

[1] This fact and its proof are omitted here, due to their length.

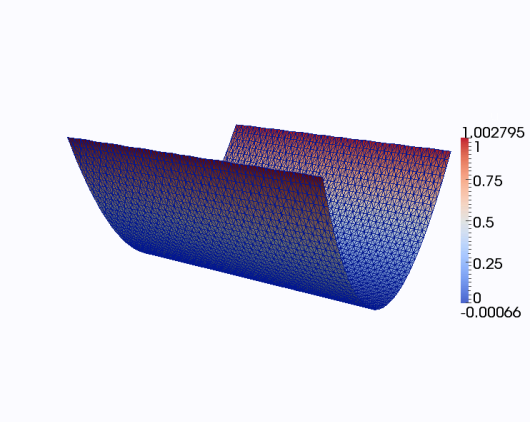
Figure 4.1: Three successive uniform refinements of the nonvariational finite element approximation to the PDE given in (4.146). The problem data is chosen such that the solution of this PDE is given by $u(\boldsymbol{x}) = x_1^2$. We are using the method given in §4.4.2 to enforce the boundary conditions. Notice the boundary conditions are not enforced exactly.



(a) Iterate 1, $\dim \mathbb{V} = 289$.



(b) Iterate 2, $\dim \mathbb{V} = 1089$.



(c) Iterate 3, $\dim \mathbb{V} = 4225$.

Computationally it is clear that Method 2 has a higher complexity than Method 1 since the resultant matrix is larger. Notice here however that it is not guaranteed that $\mathring{\mathbf{u}} = \mathring{\mathbf{g}}$, this is observed numerically (Figure 4.1). Both methods yield an optimal convergence however.

4.5 Numerical applications

In this section we study the numerical behavior of the scheme presented in Theorem 4.1.3.5. All our computations were carried out in Matlab ® (code available on request).

We present a set of linear benchmark problems, for which the solution is known. We take Ω to be the square $(-1,1)^2 \subset \mathbb{R}^2$ and in tests 4.5.1 and 4.5.2 consider the operator

$$\boldsymbol{A}(\boldsymbol{x}) = \begin{bmatrix} 1 & b(\boldsymbol{x}) \\ b(\boldsymbol{x}) & a(\boldsymbol{x}) \end{bmatrix} \tag{4.142}$$

varying the coefficients $a(\boldsymbol{x})$ and $b(\boldsymbol{x})$.

4.5.1 Test problem with a nondifferentiable operator

For the first test problem we choose the operator in such a way that (1.2) does not hold in the classical sense, that is, the components of $\boldsymbol{A}$ are nondifferentiable on Ω. Namely we take

$$a(\boldsymbol{x}) = (x_1^2 x_2^2)^{1/3} + 1 \tag{4.143}$$

$$b(\boldsymbol{x}) = 0. \tag{4.144}$$

A visualisation of the coefficient $a(\boldsymbol{x})$ is given in Figure 4.2. We choose the problem's source term f such that the exact solution to Problem 4.5.1 is given by:

$$u(\boldsymbol{x}) = \exp(-10\,|\boldsymbol{x}|^2). \tag{4.145}$$

We discretise the problem given by (4.143) under the algorithm set out in §4.1.3. Experimental convergence rates are shown in Figure 4.4.

4.5.2 Test problem with convection dominated operator

The second test problem demonstrates the ability to overcome oscillations introduced into the standard finite element when rewriting the operator in divergence form. Take

$$a(\boldsymbol{x}) = \arctan\Big(k(|\boldsymbol{x}|^2 - 1)\Big) + 2 \tag{4.146}$$

$$b(\boldsymbol{x}) = 0 \tag{4.147}$$

with $k \in \mathbb{R}^+$. Rewriting in divergence form gives

$$\boldsymbol{A}{:}\mathrm{D}^2 u = \operatorname{div} \boldsymbol{A}\nabla u - \operatorname{div} \boldsymbol{A}\nabla u, \tag{4.148}$$

where we are using

$$\operatorname{div} \boldsymbol{A} := \begin{bmatrix} \partial_1 & \dots & \partial_d \end{bmatrix} \begin{bmatrix} a^{1,1}(\boldsymbol{x}) & \dots & a^{1,d}(\boldsymbol{x}) \\ \vdots & \ddots & \vdots \\ a^{d,1}(\boldsymbol{x}) & \dots & a^{d,d}(\boldsymbol{x}) \end{bmatrix}. \tag{4.149}$$

The derivatives

$$\partial_\alpha a(\boldsymbol{x}) = \frac{dkx_\alpha}{1 + k\left(|\boldsymbol{x}|^2 - 1\right)} \tag{4.150}$$

can be made arbitrarily large on the unit circle by choosing k appropriately (see Figure 4.2). We choose our problem's source term f such that the exact solution to the problem is given by:

$$u(\boldsymbol{x}) = \sin(\pi x_1)\sin(\pi x_2). \tag{4.151}$$

We then construct the standard finite element method around (4.148), that is find $U \in \mathring{\mathbb{V}}$ such that for each $\mathring{\Phi} \in \mathring{\mathbb{V}}$

$$\left\langle \boldsymbol{A}\nabla U, \nabla\mathring{\Phi} \right\rangle - \left\langle \operatorname{div} \boldsymbol{A}\nabla U, \mathring{\Phi} \right\rangle = \left\langle f, \mathring{\Phi} \right\rangle. \tag{4.152}$$

If k is chosen small enough the standard finite element method converges optimally. If we increase the value of k oscillations become apparent in the finite element solution along the unit circle. Figure 4.6 demonstrates the oscillations arising from this method compared to discretising using the nonvariational finite element method.

Figure 4.5 shows the numerical convergence rates of the nonvariational finite element method applied to this problem.

4.5.3 Test problem choosing a solution with nonsymmetric Hessian

In this test we choose the operator such that $b(\boldsymbol{x})$ is nonzero. To maintain ellipticity in this problem we must choose $a(\boldsymbol{x})$ such that the trace of $\boldsymbol{A}$ dominates its determinant. We choose

$$a(\boldsymbol{x}) = 2 \tag{4.153}$$

$$b(\boldsymbol{x}) = (x_1^2 x_2^2)^{1/3}. \tag{4.154}$$

We choose the problem data such that the exact solution is given by

$$u(\boldsymbol{x}) = \begin{cases} \frac{x_1 x_2 (x_1^2 - x_2^2)}{x_1^2 + x_2^2} & \boldsymbol{x} \neq \mathbf{0} \\ 0 & \boldsymbol{x} = \mathbf{0}. \end{cases} \tag{4.155}$$

This function has a nonsymmetric Hessian at the point $\mathbf{0}$. The nontrivial Dirichlet boundary is dealt with using the direct method described in §4.4.1 [2]. Figure 4.8 shows numerical results for this problem.

4.5.4 Test problem for an irregular solution

In this test we choose the operator

$$a(\boldsymbol{x}) = \sin\left(\frac{1}{|x_1| + |x_2| + 10^{-15}}\right) \tag{4.156}$$

$$b(\boldsymbol{x}) = 0. \tag{4.157}$$

Notice the operator oscillates heavily near $\mathbf{0}$. Figure 4.3 shows a surface plot of the operator (4.156) and a cross section through $x_1 = 0$ illustrating the oscillations near the origin.

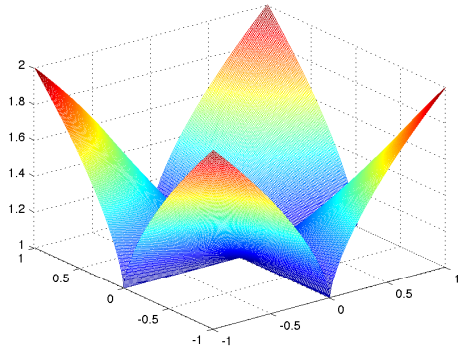
We choose the problem data such that the solution is given by

$$u(\boldsymbol{x}) = -\sqrt{2 - x_1^2 - x_2^2}. \tag{4.158}$$

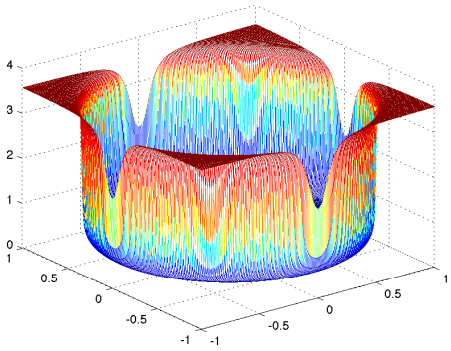
The solution is singular on the corners of Ω and the convergence rates reflect that as can be seen in Figure 4.9.

[2] The natural method described in §4.4.2 yields the same results qualitatively

Figure 4.2: A visualisation of the coefficient of the operators (4.143) (on the left) and (4.146) (on the right).

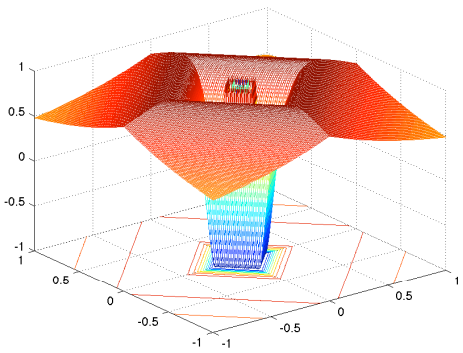


(a) The function $(x_1^2 x_2^2)^{1/3} + 1$ over Ω. Note the derivatives are singular at $x_1 = 0$ and $x_2 = 0$.

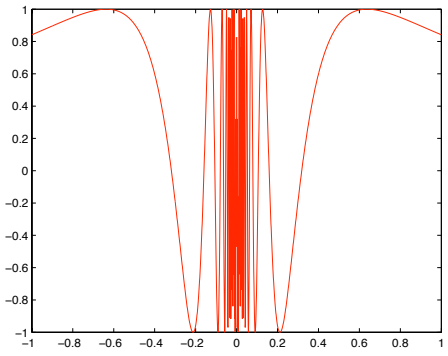


(b) The function $\arctan\left(5000(|x|^2 - 1)\right)$ over Ω. Note the derivatives are very large on the unit circle.

Figure 4.3: A visualisation of the coefficient of the operator from Problem (4.156) and a cross section through the coordinate axis.

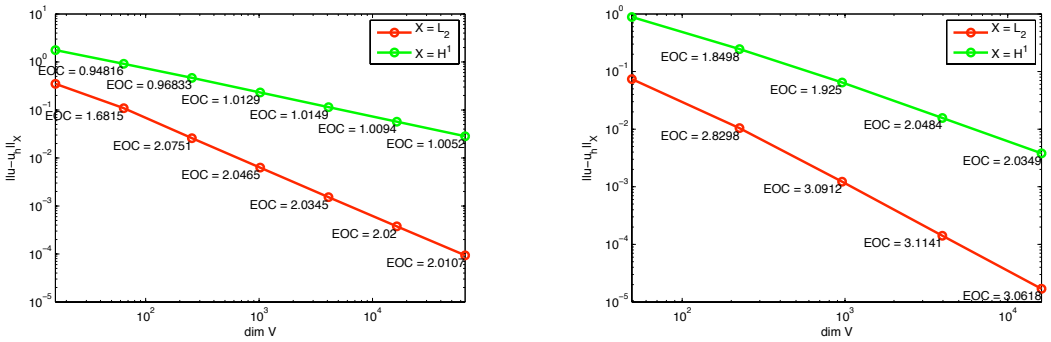


(a) The function $\sin\left(\frac{1}{|x_1|+|x_2|+10^{-15}}\right)$ over Ω. Note the function oscillates heavily near $\mathbf{0}$.



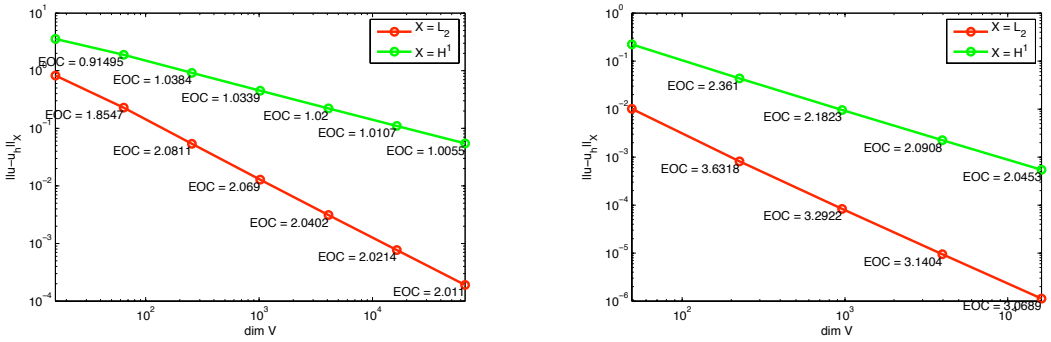
(b) A cross section through the first coordinate axis demonstrating the oscillations.

Figure 4.4: Problem 4.5.1. $\mathrm{L}_2(\Omega)$ and $\mathrm{H}^1(\Omega)$ errors and convergence rates for the NVFEM applied to a nondivergence form operator (4.143), choosing f appropriately such that $u(\boldsymbol{x}) = \exp(-10\,|\boldsymbol{x}|)$. The convergence rates are optimal, that is for $\mathbb{P}^1$-elements (on the left) $\|u - U\| = \mathrm{O}(h^2)$ and $|u - U|_1 = \mathrm{O}(h)$. For $\mathbb{P}^2$-elements (on the right) $\|u - U\| = \mathrm{O}(h^3)$ and $|u - U|_1 = \mathrm{O}(h^2)$.



(a) $\mathbb{P}^1$-elements (b) $\mathbb{P}^2$-elements

Figure 4.5: Problem 4.5.2. $\mathrm{L}_2(\Omega)$ and $\mathrm{H}^1(\Omega)$ errors and convergence rates for the NVFEM applied to a nondivergence form operator (4.146) with $k = 5000$, choosing f appropriately such that $u(\boldsymbol{x}) = \sin(\pi x_1)\sin(\pi x_2)$. The convergence rates are optimal, that is for $\mathbb{P}^1$-elements (on the left) $\|u - U\| = \mathrm{O}(h^2)$ and $|u - U|_1 = \mathrm{O}(h)$. For $\mathbb{P}^2$-elements (on the right) $\|u - U\| = \mathrm{O}(h^3)$ and $|u - U|_1 = \mathrm{O}(h^2)$.



(a) $\mathbb{P}^1$-elements (b) $\mathbb{P}^2$-elements

Figure 4.6: Test 4.5.2. On the left we present $\left\|u-\tilde{U}\right\|_{\mathrm{L}_\infty(K)}$ plotted as a function over Ω. This represents the maximum error of the standard FE solution, $\tilde{U}$, to problem (4.146) with 16384 DOFs. Notice the oscillations apparent on the unit circle centered at the origin. On the right we show $\|u-U\|_{\mathrm{L}_\infty(K)}$ plotted as a function over Ω, the maximum error of the NVFE solution, U, to problem (4.146) with 16384 DOFs ($h = 1/32$).

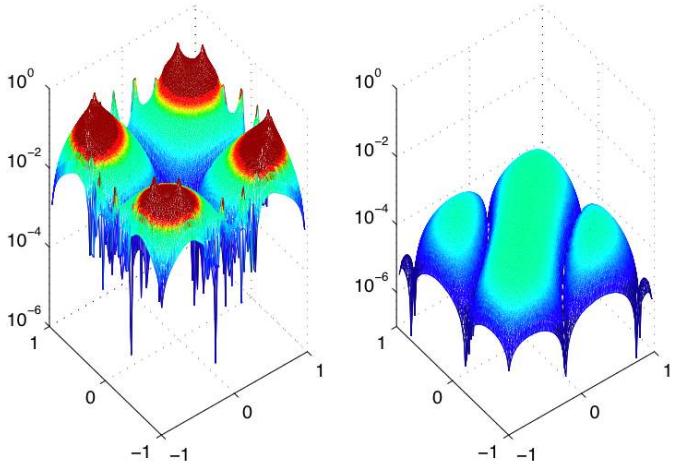


Figure 4.7: Test 4.5.2. On top we realise the FE solution on the unit circle centered at the origin as a 1 dimensional function of θ. Below we show the $\mathrm{L}_\infty(\Omega)$ error over the same domain.

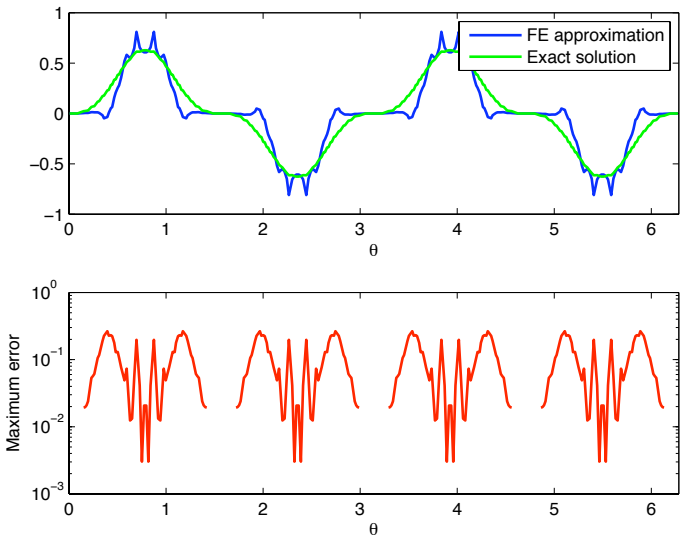
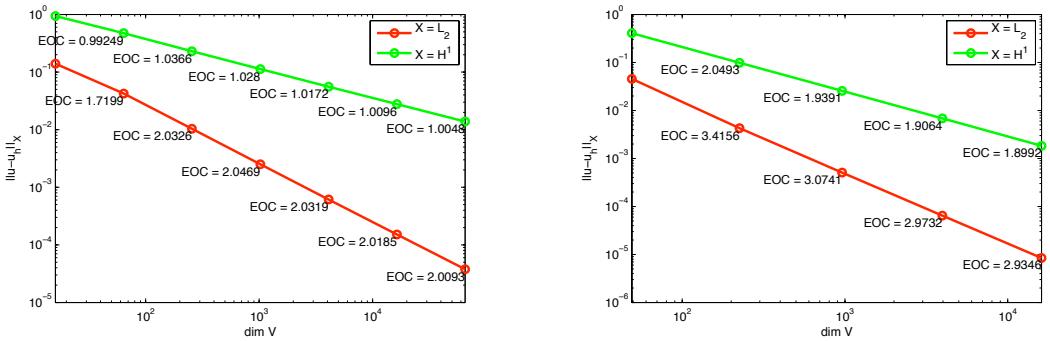
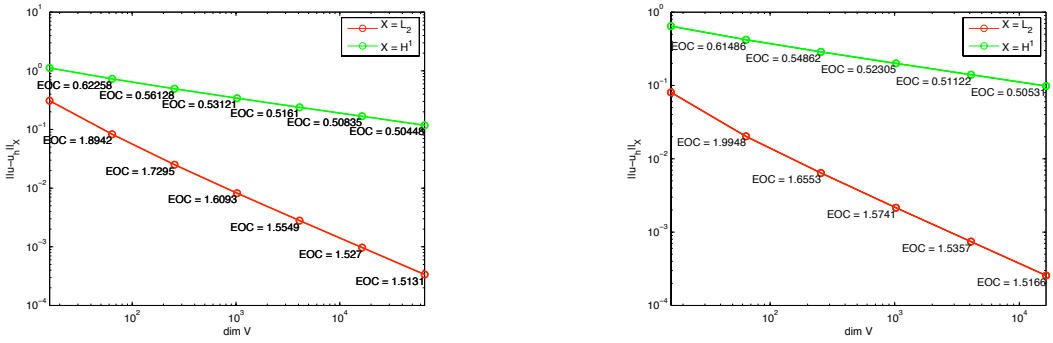


Figure 4.8: Problem 4.5.3. $\mathrm{L}_2(\Omega)$ and $\mathrm{H}^1(\Omega)$ errors and convergence rates for the NVFEM on an operator (4.153), choosing f appropriately such that $u(\boldsymbol{x}) = \frac{x_1 x_2 (x_1^2 - x_2^2)}{x_1^2 + x_2^2}$ if $\boldsymbol{x} \neq \mathbf{0}$, or $u(\boldsymbol{x}) = 0$ otherwise. The convergence rates are optimal, that is for $\mathbb{P}^1$-elements (on the left) $\|u - U\| = \mathrm{O}(h^2)$ and $|u - U|_1 = \mathrm{O}(h)$. For $\mathbb{P}^2$-elements (on the right) $\|u - U\| = \mathrm{O}(h^3)$ and $|u - U|_1 = \mathrm{O}(h^2)$.



(a) $\mathbb{P}^1$-elements (b) $\mathbb{P}^2$-elements

Figure 4.9: Problem 4.5.4. $\mathrm{L}_2(\Omega)$ and $\mathrm{H}^1(\Omega)$ errors and convergence rates for the NVFEM on an operator (4.156), choosing f appropriately such that $u(\boldsymbol{x}) = -\sqrt{2 - x_1^2 - x_2^2}$. The convergence rates are suboptimal due to the singular nature of the solution near the corners of $\Omega = (-1, 1)^2$, that is for both $\mathbb{P}^1$ (on the left) and $\mathbb{P}^2$-elements (on the right) $\|u - U\| = \mathrm{O}(h^{1.5})$ and $|u - U|_1 = \mathrm{O}(h^{0.5})$.



(a) $\mathbb{P}^1$-elements (b) $\mathbb{P}^2$-elements

Table 4.1: Problem 4.5.1. On the condition number of $\mathbf{E}$ upon discretising problem (4.143) using $\mathbb{P}^1$ finite elements. As claimed in Remark 4.3.0.15 $\kappa(\mathbf{E}) \approx Ch^{-2}$.

$\dim \mathbb{V}$	h	$\kappa(\mathbf{E})$	$h^2\kappa(\mathbf{E})$
16	0.4714	4.904×10^1	10.898
64	0.202	6.594×10^2	26.952
256	0.0943	3.665×10^3	32.633
1024	0.0456	1.722×10^4	35.833
4096	0.0224	6.894×10^4	34.737
16384	0.0111	3.383×10^5	41.949
65536	0.0055	1.337×10^6	40.43

4.6 Error analysis

In this section we analyse the method derived in §4.1.3 in both a apriori and aposteriori sense. We then propose an adaptive algorithm based on the estimator, conducting numerous numerical experiments in the process.

4.6.0.1 Lemma (Galërkin orthogonality)**.** *Let $u \in \mathrm{H}^2(\Omega) \cap \mathrm{H}^1_0(\Omega)$ be the solution of the continuous problem (4.1). Suppose also that $U \in \mathring{\mathbb{V}}$ is the solution of the discrete problem (4.21). Then*

$$\left\langle \boldsymbol{A}{:}(\mathrm{D}^2 u - \boldsymbol{H}[U]), \mathring{\Phi} \right\rangle = 0 \quad \forall\, \mathring{\Phi} \in \mathring{\mathbb{V}}. \tag{4.159}$$

Equivalently due to the definition of $\boldsymbol{H}[U]$ (see Definition 4.1.3.4)

$$\left\langle \boldsymbol{A}{:}(\mathrm{D}^2 u - \mathrm{D}^2 U) \,|\, \mathring{\Phi} \right\rangle = 0 \quad \forall\, \mathring{\Phi} \in \mathring{\mathbb{V}}. \tag{4.160}$$

Proof It suffices to note that equation (4.18) holds for each $\Phi \in \mathring{\mathbb{V}} \subset \mathrm{H}^1_0(\Omega)$. Taking the difference between (4.21) and (4.18) restricted to the subspace $\mathring{\mathbb{V}}$ gives the desired result. □

4.6.0.2 Lemma (quasioptimality)**.** *Let u and U be defined as in Theorem 4.6.0.1 then*

$$\left\| \boldsymbol{A}{:}\big(\mathrm{D}^2 u - \boldsymbol{H}[U]\big) \right\| = \min_{V \in \mathring{\mathbb{V}}} \left\| \boldsymbol{A}{:}\big(\mathrm{D}^2 u - \boldsymbol{H}[V]\big) \right\|. \tag{4.161}$$

Proof By Lemma 4.6.0.1 we can write

$$\begin{aligned}\left\|\boldsymbol{A}{:}\left(\mathrm{D}^2u-\boldsymbol{H}[U]\right)\right\|^2&=\left\langle\boldsymbol{A}{:}\left(\mathrm{D}^2u-\boldsymbol{H}[U]\right),\boldsymbol{A}{:}\left(\mathrm{D}^2u-\boldsymbol{H}[U]\right)\right\rangle\\&=\left\langle\boldsymbol{A}{:}\left(\mathrm{D}^2u-\boldsymbol{H}[U]\right),\boldsymbol{A}{:}\mathrm{D}^2u\right\rangle\\&=\left\langle\boldsymbol{A}{:}\left(\mathrm{D}^2u-\boldsymbol{H}[U]\right),\boldsymbol{A}{:}\left(\mathrm{D}^2u-\boldsymbol{H}[V]\right)\right\rangle\\&\leq\left\|\boldsymbol{A}{:}\left(\mathrm{D}^2u-\boldsymbol{H}[U]\right)\right\|\left\|\boldsymbol{A}{:}\left(\mathrm{D}^2u-\boldsymbol{H}[V]\right)\right\|,\end{aligned}\tag{4.162}$$

where we have used the Cauchy–Bunyakovskii–Schwartz inequality. The result follows after simplification. □

4.6.0.3 Lemma (convergence of $\mathrm{L}_2(\Omega)$-projection in a negative norm)**.** *Let $v\in\mathrm{H}^j(\Omega)$ for some $0\leq j\leq p+1$ and $P^{\mathbb{V}}v$ denote the $\mathrm{L}_2(\Omega)$-projection of v onto $\mathbb{V}$ then there exists a $C>0$ such that the following bound holds*

$$\left\|v-P^{\mathbb{V}}v\right\|_{\mathrm{H}^{-1}(\Omega)}\leq Ch^{j+1}\left|v\right|_j.\tag{4.163}$$

Proof From the definition of the $\mathrm{H}^{-1}(\Omega)$ norm we have

$$\left\|v-P^{\mathbb{V}}v\right\|_{\mathrm{H}^{-1}(\Omega)}=\sup_{0\neq\phi\in\mathrm{H}^1_0(\Omega)}\frac{\left\langle v-P^{\mathbb{V}}v,\phi\right\rangle}{\left|\phi\right|_1}.\tag{4.164}$$

By definition for any v, $v-P^{\mathbb{V}}v$ is orthogonal to $\mathring{\mathbb{V}}$ in $\mathrm{L}_2(\Omega)$

$$\begin{aligned}\left\|v-P^{\mathbb{V}}v\right\|_{\mathrm{H}^{-1}(\Omega)}&=\sup_{0\neq\phi\in\mathrm{H}^1_0(\Omega)}\frac{\left\langle v-P^{\mathbb{V}}v,\phi-P^{\mathbb{V}}\phi\right\rangle}{\left|\phi\right|_1}\\&\leq\sup_{0\neq\phi\in\mathrm{H}^1_0(\Omega)}\frac{\left\|v-P^{\mathbb{V}}v\right\|\left\|\phi-P^{\mathbb{V}}\phi\right\|}{\left|\phi\right|_1}\\&\leq\sup_{0\neq\phi\in\mathrm{H}^1_0(\Omega)}\frac{C_1h^j\left|v\right|_j h\left|\phi\right|_1}{\left|\phi\right|_1}\\&\leq C_1h^{j+1}\left|v\right|_j,\end{aligned}\tag{4.165}$$

giving the desired result. □

4.6.0.4 Theorem (convergence in a negative norm)**.** *Let u and U be defined as in Theorem 4.6.0.1. Suppose also that $f\in\mathrm{H}^j(W)$ for some $j=0,\ldots,p+1$ then there exists a constant C depending on $\mathbb{V}$ and Ω such that*

$$\left\|\boldsymbol{A}{:}(\mathrm{D}^2u-\boldsymbol{H}[U])\right\|_{\mathrm{H}^{-1}(\Omega)}\leq Ch^{j+1}\left|f\right|_j.\tag{4.166}$$

Proof Lemma 4.6.0.1 implies

$$\boldsymbol{A}{:}\boldsymbol{H}[U] = P^{\mathbb{V}}(\boldsymbol{A}{:}\mathrm{D}^2 u). \tag{4.167}$$

The result then follows from Lemma 4.6.0.3 and the PDE (4.18). □

4.6.0.5 Definition (dual problem)**.** We will conduct the aposteriori analysis making use of a duality argument. The *dual* or *adjoint problem* to (4.18) is given as follows: find v such that

$$\left\langle \mathrm{D}^2{:}[v\boldsymbol{A}] \,|\, \phi \right\rangle = \langle g, \phi \rangle \qquad \forall\, \phi \in \mathrm{H}^1_0(\Omega), \tag{4.168}$$

with $\mathrm{D}^2{:}[v\boldsymbol{A}]$ defined as

$$\mathrm{D}^2{:}[v\boldsymbol{A}] := \sum_{\alpha=1}^{d}\sum_{\beta=1}^{d} \partial_\alpha \partial_\beta \left[v\boldsymbol{A}^{\alpha,\beta} \right]. \tag{4.169}$$

Note that the dual problem satisfies the following identity, given $v \in \mathrm{H}^1_0(\Omega)$

$$\begin{aligned} \left\langle \mathrm{D}^2{:}(v\boldsymbol{A}) \,|\, \phi \right\rangle &= -\left\langle \mathrm{div}\,(v\boldsymbol{A}), \nabla\phi \right\rangle \\ &= \left\langle v \,|\, \boldsymbol{A}{:}\mathrm{D}^2\phi \right\rangle \qquad \forall\, \phi \in \mathrm{H}^1_0(\Omega), \end{aligned} \tag{4.170}$$

hence we are able to invoke the following regularity result.

4.6.0.6 Theorem (regularity of the dual problem [GT83, Thm 5.8])**.** *Let $\Omega \subset \mathbb{R}^d$ be an open, bounded, Lipschitz domain. Given $g \in \mathrm{L}_2(\Omega)$, $\boldsymbol{A} \in \mathrm{W}^1_\infty(\Omega)^{d\times d}$ such that the dual problem (4.168) is uniformly elliptic. Then $v \in \mathrm{H}^2(\Omega)$ and there exists a C such that*

$$|v|_2 \leq C \|g\|. \tag{4.171}$$

4.6.0.7 Assumption (regularity of the dual problem's coefficients)**.** To invoke the regularity result we will assume the dual problem's coefficients (4.168) are "sufficiently regular" to rewrite it in divergence form, that is $\boldsymbol{A} \in \mathrm{W}^1_\infty(\Omega)^{d\times d}$.

4.6.0.8 Lemma (Clément interpolant [Clé75])**.** *We introduce the Clément interpolation operator $\mathit{\Pi} : \mathrm{H}^1_0(\Omega) \to \mathbb{V}$ which under necessary regularity assumptions on ϕ satisfies the following local interpolation bounds for each $j \leq p+1$*

$$\|\phi - \mathit{\Pi}\phi\|_{\mathrm{L}_2(K)} \leq C h_K^j \, |\phi|_{\mathrm{H}^j(\hat{K})} \tag{4.172}$$

$$\|\phi - \mathit{\Pi}\phi\|_{\mathrm{L}_2(S)} \leq C h_K^{j-1/2} \, |\phi|_{\mathrm{H}^j(\hat{K})} \tag{4.173}$$

where K is a simplex and S is a common wall shared between two simplexes and $\hat{K}$ is a localised neighbourhood of K.

4.6.0.9 Lemma (local duality error bound)**.** *Let u and U be defined as in Theorem 4.6.0.1. Let the conditions of Assumption 4.6.0.7 hold. Assume also that the conditions of Theorem 4.6.0.6 hold, then there exists a $C > 0$ such that*

$$\|u - U\| \leq C \sum_{K \in \mathscr{T}} h_K \left\| \boldsymbol{A} : \left(\mathrm{D}^2 u - \mathrm{D}^2 U \right) \right\|_{\mathrm{H}^{-1}(K)} . \tag{4.174}$$

Proof Let v be the solution to the dual problem (4.168). Using the Galërkin orthogonality result for Theorem 4.6.0.1 and the apriori regularity result (Theorem 4.6.0.6)

$$\begin{aligned}
\langle u - U, g \rangle &= \left\langle u - U, \mathrm{D}^2 : (v\boldsymbol{A}) \right\rangle \\
&= \left\langle \boldsymbol{A} : \mathrm{D}^2 (u - U) \,|\, v \right\rangle \\
&= \sum_{K \in \mathscr{T}} \int_K \boldsymbol{A} : \mathrm{D}^2 (u - U) v \\
&= \sum_{K \in \mathscr{T}} \int_K \boldsymbol{A} : \mathrm{D}^2 (u - U)(v - \Pi v) \\
&\leq \sum_{K \in \mathscr{T}} \left\| \boldsymbol{A} : \mathrm{D}^2 (u - U) \right\|_{\mathrm{H}^{-1}(K)} |v - \Pi v|_{\mathrm{H}^1(K)} \\
&\leq C \sum_{K \in \mathscr{T}} h_K \, |v|_{\mathrm{H}^2(\hat{K})} \left\| \boldsymbol{A} : \mathrm{D}^2 (u - U) \right\|_{\mathrm{H}^{-1}(K)} \\
&\leq C \, \|g\|_{\mathrm{L}_2(\Omega)} \sum_{K \in \mathscr{T}} h_K \left\| \boldsymbol{A} : \mathrm{D}^2 (u - U) \right\|_{\mathrm{H}^{-1}(K)} .
\end{aligned} \tag{4.175}$$

The result follows by noting g was a generic test function. □

4.6.0.10 Corollary (global duality error bound)**.** *We may use the technique in Lemma 4.6.0.9 to derive a global duality error bound. To that end, let u and U be defined as in Theorem 4.6.0.1 and let the conditions of Assumption 4.6.0.7 hold. Then there exists a $C > 0$ such that*

$$\|u - U\| \leq Ch \left\| \boldsymbol{A} : \mathrm{D}^2 u - \mathrm{D}^2 U \right\|_{-1} \tag{4.176}$$

Proof The proof follows the same lines as that of Lemma 4.6.0.9 with the exception that the integral is not divided into its elementwise contributions. □

4.6.0.11 Theorem (aposteriori residual upper error bound)**.** *Let u and U be defined as in Theorem 4.6.0.1. then there exists a $C > 0$ such that*

$$\left\|f - \boldsymbol{A}{:}\mathrm{D}^2 U\right\|_{\mathrm{H}^{-1}(\Omega)} \leq C \left(\sum_{K \in \mathscr{T}} h_K \left\|\mathcal{R}[U, \boldsymbol{A}, f]\right\|_{\mathrm{L}_2(K)} + \sum_{S \in \mathscr{S}} h_K^{1/2} \left\|\mathcal{J}[U, \boldsymbol{A}]\right\|_{\mathrm{L}_2(S)} \right) \tag{4.177}$$

where the interior residual, $\mathcal{R}[U, \boldsymbol{A}, f]$, over a simplex K and jump residual, $\mathcal{J}[U, \boldsymbol{A}]$, over a common wall $S = \overline{K}^+ \cap \overline{K}^-$ of two simplexes, K^+ and K^- are defined as

$$\left\|\mathcal{R}[U, \boldsymbol{A}, f]\right\|^2_{\mathrm{L}_2(K)} = \int_K \left(f - \boldsymbol{A}{:}\mathrm{D}^2 U\right)^2, \tag{4.178}$$

$$\left\|\mathcal{J}[U, \boldsymbol{A}]\right\|^2_{\mathrm{L}_2(S)} = -\int_S \left(\boldsymbol{A}{:}\left(\nabla U\big|_{K^+} \otimes \boldsymbol{n}_{K^+}\right) + \boldsymbol{A}{:}\left(\nabla U\big|_{K^-} \otimes \boldsymbol{n}_{K^-}\right) \right)^2, \tag{4.179}$$

with $\boldsymbol{n}_{K^+}$ and $\boldsymbol{n}_{K^-}$ denoting the outward pointing normals to K^+ and K^- respectively.

Proof By the definition of the $\mathrm{H}^{-1}(\Omega)$ norm it follows

$$\begin{aligned}
\left\langle f - \boldsymbol{A}{:}\mathrm{D}^2 U \,|\, \phi \right\rangle &= \sum_{K \in \mathscr{T}} \int_K \left(f - \boldsymbol{A}{:}\mathrm{D}^2 U\right) \phi \\
&= \sum_{K \in \mathscr{T}} \int_K f\phi - \mathrm{D}^2 U{:}\phi\boldsymbol{A} \\
&= \sum_{K \in \mathscr{T}} \int_K f\phi - \sum_{\alpha,\beta=1}^{d} \partial_\alpha \partial_\beta U \phi \boldsymbol{A}^{\alpha,\beta} \\
&= \sum_{K \in \mathscr{T}} \int_K f\phi + \sum_{\alpha,\beta=1}^{d} \partial_\beta U \partial_\alpha \left(\phi \boldsymbol{A}^{\alpha,\beta}\right) \\
&= \sum_{K \in \mathscr{T}} \int_K f\phi - \sum_{\alpha,\beta=1}^{d} \left(\partial_\alpha \partial_\beta U \phi \boldsymbol{A}^{\alpha,\beta} + \int_{\partial K \setminus \partial\Omega} \partial_\beta U \boldsymbol{n}_\alpha \phi \boldsymbol{A}^{\alpha,\beta} \right) \\
&= \sum_{K \in \mathscr{T}} \int_K f\phi - \mathrm{D}^2 U{:}\phi\boldsymbol{A} + \int_{\partial K \setminus \partial\Omega} \boldsymbol{A}{:}\left(\nabla U \boldsymbol{n}_K^{\intercal}\right)\phi.
\end{aligned} \tag{4.180}$$

Utilising the definition of interior and jump residuals and noting from Lemma 4.6.0.1 that $f - \boldsymbol{A}{:}\mathrm{D}^2 U$ is polar to $\mathbb{V}$

$$\begin{aligned}
\left\langle f - \boldsymbol{A}{:}\mathrm{D}^2 U \,|\, \phi \right\rangle &= \sum_{K \in \mathscr{T}} \int_K \mathcal{R}[U, \boldsymbol{A}, f]\phi - \sum_{S \in \mathscr{S}} \int_S \mathcal{J}[U, \boldsymbol{A}]\phi \\
&= \sum_{K \in \mathscr{T}} \int_K \mathcal{R}[U, \boldsymbol{A}, f]\left(\phi - \Pi\phi\right) - \sum_{S \in \mathscr{S}} \int_S \mathcal{J}[U, \boldsymbol{A}]\left(\phi - \Pi\phi\right) \\
&\leq \sum_{K \in \mathscr{T}} \left\|\mathcal{R}[U, \boldsymbol{A}, f]\right\|_{\mathrm{L}_2(K)} \left\|\phi - \Pi\phi\right\|_{\mathrm{L}_2(K)} \\
&\quad + \sum_{S \in \mathscr{S}} \left\|\mathcal{J}[U, \boldsymbol{A}]\right\|_{\mathrm{L}_2(S)} \left\|\phi - \Pi\phi\right\|_{\mathrm{L}_2(S)}.
\end{aligned} \tag{4.181}$$

The properties of Π (4.172) now infer

$$\left\langle f - \boldsymbol{A}{:}\mathrm{D}^2 U \,|\, \phi \right\rangle \leq C\,|\phi|_1 \left(\sum_{K\in\mathscr{T}} h_K \left\| \mathcal{R}[U, \boldsymbol{A}, f] \right\|_{\mathrm{L}_2(K)} + \sum_{S\in\mathscr{S}} h_K^{1/2} \left\| \mathcal{J}[U, \boldsymbol{A}] \right\|_{\mathrm{L}_2(S)} \right) \tag{4.182}$$

giving the desired result. □

4.6.0.12 Theorem ($\mathrm{L}_2(\Omega)$ aposteriori error bound)**.** *We have the following aposteriori bound on the* $\mathrm{L}_2(\Omega)$ *error.*

$$\|u - U\| \leq Ch \left(\sum_{K\in\mathscr{T}} h_K \left\| \mathcal{R}[U, \boldsymbol{A}, f] \right\|_{\mathrm{L}_2(K)} + \sum_{S\in\mathscr{S}} h_K^{1/2} \left\| \mathcal{J}[U, \boldsymbol{A}] \right\|_{\mathrm{L}_2(S)} \right). \tag{4.183}$$

Proof From Corollary 4.6.0.10 we see

$$\|u - U\| \leq h \left\| \boldsymbol{A}{:}\big(\mathrm{D}^2 u - \mathrm{D}^2 U\big) \right\|_{-1}. \tag{4.184}$$

Then applying Theorem 4.6.0.11 gives

$$\|u - U\| \leq Ch \left(\sum_{K\in\mathscr{T}} h_K \left\| \mathcal{R}[U, \boldsymbol{A}, f] \right\|_{\mathrm{L}_2(K)} + \sum_{S\in\mathscr{S}} h_K^{1/2} \left\| \mathcal{J}[U, \boldsymbol{A}] \right\|_{\mathrm{L}_2(S)} \right), \tag{4.185}$$

as required. □

4.6.0.13 Remark (improving the aposteriori bound)**.** The $\mathrm{L}_2(\Omega)$ bound given in Theorem 4.6.0.12 can be improved to give one which is more reminiscent of the standard residual estimators from the FEM if we can find a way to tie Lemma 4.6.0.9 with Theorem 4.6.0.12. To do this we would need a bound of the following form: Given $v \in \mathrm{H}^{-1}(\Omega)$ there exists a $C > 0$ such that

$$\sum_K \|v\|_{\mathrm{H}^{-1}(K)} \leq C \|v\|_{-1}. \tag{4.186}$$

Work has been done to this end, albeit in the context of boundary element methods, by Faermann [Fae00].

4.6.0.14 Remark (the regularity assumption $\boldsymbol{A} \in \mathrm{W}^1_\infty(\Omega)$ (Assumption 4.6.0.7))**.** This assumption is only necessary for the duality argument from Lemma 4.6.0.9 to hold. The assumption is not needed to prove the residual bound in $\mathrm{H}^{-1}(\Omega)$ arising from Theorem 4.6.0.11. In fact as will be numerically demonstrated in §4.7 the estimator (4.183) is both efficient and reliable even if $\boldsymbol{A} \notin \mathrm{W}^1_\infty(\Omega)$.

4.7 Numerical experiments

In this section we benchmark the problem with respect to the $\mathrm{H}^{-1}(\Omega)$ error, to check Theorem 4.6.0.4. In addition we study the numerical behaviour of the residual error estimator given in Theorem 4.6.0.12 and compare it with the error on three model problems.

To numerically demonstrate convergence of the error in the norm given in Theorem 4.6.0.4 we make use of the following Lemma, which shows how to practically approximate the $\mathrm{H}^{-1}(\Omega)$ norm of an arbitrary given function $v \in \mathrm{L}_2(\Omega)$.

4.7.0.15 Lemma (computing the $\mathrm{H}^{-1}(\Omega)$ norm [LP10d])**.** *Let $v \in \mathrm{L}_2(\Omega)$, consider the function $\Psi \in \mathbb{V}$ such that*

$$A\Psi = Pv, \tag{4.187}$$

where A and P are the discrete Laplacian and the $\mathrm{L}_2(\Omega)$ projection on $\mathbb{V}$, respectively. Then we have

$$\|v\|^2_{\mathrm{H}^{-1}(\Omega)} = \|\Psi\|^2_{\mathrm{H}^1_0(\Omega)} + \zeta[\Psi, v]^2, \text{ where } \zeta[\Psi, v] \leq \mathscr{E}[\Psi] \tag{4.188}$$

where $\mathscr{E}$ is a fully computable aposteriori estimator functional (see §2.4 for example).

Proof Let $\psi \in \mathrm{H}^1_0(\Omega)$ such that $-\Delta\psi = v$ and Ψ given in (4.187) we have $\Phi \in \mathbb{V}$

$$\langle -\Delta\psi - A\Psi \,|\, \Phi\rangle = \langle v - Pv, \Phi\rangle = 0, \tag{4.189}$$

i.e., that $\psi - \Psi$ is Galërkin-orthogonal to $\mathbb{V}$. Also, we have

$$\|v\|_{\mathrm{H}^{-1}(\Omega)} = \|\psi\|_{\mathrm{H}^1_0(\Omega)}\,. \tag{4.190}$$

Indeed, on the one hand

$$\begin{aligned} \|v\|_{\mathrm{H}^{-1}(\Omega)} &:= \sup_{\phi\in\mathrm{H}^1_0(\Omega)} \frac{\langle v, \phi\rangle}{\|\phi\|_{\mathrm{H}^1_0(\Omega)}} = \sup_{\phi\in\mathrm{H}^1_0(\Omega)} \frac{\langle \nabla\psi, \nabla\phi\rangle}{\|\phi\|_{\mathrm{H}^1_0(\Omega)}} \\ &\leq \sup_{\phi\in\mathrm{H}^1_0(\Omega)} \frac{\|\psi\|_{\mathrm{H}^1_0(\Omega)}\,\|\phi\|_{\mathrm{H}^1_0(\Omega)}}{\|\phi\|_{\mathrm{H}^1_0(\Omega)}} = \|\psi\|_{\mathrm{H}^1_0(\Omega)}\,, \end{aligned} \tag{4.191}$$

and, on the other hand

$$\|v\|_{\mathrm{H}^{-1}(\Omega)} := \sup_{\phi\in\mathrm{H}^1_0(\Omega)} \frac{\langle \nabla\psi, \nabla\phi\rangle}{\|\phi\|_{\mathrm{H}^1_0(\Omega)}} \geq \frac{\langle \nabla\psi, \nabla\psi\rangle}{\|\psi\|_{\mathrm{H}^1_0(\Omega)}} = \|\psi\|_{\mathrm{H}^1_0(\Omega)}\,. \tag{4.192}$$

By the above, Galërkin-orthogonality and Pythagoras's Theorem, we have

$$\|v\|^2_{\mathrm{H}^{-1}(\Omega)} = \|\psi\|^2_{\mathrm{H}^1_0(\Omega)} = \|\psi - \Psi\|^2_{\mathrm{H}^1_0(\Omega)} + \|\Psi\|^2_{\mathrm{H}^1_0(\Omega)}\,. \tag{4.193}$$

The term $\|\psi - \Psi\|_{\mathrm{H}^1_0(\Omega)}$ is the error of a function and its Ritz projection. This can be easily estimated with a fully computable aposteriori estimator functional $\mathscr{E}$ such that

$$\|\psi - \Psi\|_{\mathrm{H}^1_0(\Omega)} \leq \mathscr{E}[\Psi, v, \mathbb{V}] = \mathrm{O}(h^r_{\mathbb{V}}), \tag{4.194}$$

where $h_{\mathbb{V}}$ is the "mesh-size" of the space $\mathbb{V}$. □

4.7.0.16 Remark (our use of Lemma 4.7.0.15)**.** In our case we wish to compute $\|e\|_{-1} = \big\|\boldsymbol{A}{:}\big(\mathrm{D}^2 u - \boldsymbol{H}[U]\big)\big\|_{-1}$ which our analysis shows for regular u can be as small as $\mathrm{O}(h^{p+2}_{\mathbb{V}})$. To compute $\|e\|_{-1}$ effectively we must take $r \geq p+2$. The discrete formulation then becomes: Find $\Psi \in \mathrm{H}^1_0(\Omega)$ such that

$$A\Psi = P^{\mathbb{W}} \Lambda e, \tag{4.195}$$

where $\Lambda : \mathbb{V} \to \mathbb{W}$ is the Lagrange interpolant. For clarity we give a pseudocode for the method.

4.7.1 Approximating convergence rates from Lemma 4.6.0.4

Require: $(\mathscr{T}_0, p, r, K_{\max})$

Ensure: $(\{\|e_k\|_{-1}\}_{k=0}^{K_{\max}})$ a sequence of approximations to the left hand side of (4.166)

$k = 0$

while $k \leq K_{\max}$ **do**

$\mathbb{V}_k = \mathsf{Fe\ Space}(\mathscr{T}_k, p)$

$U = \mathsf{NVFEM}(\mathbb{V}_k, \boldsymbol{A}, f, g)$

$e_k := \boldsymbol{A}{:}\big(\mathrm{D}^2 u - \boldsymbol{H}[U]\big)$

$\mathbb{W}_k = \mathsf{Fe\ Space}(\mathscr{T}_k, r)$

$\Lambda v := \mathsf{Interpolate}(e_k, \mathbb{V}_k, \mathbb{W}_k)$

$E_k = \mathsf{FEM}(\mathbb{W}_k, -\boldsymbol{I}, \Lambda e, 0)$

$\|e_k\|_{-1} := \|\nabla E_k\|$

$\mathscr{T}_{k+1} = \mathsf{Global\ Refine}(\mathscr{T})$

$k := k + 1$

end while

We make use of a Matlab ® code in order to compute the NVFE solution. To numerically compute the estimated convergence rates of $\|e\|_{-1}$ we interface this code with one based on the adaptive FEM library ALBERTA [SS05]. In all numerical experiments the quadrature error is made negligible by taking a quadrature which is exact on polynomials of degree 9 and less.

Table 4.2: Numerical results for the apriori convergence given in Theorem 4.6.0.4. We have applied Algorithm 4.7.1 to compute the $\mathrm{H}^{-1}(\Omega)$ norm with $p = 1$, $r = 3$ and $K_{max} = 8$. In this problem the data is chosen in such a way that $u(\boldsymbol{x}) = \exp(-10\,|\boldsymbol{x}|^2)$.

$\dim \mathbb{V}$	$\dim \mathbb{W}$	$\|\boldsymbol{A}:(\mathrm{D}^2 u - \boldsymbol{H}[U])\|_{-1}$	$\mathrm{EOC}[\|\boldsymbol{A}:(\mathrm{D}^2 u - \boldsymbol{H}[U])\|_{-1}]$
5	25	4.93×10^{-1}	
13	85	9.90×10^{-2}	2.315135
41	313	1.75×10^{-3}	5.818990
145	1201	1.06×10^{-3}	0.726793
545	4705	1.38×10^{-4}	2.944427
2113	18625	1.72×10^{-5}	3.001582
8321	74113	2.15×10^{-6}	3.000756
33025	295681	2.69×10^{-7}	3.000192
131585	1181185	3.36×10^{-8}	3.000027

4.7.2 Effectivity of the estimator given in Theorem 4.6.0.12

We test the effectivity of

$$\mathscr{E}[U, \boldsymbol{A}, f] := h \left(\sum_{K \in \mathscr{T}} h_K \, \|\mathcal{R}[U, \boldsymbol{A}, f]\|_{\mathrm{L}_2(K)} + \sum_{S \in \mathscr{S}} h_K^{1/2} \, \|\mathcal{J}[U, \boldsymbol{A}]\|_{\mathrm{L}_2(S)} \right) \tag{4.196}$$

where the interior and jump residuals are given as

$$\|\mathcal{R}[U, \boldsymbol{A}, f]\|^2_{\mathrm{L}_2(K)} = \int_K \left(f - \boldsymbol{A}:\mathrm{D}^2 U\right)^2, \tag{4.197}$$

$$\|\mathcal{J}[U, \boldsymbol{A}]\|^2_{\mathrm{L}_2(S)} = \int_S \left(-\boldsymbol{A}:\left(\nabla U\big|_{K^+} \otimes \boldsymbol{n}_{K^+}\right) - \boldsymbol{A}:\left(\nabla U\big|_{K^-} \otimes \boldsymbol{n}_{K^-}\right)\right)^2. \tag{4.198}$$

This is an upper bound for the $\mathrm{L}_2(\Omega)$ norm of the error.

We proceed by conducting the first three tests as performed in §4.5. That is we consider the operator

$$\boldsymbol{A}(\boldsymbol{x}) = \begin{bmatrix} 1 & b(\boldsymbol{x}) \\ b(\boldsymbol{x}) & a(\boldsymbol{x}) \end{bmatrix} \tag{4.199}$$

and vary the coefficients $a(\boldsymbol{x})$ and $b(\boldsymbol{x})$.

4.7.3 Test problem with a nondifferentiable operator

For the first test problem we choose the operator in such a way that (1.2) does not hold, that is the components of $\boldsymbol{A}$ are non-differentiable on Ω, in this case we take

$$a(\boldsymbol{x}) = (x_1^2 x_2^2)^{1/3} + 1 \tag{4.200}$$

$$b(\boldsymbol{x}) = 0 \tag{4.201}$$

and take the problem data such that the exact solution is given by

$$u(\boldsymbol{x}) = \exp(-10\,|\boldsymbol{x}|^2). \tag{4.202}$$

Figure 4.10 shows numerical results for this problem.

4.7.4 Test problem with convection dominated operator

In this case we choose

$$a(\boldsymbol{x}) = \arctan\Big(K(|\boldsymbol{x}|^2 - 1)\Big) + 2 \tag{4.203}$$

$$b(\boldsymbol{x}) = 0. \tag{4.204}$$

The operator (4.203) can be rewritten in divergence form however the derivatives can be arbitrarily large and hence a conforming FEM may be unstable. We choose the problem data such that the exact solution to the problem is given by:

$$u(\boldsymbol{x}) = \sin(\pi x_1)\sin(\pi x_2)\,. \tag{4.205}$$

Figure 4.11 shows numerical results for this problem.

4.7.5 Test problem choosing a solution with nonsymmetric Hessian

We choose

$$a(\boldsymbol{x}) = 2 \tag{4.206}$$

$$b(\boldsymbol{x}) = (x_1^2 x_2^2)^{1/3}. \tag{4.207}$$

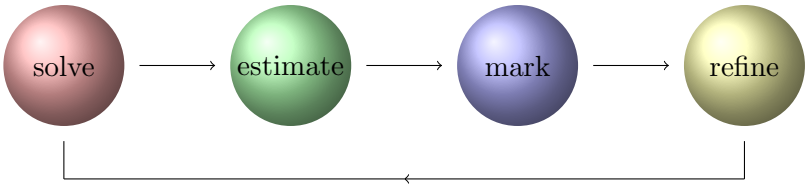
We choose the problem data such that the exact solution is given by

$$u(\boldsymbol{x}) = \begin{cases} \frac{x_1 x_2 (x_1^2 - x_2^2)}{x_1^2 + x_2^2} & \boldsymbol{x} \neq \mathbf{0} \\ 0 & \boldsymbol{x} = \mathbf{0}. \end{cases} \tag{4.208}$$

This function has a nonsymmetric Hessian at the point $\mathbf{0}$. Figure 4.12 shows numerical results for this problem.

4.7.6 Adaptivity

The adaptive algorithm we make use of is of standard type, that is



The *solve* algorithm is assemblage and solution of the NDFEM discrete system detailed in §4.1.3 to §4.2. The *estimate* is the $\mathrm{L}_2(\Omega)$ residual estimator (4.196). The *mark*ing strategy we use is the maximum strategy, that is all elements $K \in \mathscr{T}$ satisfying

$$\eta_K^2 \geq \xi \max_{L \in \mathscr{T}} \eta_L^2 \tag{4.209}$$

are marked for refinement, where ξ is a user defined constant. We *refine* using the newest vertex bisection, as described in [SS05] §1.1 (and summarised in §3.8), where each element marked for refinement is divided d-times. We pseudocode the algorithm as follows:

4.7.7 ANVFEM

Require: $(\mathbb{V}_0, \mathrm{tol}, k_{\max}, \xi)$

Ensure: $(U, \mathbb{V})$ solution of (4.21)

$\mathscr{T}_0 := \mathsf{Mesh}(\mathbb{V}_0)$

$U_0 := \mathsf{Solve}(\mathscr{T}, \boldsymbol{A}, f)$

$k := 0$

$\mathscr{E} := \mathsf{Estimate}(\boldsymbol{A}, f, \mathscr{T}_0)$

$\mathscr{R} := \emptyset$

while $\mathscr{E} >$ tol and $k \leq k_{\max}$ **do**

  for all $K \in \mathscr{T}_k$ **do**

    if $\mathscr{E}^2 \geq \xi \max_{L \in \mathscr{T}} \mathscr{E}^2$ **then**

      $\mathscr{R} := \{K\} \cup \mathscr{R}$

    end if

  end for

  $\mathscr{T}_{k+1} := \mathsf{Refine}(\mathscr{T}_k, \mathscr{R})$ using [SS05, §1.1.1]

  $U_{k+1} := \mathsf{Solve}(\mathscr{T}_{k+1}, \boldsymbol{A}, f)$

  $\mathscr{E} = \mathsf{Estimate}(\mathscr{T}_{k+1}, \boldsymbol{A}, f)$

  $k := k + 1$

end while

$U := U_k$

$\mathbb{V} := \mathsf{Mesh}^{-1}(\mathscr{T}_k)$

return $(U, \mathbb{V})$

Note we are not incorporating any mesh coarsening into the algorithm as we are not concerned with mesh optimality, or any convergence proofs for the adaptive scheme [BDD04].

4.7.8 Test problem for an irregular solution

In this test we choose

$$a(\boldsymbol{x}) = \sin\left(\frac{1}{|x_1| + |x_2| + 10^{-15}}\right) \tag{4.210}$$

$$b(\boldsymbol{x}) = 0. \tag{4.211}$$

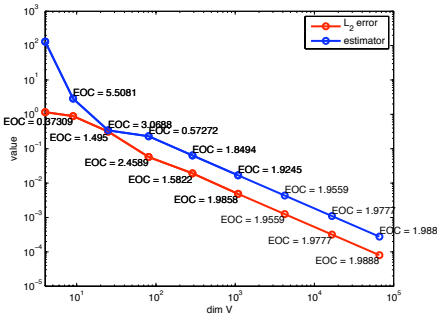
The problem data is chosen such that the solution is given by

$$u(\boldsymbol{x}) = -\sqrt{2 - x_1^2 - x_2^2}. \tag{4.212}$$

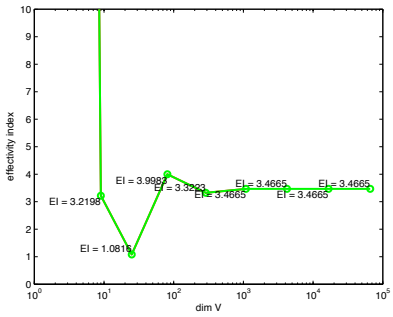
Note that the function has singular derivatives on the boundary. We run the Algorithm 4.7.7 with the given problem data with tol $= 0.01$ and $\xi = 0.5$. Figure 4.13 shows a surface

plot of the ANVFE solution together with the mesh generated. Figure 4.14 demonstrates we regain optimal convergence under the adaptive strategy.

Figure 4.10: A numerical study on the performance of the residual estimator given in (4.196) on a problem with a nondifferentiable operator (4.200) with data chosen such that $u(\boldsymbol{x}) = \exp(-10\,|\boldsymbol{x}|^2)$.

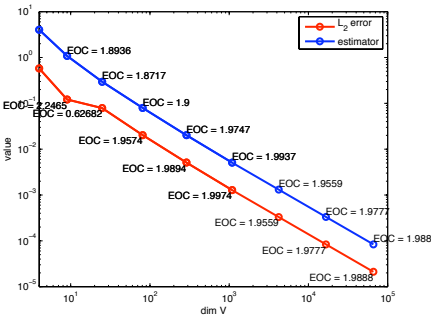


(a) The values on a logarithmic scale of the $\mathscr{E}$ (blue) and the $\mathrm{L}_2(\Omega)$ error (red) together with their EOC s. Notice both quantities converge to zero at a rate of $\mathrm{O}\,(h^2)$.

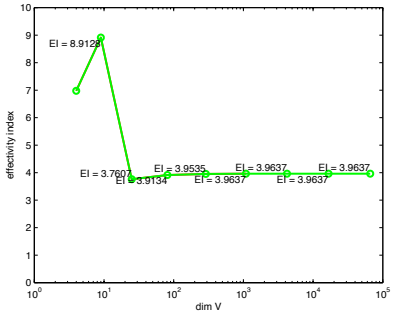


(b) The effectivity index of $\mathscr{E}$. The values show the estimator overestimates the error, as expected with residual estimates.

Figure 4.11: A numerical study on the performance of the residual estimator given in (4.196) on a problem with a convection dominated operator (4.203) with data chosen such that $u(\boldsymbol{x}) = \sin(\pi x_1)\sin(\pi x_2)$.

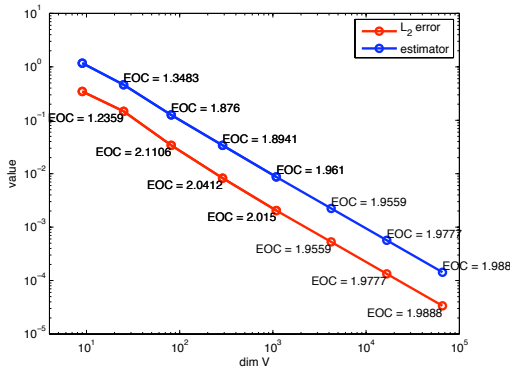


(a) The values on a logarithmic scale of the $\mathscr{E}$ (blue) and the $L_2(\Omega)$ error (red) together with their EOC s. Notice both quantities converge to zero at a rate of $O(h^2)$.

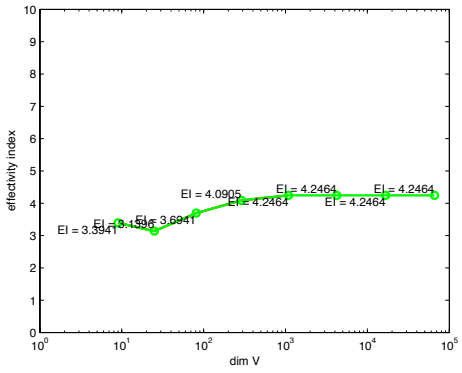


(b) The effectivity index of $\mathscr{E}$. The values show the estimator overestimates the error, as expected with residual estimates.

Figure 4.12: A numerical study on the performance of the residual estimator given in (4.196) on a problem with a convection dominated operator (4.203) with data chosen such that $u(\boldsymbol{x}) = \frac{x_1 x_2 (x_1^2 - x_2^2)}{x_1^2 + x_2^2}$ if $\boldsymbol{x} \neq \boldsymbol{0}$, or $u(\boldsymbol{x}) = 0$ otherwise.

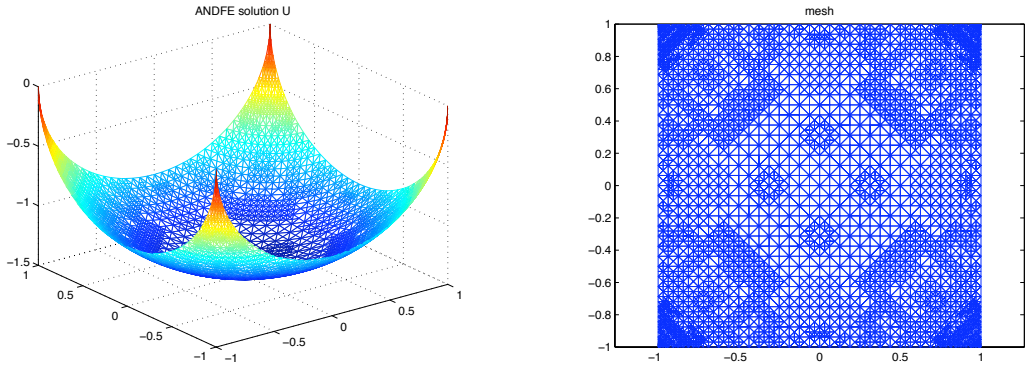


(a) The values on a logarithmic scale of the $\mathscr{E}$ (blue) and the $L_2(\Omega)$ error (red) together with their EOC s. Notice both quantities converge to zero at a rate of $O(h^2)$.



(b) The effectivity index of $\mathscr{E}$. The values show the estimator overestimates the error, as expected with residual estimates.

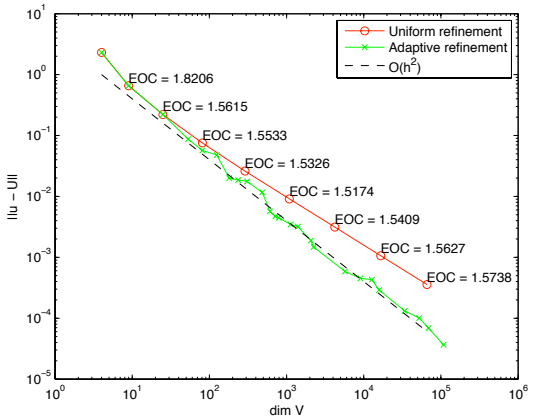
Figure 4.13: A plot of the ANDFE approximation to a solution, u, of problem (4.210) where $u \notin \mathrm{H}^2(\Omega)$. The singularities occur on the corners of Ω, notice the mesh is well refined there.



(a) The ANDFE approximation to $u(\boldsymbol{x}) = -\sqrt{2 - x_1^2 - x_2^2}$

(b) The underlying mesh.

Figure 4.14: Here we study the adaptive strategy given in Algorithm 4.7.7. We formulate problem (4.210) and plot the error for the uniform refinement strategy and compare it to that of the adaptive strategy. Notice under the adaptive strategy we regain optimal convergence.



4.8 Quasilinear PDEs in nondivergence form

In this section we are applying the formulation given in §4.1.3 in the context of linear PDEs to accommodate general quasilinear PDEs.

4.8.0.1 Definition (Quasilinear PDE)**.** A quasilinear PDE is one which is linear with respect to its highest order derivative.

We consider the problem

$$\begin{aligned} \boldsymbol{A}(\nabla u, u, \boldsymbol{x}) : \mathrm{D}^2 u = f(\nabla u, u, \boldsymbol{x}) \qquad & \text{in } \Omega \\ u = 0 \qquad & \text{on } \partial\Omega, \end{aligned} \tag{4.213}$$

and apply a fixed point linearisation to the problem (4.213). This results in a sequence of linear PDEs. Given an initial guess u^0, for each $n \in \mathbb{N}_0$ we wish to find u^{n+1} such that

$$\begin{aligned} \boldsymbol{A}(\nabla u^n, u^n, \boldsymbol{x}) : \mathrm{D}^2 u^{n+1} = f(\nabla u^n, u^n, \boldsymbol{x}) \qquad & \text{in } \Omega \\ u = 0 \qquad & \text{on } \partial\Omega. \end{aligned} \tag{4.214}$$

The discretisation we propose is a simple extension of that set out in §4.1.3. That is, given an initial guess $U^0 \in \mathring{\mathbb{V}}$, find $U^{n+1} \in \mathring{\mathbb{V}}$ such that

$$\left\langle \boldsymbol{A}(\nabla U^n, U^n, \boldsymbol{x}) : \boldsymbol{H}[U^{n+1}], \mathring{\Phi} \right\rangle = \left\langle f, \mathring{\Phi} \right\rangle \quad \forall\, \mathring{\Phi} \in \mathring{\mathbb{V}}. \tag{4.215}$$

As a model problem we use the equation of prescribed mean curvature which is a quasilinear PDE arising from differential geometry:

$$\sqrt{1 + |\nabla u|^2}\, \mathrm{div} \left(\frac{\nabla u}{\sqrt{1 + |\nabla u|^2}} \right) = f \tag{4.216}$$

where $\sqrt{1 + |\nabla u|^2}$ is the area element. Here we are using $|\nabla u|^2 = \mathrm{D}u \nabla u$.

We may work on this problem combining the two nonlinear terms. To do so we must

first rewrite (4.216) into the form $A(\nabla u, u, \boldsymbol{x}){:}\mathrm{D}^2 u = f$.

$$\begin{aligned} f =& \sqrt{1+|\nabla u|^2}\,\mathrm{div}\left(\frac{\nabla u}{\sqrt{1+|\nabla u|^2}}\right) \\ =& \sqrt{1+|\nabla u|^2}\left(\frac{\Delta u}{\sqrt{1+|\nabla u|^2}} + \frac{\mathrm{D}\left(1+|\nabla u|^2\right)}{2\left(1+|\nabla u|^2\right)^{3/2}}\nabla u\right) \\ =& \Delta u + \frac{\mathrm{D}u\mathrm{D}^2 u\nabla u}{1+|\nabla u|^2} \\ =& \left(\boldsymbol{I} + \frac{\nabla u\otimes\nabla u}{1+|\nabla u|^2}\right){:}\mathrm{D}^2 u. \end{aligned} \tag{4.217}$$

Applying the fixed point linearisation from (4.214), given an initial guess u^0 for each $n \in \mathbb{N}_0$ we seek u^{n+1} such that

$$\left(\boldsymbol{I} + \frac{\nabla u^n \otimes \nabla u^n}{1+|\nabla u^n|^2}\right){:}\mathrm{D}^2 u^{n+1} = f. \tag{4.218}$$

Discretising the problem is then similar to that set out in §4.1.3. Restricting our attention to $\mathring{\mathbb{V}}$ the problem becomes given $U^0 = \Lambda^{\mathbb{V}} u^0$ find $U^{n+1} \in \mathring{\mathbb{V}}$ such that

$$\left\langle \left(\boldsymbol{I} + \frac{\nabla U^n \otimes \nabla U^n}{1+|\nabla U^n|^2}\right){:}\boldsymbol{H}[U^{n+1}], \mathring{\Phi}\right\rangle = \left\langle f, \mathring{\Phi}\right\rangle. \tag{4.219}$$

The equivalent linear algebra problem is: Find $\mathring{\mathbf{u}}^{n+1}$ such that

$$\mathring{\mathring{\mathbf{D}}}^n \mathring{\mathbf{u}} := \sum_{\alpha=1}^{d}\sum_{\beta=1}^{d} \mathring{\mathbf{B}}_n^{\alpha\beta}\mathbf{M}^{-1}\mathring{\mathbf{C}}_{\alpha\beta}\mathring{\mathbf{u}}^{n+1} = \mathring{\mathbf{f}}. \tag{4.220}$$

The component matrixes $\mathbf{M}$ and $\mathring{\mathbf{C}}_{\alpha\beta}$ are problem independent, $\mathring{\mathbf{B}}_n^{\alpha\beta}$ are defined as

$$\mathring{\mathbf{B}}_n^{\alpha\beta} = \begin{cases} \left\langle \mathring{\boldsymbol{\Phi}}, \left(1 + \frac{\partial_\alpha U^n \partial_\beta U^n}{1+|\nabla U^n|^2}\right)\boldsymbol{\Phi}^{\mathsf{T}}\right\rangle & \text{for } \alpha = \beta, \\ \left\langle \mathring{\boldsymbol{\Phi}}, \frac{\partial_\alpha U^n \partial_\beta U^n}{1+|\nabla U^n|^2}\boldsymbol{\Phi}^{\mathsf{T}}\right\rangle & \text{for } \alpha \neq \beta. \end{cases} \tag{4.221}$$

We pseudocode the algorithm proposed as follows:

4.8.1 The NVFEM for general quasilinear problems

Require: $(\mathscr{T}_0, u^0, p, N, K_{\max}, \mathrm{tol})$

Ensure: $(U, \mathbb{V})$ the NVFE approximation of (4.216)

$k = 0$
while $k \leq K_{\max}$ **do**
  $\mathbb{V}_k = \mathsf{Fe\ Space}(\mathscr{T}_k, p)$ (§2)
  if $k = 0$ **then**
    $U^0 = \Lambda^{\mathbb{V}_0} u^0$
  end if
  $n = 0$
  while $n \leq N$ **do**
    $U^{n+1} = \mathsf{NVFEM}(\mathbb{V}_k, \boldsymbol{A}^n[U^n], f)$ (§4.1.3)
    if $\left\|U^{n+1} - U^n\right\| \leq \mathrm{tol}$ **then**
      break
    end if
    $n = n + 1$
  end while
  $\mathscr{T}_{k+1} = \mathsf{Global\ Refine}(\mathscr{T}_k)$
  $k = k + 1$
end while

Of course we are able to work on the equation (4.216) in divergence form and make use of standard FE techniques. We could apply a fixed point linearisation as follows: Given an initial guess u^0 for each $n \in \mathbb{N}_0$ we seek u^{n+1} such that

$$\operatorname{div}\left(\frac{\nabla u^{n+1}}{\sqrt{1+|\nabla u^n|^2}}\right) = \frac{f}{\sqrt{1+|\nabla u^n|^2}}. \tag{4.222}$$

Applying a standard finite element discretisation of (4.222) yields: Given $U^0 \in \mathring{\mathbb{V}}$, find $U^{n+1} \in \mathring{\mathbb{V}}$ such that for each $\mathring{\Phi} \in \mathring{\mathbb{V}}$

$$\left\langle \frac{\nabla U^{n+1}}{\sqrt{1+|\nabla U^n|^2}}, \nabla \mathring{\Phi} \right\rangle = \left\langle \frac{f}{\sqrt{1+|\nabla U^n|^2}}, \mathring{\Phi} \right\rangle. \tag{4.223}$$

Table 4.3 compares the two linearisations (4.222) and (4.218). Figure 4.15 show asymptotic numerical convergence results for NVFEM applied to (4.218) under Algorithm 4.8.1.

Table 4.3: Test 4.8. Comparison of the fixed point linearisation in variational form (4.222) and in nonvariational form (4.218). We fix f appropriately such that $u(\boldsymbol{x}) = \sin(\pi x_1)\sin(\pi x_2)$. Taking initial guesses $U^0 = \tilde{U^0} = 0$ we discretise problem (4.216) using a standard FEM and using the NVFEM. Denoting U_i and $\tilde{U}_i$ to be the NVFE-solution FE-solution respectively we run both linearisations for until a tolerance $\|U_{n+1} - U_n\|$ (resp. $\left\|\tilde{U}_{n+1} - \tilde{U}_n\right\|$) $\leq h^2$ is achieved. We compute both the stagnation point—which is the iteration at which the prescribed tolerance is achieved—and the total CPU time. Notice there is significant savings in the number of iterations required to reach the stagnation point using the NVFEM over the standard FEM, however each iteration is computationally more costly using the NVFEM since the system is larger and more complicated to solve. The CPU cost for the entire algorithm is comparable for each fixed h.

	h	$\sqrt{2}/5$	$\sqrt{2}/10$	$\sqrt{2}/20$	$\sqrt{2}/40$	$\sqrt{2}/80$	$\sqrt{2}/160$
FEM	Stag. Point	5	13	16	26	32	36
	CPU Time	0.50	4.02	17.51	117.58	796.58	5308.81
NDFEM	Stag. Point	4	6	7	8	10	12
	CPU Time	0.72	3.40	16.49	97.93	838.8	5256.84

4.8.1.1 Lemma (on aposteriori estimation for mean curvature)**.** *We may follow the same error analysis as in Theorem 4.6.0.11. Let*

$$\boldsymbol{A}^n := \boldsymbol{I} + \frac{\nabla U^n \otimes \nabla U^n}{1 + |\nabla U^n|^2}. \tag{4.224}$$

Then the following bound holds

$$\left\|f - \boldsymbol{A}^n{:}\mathrm{D}^2 U^{n+1}\right\|_{-1} \leq C\left(\sum_{K\in\mathscr{T}} h_K \left\|\mathcal{R}_n[U^{n+1}]\right\|_{\mathrm{L}_2(K)} + \sum_{S\in\mathscr{S}} h_K^{1/2}\left\|\mathcal{J}_n[U^{n+1}]\right\|_{\mathrm{L}_2(S)}\right), \tag{4.225}$$

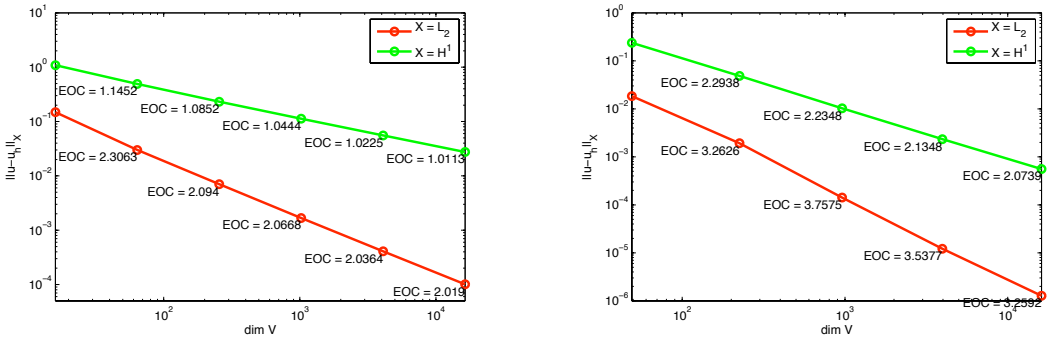
where the interior residual over a simplex K and jump residual over a common wall S of two simplexes, K^+ and K^- are defined as

$$\left\|\mathcal{R}_n[U^{n+1}]\right\|^2_{\mathrm{L}_2(K)} = \int_K \left(f - \boldsymbol{A}^n{:}\mathrm{D}^2 U^{n+1}\right)^2, \tag{4.226}$$

$$\left\|\mathcal{J}_n[U^{n+1}]\right\|^2_{\mathrm{L}_2(S)} = \int_S \left(-\boldsymbol{A}^n{:}\left(\nabla U^{n+1}\big|_{K^+} \otimes \boldsymbol{n}_{K^+}\right) - \boldsymbol{A}^n{:}\left(\nabla U^{n+1}\big|_{K^-} \otimes \boldsymbol{n}_{K^-}\right)\right)^2 \tag{4.227}$$

with $\boldsymbol{n}_{K^+}$ and $\boldsymbol{n}_{K^-}$ denoting the outward pointing normals to K^+ and K^- respectively.

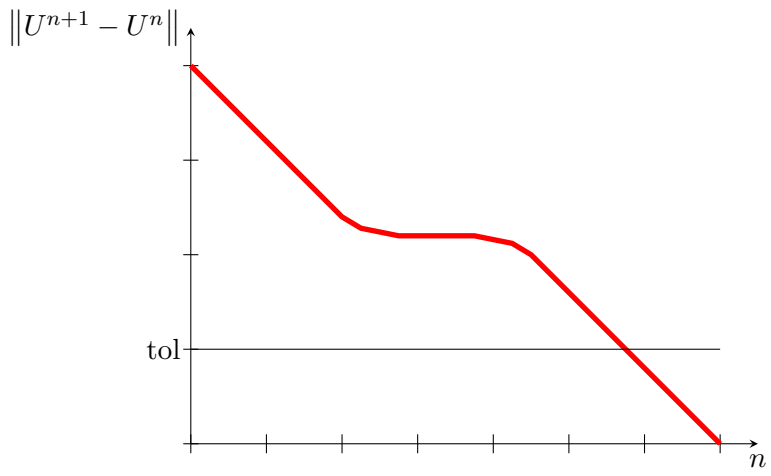
Figure 4.15: Test 4.8. Errors and convergence rates for NVFEM applied to (4.216), a quasilinear PDE under a fixed point linearisation. We fix f appropriately such that $u(\boldsymbol{x}) = \sin(\pi x_1)\sin(\pi x_2)$, taking an initial guess $u^0 = 0$. The convergence rates are optimal, that is for $\mathbb{P}^1$-elements (on the left) $\|u - U\| = \mathrm{O}(h^2)$ and $|u - U|_1 = \mathrm{O}(h)$. For $\mathbb{P}^2$-elements (on the right) $\|u - U\| = \mathrm{O}(h^3)$ and $|u - U|_1 = \mathrm{O}(h^2)$.



(a) $\mathbb{P}^1$-elements (b) $\mathbb{P}^2$-elements

4.8.1.2 Remark (lower bounds to the residual)**.** This estimator is an upper bound but may not be a lower bound. In the case of the quasilinear mean curvature flow (and in fact general nonlinear operators) the linearised operator may become degenerate, i.e., in this case $|\nabla u^n| \to \infty$. Hence the estimator may prove unreliable. An idea motivated by Veeser [FV03] is to add contributions to the estimator that are functions of the element area.

4.8.1.3 Remark (stopping criterion for the linearisation)**.** In our numerical experiments we use $\left\|U^{n+1} - U^n\right\| \leq \mathrm{tol}$ as a stopping criterion. This however does not guarantee the method has converged to the fixed point. Indeed if the linearisation were to stagnate as depicted in the following sketch



the algorithm would terminate prematurely.

A better approach would be to use the relative error of the aposteriori residual. That is

$$\frac{\mathscr{E}[U^n]}{\mathscr{E}[U^0]} \leq \text{tol}\,. \tag{4.228}$$

This guarantees a reduction in error (due to the upper bound of $\mathscr{E}$ from 4.8.1.1).

Chapter 5

A numerical method for second order fully nonlinear elliptic PDEs

In this chapter we will present a novel method for the approximation of fully nonlinear second order elliptic PDEs which do not have constraints. This is a Newton linearisation together with the finite element method developed in §4 for use in nonvariational elliptic problems.

In the case that the PDE does have constraints we will illustrate the difficulties in passing these down to the discrete level by considering the Monge–Ampère equation as a model.

We will make use of the definition of *finite element convexity* from [AM09] and the concept of semidefinite programming [VB96] to enforce convexity on the discrete solution at each Newton iterate.

5.0.1.1 Definition (fully nonlinear PDE)**.** A PDE is *fully nonlinear* if it is nonlinear with respect to its highest order derivative, i.e.,

$$\mathscr{N}[u] = F(D^2u, \nabla u, u, \boldsymbol{x}) - f(\boldsymbol{x}) = 0 \tag{5.1}$$

where $F : \mathrm{Sym}^+(\mathbb{R}^{d\times d}) \times \mathbb{R}^d \times \mathbb{R} \times \overline{\Omega} \to \mathbb{R}$ is nonlinear with respect to its first argument.

The difficulty for fully nonlinear equations is in dealing with the highest order term. We will not deal with first and zeroth order terms here and we restrict F to be a function of its first argument, that is

$$\mathscr{N}[u] = F(D^2u) - f(\boldsymbol{x}) = 0. \tag{5.2}$$

Recall from §4.1.2 a function, $u \in \mathrm{C}^2(\overline{\Omega})$, is *a classical solution* of (5.2) if the problem is satisfied pointwise. In a similar context to that of linear equations we call a function that is a twice weakly differentiable a *strong solution* of (5.2) if it is satisfied almost everywhere in Ω.

5.0.1.2 Definition (ellipticity [CC95])**.** The problem (5.2) is said to be uniformly elliptic if for any $\boldsymbol{M} \in \mathrm{Sym}^+(\mathbb{R}^{d\times d})$ there exist *ellipticity constants* $\lambda, \Lambda > 0$ such that:

$$\lambda \sup_{|\boldsymbol{x}|=1} |\boldsymbol{N}\boldsymbol{x}| \leq F(\boldsymbol{M}+\boldsymbol{N}) - F(\boldsymbol{M}) \leq \Lambda \sup_{|\boldsymbol{x}|=1} |\boldsymbol{N}\boldsymbol{x}| \quad \forall\, \boldsymbol{N} \in \mathrm{Sym}^+(\mathbb{R}^{d\times d}) \tag{5.3}$$

Another more familiar way of considering (5.3) is via the derivative of F. If F is differentiable at a point $\boldsymbol{R} \in \mathrm{Sym}^+(\mathbb{R}^{d\times d})$ we can write

$$\boldsymbol{R} = \begin{bmatrix} r_{1,1} & \cdots & r_{1,d} \\ \vdots & \ddots & \vdots \\ r_{d,1} & \cdots & r_{d,d} \end{bmatrix} \tag{5.4}$$

and the derivative of F is a matrix valued function

$$F'(\boldsymbol{R}) = \begin{bmatrix} \frac{\partial F(\boldsymbol{R})}{\partial r_{1,1}} & \cdots & \frac{\partial F(\boldsymbol{R})}{\partial r_{d,1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial F(\boldsymbol{R})}{\partial r_{1,d}} & \cdots & \frac{\partial F(\boldsymbol{R})}{\partial r_{d,d}} \end{bmatrix}. \tag{5.5}$$

Then if there exists a constant $\lambda > 0$ such that

$$\boldsymbol{\xi}^\intercal F'(\boldsymbol{N})\boldsymbol{\xi} \geq \lambda\, |\boldsymbol{\xi}|^2 \quad \forall\, \boldsymbol{N} \in \mathrm{Sym}^+(\mathbb{R}^{d\times d}), \boldsymbol{\xi} \in \mathbb{R}^d \tag{5.6}$$

the problem (5.2) is elliptic.

5.0.1.3 Definition (The action of $F'(\boldsymbol{R})$ on $\boldsymbol{M}$)**.** If F is differentiable, the action of the derivative $F'(\boldsymbol{R})$ on an increment $\boldsymbol{M}$ is defined as

$$F'(\boldsymbol{R})\boldsymbol{M} := F'(\boldsymbol{R}){:}\boldsymbol{M}. \tag{5.7}$$

This is crucial in defining Newton's method correctly (cf. §5.1.1).

5.1 On the linearisation of fully nonlinear problems

In this work we will study Newton's method, although noting that fixed point methods can be used due to the relation between fully nonlinear problems and nonvariational problems as characterised in the following remark.

5.1.0.4 Remark (fixed point methods)**.** We can rewrite (5.2) into a more familiar form (a Frobenius product), the fixed point linearisation then follows from this. By the chain rule and the fundamental theorem of calculus

$$\mathscr{N}[u] = \left[\int_0^1 F'(t\mathrm{D}^2u)\,\mathrm{d}t\right]:\mathrm{D}^2u + F(\mathbf{0}) - f = 0. \tag{5.8}$$

Setting

$$\boldsymbol{N}(\mathrm{D}^2u) = \int_0^1 F'(t\mathrm{D}^2u)\,\mathrm{d}t, \tag{5.9}$$

$$g = f - F(\mathbf{0}), \tag{5.10}$$

then if u solves (5.2), it also solves

$$\boldsymbol{N}(\mathrm{D}^2u):\mathrm{D}^2u = g. \tag{5.11}$$

A fixed point method would then consist in: finding a sequence $(u^n)_{n\in\mathbb{N}_0}$ such that for each $n \in \mathbb{N}_0$

$$\boldsymbol{N}(\mathrm{D}^2u^n):\mathrm{D}^2u^{n+1} = g, \tag{5.12}$$

with u^0 given.

5.1.1 Newton's method

Given an initial guess u^0, we define the Newton step for (5.2) as: For $n \in \mathbb{N}_0$ find u^{n+1} such that:

$$\mathscr{N}'[u^n]\left(u^{n+1} - u^n\right) = -\mathscr{N}[u^n]. \tag{5.13}$$

Rewriting it in terms of the nonlinear operator.

$$\begin{aligned}\mathscr{N}'[u]v &= \lim_{\epsilon\to 0}\frac{\mathscr{N}[u+\epsilon v] - \mathscr{N}[u]}{\epsilon}\\ &= \lim_{\epsilon\to 0}\frac{F(\mathrm{D}^2u + \epsilon\mathrm{D}^2v) - F(\mathrm{D}^2u)}{\epsilon}\\ &= F'(\mathrm{D}^2u) : \mathrm{D}^2v\end{aligned} \tag{5.14}$$

Combining (5.13) and (5.14) then results in the following nonvariational sequence of linear PDEs. Given u^0 for each $n \in \mathbb{N}$ find u^{n+1} such that

$$F'(\mathrm{D}^2 u^n) : \mathrm{D}^2 \left(u^{n+1} - u^n\right) = f - F(\mathrm{D}^2 u^n). \tag{5.15}$$

5.1.1.1 Remark (constraints)**.** Many fully nonlinear elliptic PDEs must be constrained in order to admit a unique solution. For example the Monge–Ampère–Dirichlet (MAD) problem

$$\begin{aligned} \det \mathrm{D}^2 u &= f \qquad \text{in } \Omega \\ u &= 0 \qquad \text{on } \partial\Omega \end{aligned} \tag{5.16}$$

admits a unique solution in the cone of convex functions (see A.3). We will study the MAD problem in further detail in §5.4.

Due to difficulties arising from the passing of these constraints from the continuous level down to the discrete we will initially study fully nonlinear PDEs which have no such constraint.

5.2 Unconstrained fully nonlinear PDEs

Before we discretise the model problem (5.2) in general we will illustrate the method we will propose with a simple example. For this we follow the framework set out in §4.

5.2.0.2 Example (a simple fully nonlinear PDE)**.** We consider the problem

$$\begin{aligned} \mathscr{N}[u] := |\Delta u| + 2\Delta u - f = 0 \text{ in } \Omega \\ u = 0 \quad \text{on } \partial\Omega \end{aligned} \tag{5.17}$$

which is specifically constructed to be uniformly elliptic. Indeed

$$F'(\mathrm{D}^2 u) = (\operatorname{sign}(\Delta u) + 2)\, \boldsymbol{I} > 0. \tag{5.18}$$

The Newton linearisation of the problem is then: Given u^0 find u^{n+1} such that

$$(\operatorname{sign}(\Delta u^n) + 2)\, \boldsymbol{I} : \mathrm{D}^2(u^{n+1} - u^n) = f - |\Delta u^n| - 2\Delta u^n. \tag{5.19}$$

Recall the NVFEM was set up in such a way that the finite element Hessian was given as part of the solution process (see (5.41)). With that in mind we may in fact use the finite

element Hessian of the previous Newton iterate in the discrete formulation as follows: Find $U^{n+1} \in \mathbb{V}$ such that

$$\left\langle (\operatorname{sign}(\Delta_h U^n) + 2)\, \boldsymbol{I}{:}\boldsymbol{H}[U^{n+1} - U^n], \mathring{\Phi} \right\rangle = \left\langle f - |\Delta_h U^n| - 2\Delta_h U^n, \mathring{\Phi} \right\rangle \quad \forall\, \mathring{\Phi} \in \mathring{\mathbb{V}} \tag{5.20}$$

where we are using $\Delta_h v = \operatorname{trace} \boldsymbol{H}[v]$.

We now present the method for general unconstrained fully nonlinear PDEs.

5.2.0.3 Definition (nonlinear finite element method (NLFEM))**.** Given a BVP of the form, finding $u \in \mathrm{H}^2(\Omega) \cap \mathrm{H}_0^1(\Omega)$ such that

$$\begin{aligned} \mathscr{N}[u] = F(\mathrm{D}^2 u) - f = 0 \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega. \end{aligned} \tag{5.21}$$

Upon applying Newton's method to solve problem (5.21) we obtain a sequence of functions $(u^n)_{n\in\mathbb{N}_0}$ solving the following linear equations in nonvariational form

$$\boldsymbol{N}(\mathrm{D}^2 u^n){:}\mathrm{D}^2 u^{n+1} = g(\mathrm{D}^2 u^n) \tag{5.22}$$

where

$$\boldsymbol{N}(\boldsymbol{X}) := F'(\boldsymbol{X}), \tag{5.23}$$

$$g(\boldsymbol{X}) := f - F(\boldsymbol{X}) + F'(\boldsymbol{X}){:}\boldsymbol{X}. \tag{5.24}$$

Recall the notation from §4, the finite element space, $\mathring{\mathbb{V}}$, is defined in (4.10). The finite element Hessian, $\boldsymbol{H}$, is given in Definition 4.1.3.4.

The nonlinear finite element method to approximate (5.22) is defined as finding $(U^n)_{n\in\mathbb{N}_0} \in \mathring{\mathbb{V}}$ such that

$$\left\langle \boldsymbol{N}(\boldsymbol{H}[U^n]){:}\boldsymbol{H}[U^{n+1}], \mathring{\Phi} \right\rangle = \left\langle g(\boldsymbol{H}[U^n]), \mathring{\Phi} \right\rangle \quad \forall\, \mathring{\Phi} \in \mathring{\mathbb{V}}. \tag{5.25}$$

Given $U^{n+1} = \mathring{\boldsymbol{\Phi}}^{\mathsf{T}} \mathring{\mathbf{u}}^{n+1}$ the equivalent linear system is given by

$$\mathring{\mathring{\mathbf{D}}}^n \mathbf{u}^{n+1} := \sum_{\alpha=1}^{d} \sum_{\beta=1}^{d} \mathring{\mathbf{B}}_n^{\alpha\beta} \mathbf{M}^{-1} \mathring{\mathbf{C}}_{\alpha\beta} \mathring{\mathbf{u}}^{n+1} = \mathring{\mathbf{g}}^n. \tag{5.26}$$

The problem dependent components of (5.26) are given by

$$\mathring{\mathbf{B}}_n^{\alpha\beta} := \left\langle \mathring{\boldsymbol{\Phi}}, \boldsymbol{N}(\boldsymbol{H}[U^n])^{\alpha,\beta} \boldsymbol{\Phi}^{\mathsf{T}} \right\rangle \tag{5.27}$$

$$\mathring{\mathbf{g}}^n := \left\langle g(\boldsymbol{H}[U^n]), \mathring{\boldsymbol{\Phi}} \right\rangle. \tag{5.28}$$

We now give an algorithm for the general method.

5.2.1 The NVFEM for fully nonlinear problems

Require: $(\mathscr{T}_0, u^0, p, N, K_{\max}, \mathrm{tol})$

Ensure: $(U, \mathbb{V})$ the NVFE approximation of (4.216)

$k = 0$

while $k \leq K_{\max}$ **do**

$\mathbb{V}_k = \mathsf{Fe\ Space}(\mathscr{T}_k, p)$

if $k = 0$ **then**

$U^0 = \Lambda^{\mathbb{V}_0} u^0$

$\boldsymbol{H}[U^0] = \mathsf{Hessian\ Recovery}(U^0, \mathbb{V}_0)$

end if

$n = 0$

while $n \leq N$ **do**

$[U^{n+1}, \boldsymbol{H}[U^{n+1}]] = \mathsf{NVFEM}(\mathbb{V}_k, \boldsymbol{N}(\boldsymbol{H}[U^n]), f)$

if $\left\|U^{n+1} - U^n\right\| \leq \mathrm{tol}$ **then**

break

end if

$n = n + 1$

end while

$\mathscr{T}_{k+1} = \mathsf{Global\ Refine}(\mathscr{T}_k)$

$k = k + 1$

end while

5.2.1.1 Remark (quasilinear problems)**.** This method (Definition 5.2.0.3 and Algorithm 5.2.1) is reminiscent to that of the quasilinear problems in §4.8 (specifically Algorithm 4.8.1) with the added complication of dealing with the finite element Hessian.

In a general case if we apply a Newton linearisation to the quasilinear problem (4.213) the result is a sequence of nonvariational linear PDEs whose problem coefficients depend on the Hessian of the previous iterate as in the fully nonlinear case. This method further generalises that proposed in §4.8 to general quasilinear PDEs using Newton's method.

5.3 Examples

In this section we utilise the method proposed in Theorem 5.2.0.3 for some simple fully nonlinear equations without constraints.

5.3.0.2 Example. The problem under consideration here is

$$\begin{aligned} \mathscr{N}[u] := \sin(\Delta u) + 2\Delta u - f = 0 \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega. \end{aligned} \tag{5.29}$$

Computing Newton's Method (5.13) in this case is straightforward. It is easily verified that the derivative

$$\begin{aligned} \mathscr{N}'[u]\, v &= \lim_{\eta\to 0} \frac{1}{\eta}\Big(\mathscr{N}[u+\eta v] - \mathscr{N}[u] \Big) \\ &= \lim_{\eta\to 0} \frac{1}{\eta}\Big(\sin(\Delta u + \eta\Delta v) + 2\Delta u + 2\eta\Delta v - \sin(\Delta u) - 2\Delta u \Big) \\ &= \lim_{\eta\to 0} \frac{1}{\eta}\left(2\cos\left(\frac{2\Delta u + \eta\Delta v}{2}\right)\sin\left(\frac{\eta\Delta v}{2}\right) + 2\eta\Delta v \right) \\ &= \lim_{\eta\to 0} \left(\cos\left(\frac{2\Delta u + \eta\Delta v}{2}\right)\frac{\sin\left(\frac{\eta\Delta v}{2}\right)}{\frac{\eta\Delta v}{2}}\Delta v + 2\Delta v \right) \\ &= \cos(\Delta u)\,\Delta v + 2\Delta v. \end{aligned} \tag{5.30}$$

Setting

$$\boldsymbol{N}(\boldsymbol{X}) = (\cos(\operatorname{trace}\boldsymbol{X}) + 2)\,\boldsymbol{I}, \tag{5.31}$$

$$g(\boldsymbol{X}) = f - \sin(\operatorname{trace}\boldsymbol{X}) + \cos(\operatorname{trace}\boldsymbol{X})\operatorname{trace}\boldsymbol{X}. \tag{5.32}$$

Newton's method applied to this problem then reads: Given u^0 for each $n \in \mathbb{N}$ find u^{n+1} such that

$$\boldsymbol{N}(\mathrm{D}^2u^n){:}\mathrm{D}^2u^{n+1} = g(\mathrm{D}^2u^n). \tag{5.33}$$

Using the discretisation from (5.26) given $U^0 \in \mathring{\mathbb{V}}$ for each $n = 1, \ldots, M$ we seek $U^{n+1} \in \mathring{\mathbb{V}}$ such that

$$\left\langle \boldsymbol{N}(\boldsymbol{H}[U^n]){:}\boldsymbol{H}[U^{n+1}], \mathring{\Phi} \right\rangle = \left\langle g(\boldsymbol{H}[U^n]), \mathring{\Phi} \right\rangle \quad \forall\, \mathring{\Phi} \in \mathring{\mathbb{V}}. \tag{5.34}$$

The equivalent linear system can be posed with $U^{n+1} = \mathring{\boldsymbol{\Phi}}^{\mathsf{T}}\mathring{\mathbf{u}}^{n+1}$ such that $\mathring{\mathbf{u}}^{n+1} \in \mathbb{R}^{\mathring{N}}$ is the solution to the following linear system

$$\mathring{\mathring{\mathbf{D}}}^n\mathring{\mathbf{u}}^{n+1} := \left(\mathring{\mathbf{B}}_n^{11}\mathbf{M}^{-1}\mathring{\mathbf{C}}_{11} + \mathring{\mathbf{B}}_n^{22}\mathbf{M}^{-1}\mathring{\mathbf{C}}_{22}\right)\mathring{\mathbf{u}}^{n+1} = \mathring{\mathbf{g}}^n. \tag{5.35}$$

The problem dependent components of (5.35) are given by

$$\mathring{\mathbf{B}}_n^{\alpha\alpha} := \left\langle \mathring{\mathbf{\Phi}}, \left(\cos\left(\Delta_h U^n\right) + 2\right)\mathbf{\Phi}^{\intercal}\right\rangle \tag{5.36}$$

$$\mathring{\mathbf{g}}^n := \left\langle f - \sin\left(\Delta_h U^n\right) - 2\Delta_h U^n + \left(\cos\left(\Delta_h U^n\right) + 2\right)\boldsymbol{I}\text{:}\boldsymbol{H}[U^n], \mathring{\mathbf{\Phi}}\right\rangle. \tag{5.37}$$

Hence from Lemma 4.2.1.1, given

$$\mathbf{E} = \begin{bmatrix} \mathbf{M} & \mathbf{0} & -\mathring{\mathbf{C}}_{11} \\ \mathbf{0} & \mathbf{M} & -\mathring{\mathbf{C}}_{22} \\ \mathring{\mathbf{B}}_n^{11} & \mathring{\mathbf{B}}_n^{22} & \mathbf{0} \end{bmatrix}. \tag{5.38}$$

$$\mathbf{v} = \left[\mathbf{h}_{1,1}, \mathbf{h}_{2,2}, \mathring{\mathbf{u}}^{n+1}\right]^{\intercal}, \tag{5.39}$$

$$\mathbf{b} = \left[\mathbf{0}, \mathbf{0}, \mathring{\mathbf{g}}^n\right]^{\intercal}, \tag{5.40}$$

solving the system (5.35) is equivalent to solving

$$\mathbf{E}\mathbf{v} = \mathbf{b} \tag{5.41}$$

for $\mathring{\mathbf{u}}^{n+1}$. Figure 5.1 shows numerical results for the problem.

5.3.0.3 Example. In this example we return to the problem first presented in Example 5.2.0.2

$$\begin{aligned} \mathscr{N}[u] := |\Delta u| + 2\Delta u - f = 0 \text{ in } \Omega, \\ u = 0 \text{ on } \partial\Omega. \end{aligned} \tag{5.42}$$

Recall the discrete linearised problem is : Given U^0 find U^{n+1} such that

$$\left\langle \boldsymbol{N}(\boldsymbol{H}[U^n])\text{:}\boldsymbol{H}[U^{n+1}], \mathring{\Phi}\right\rangle = \left\langle g(\boldsymbol{H}[U^n]), \mathring{\Phi}\right\rangle \quad \forall\, \mathring{\Phi} \in \mathring{\mathbb{V}}, \tag{5.43}$$

where

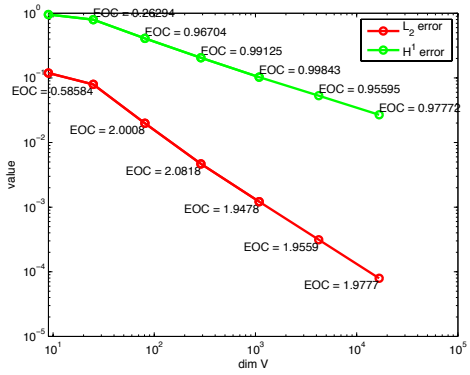
$$\boldsymbol{N}(\boldsymbol{H}[U^n]) = \left(\operatorname{sign}\left(\Delta_h U^n\right) + 2\right)\boldsymbol{I} \tag{5.44}$$

$$g(\boldsymbol{H}[U^n]) = f - |\Delta_h U^n| - 2\Delta_h U^n. \tag{5.45}$$

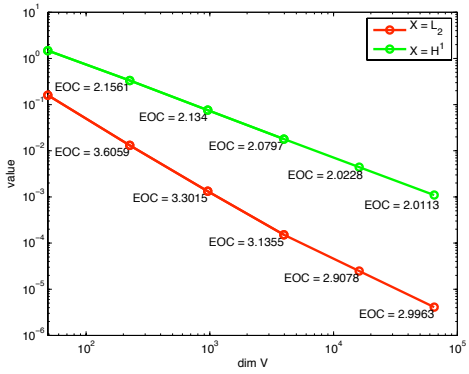
We may write the problem as a linear system using the same methodology as in the previous example.

Figure 5.2 shows numerical results for this problem.

Figure 5.1: Numerical convergence rates for problem (5.29). Choosing f appropriately such that $u(\boldsymbol{x}) = \sin(\pi x_1)\sin(\pi x_2)$. We run Algorithm 5.2.1 with an initial guess $u^0 = 0$ until $\left\|U^{n+1} - U^n\right\| \leq 10^{-3}$ upon each Newton step.

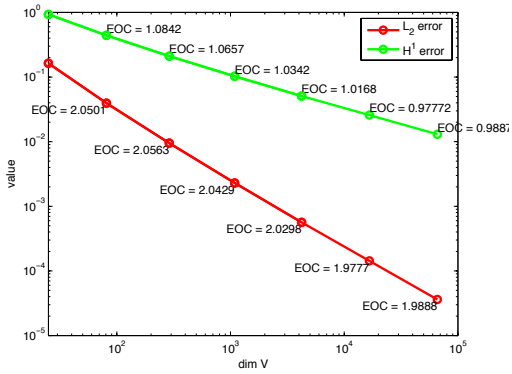


(a) Taking $\mathbb{V}$ to be the space of piecewise linear functions on Ω.

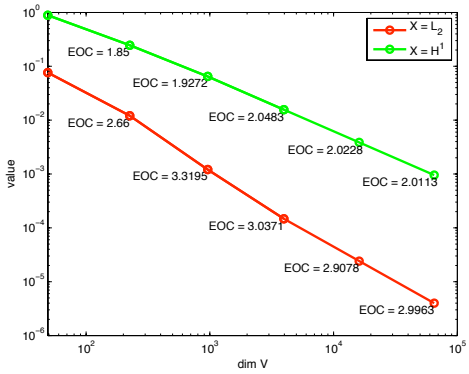


(b) Taking $\mathbb{V}$ to be the space of piecewise quadratic functions on Ω.

Figure 5.2: Numerical convergence rates for problem (5.42). Choosing f appropriately such that $u(\boldsymbol{x}) = \exp\left(-10\left|\boldsymbol{x}\right|^2\right)$.We run Algorithm 5.2.1 with an initial guess $u^0 = 0$ until $\left\|U^{n+1} - U^n\right\| \leq 10^{-3}$ upon each Newton step.



(a) Taking $\mathbb{V}$ to be the space of piecewise linear functions on Ω.



(b) Taking $\mathbb{V}$ to be the space of piecewise quadratic functions on Ω.

5.3.0.4 Example. In this example we look at an example of a uniformly elliptic fully nonlinear PDE whose solution is unknown [Kry95].

The problem is for $d = 2$

$$\begin{aligned} \mathscr{N}[u] := (\partial_{11}u)^2 + (\partial_{22}u)^2 + \frac{5}{2}\partial_{11}u\partial_{22}u - \frac{1}{2} = 0 \quad &\text{in } \Omega \\ u = 0 \quad &\text{on } \partial\Omega. \end{aligned} \tag{5.46}$$

The Newton linearisation of the problem is then: Given u^0 find $(u^n)_{n\in\mathbb{N}_0}$ such that

$$\boldsymbol{N}(\mathrm{D}^2u^n){:}\mathrm{D}^2\left(u^{n+1} - u^n\right) = g(\mathrm{D}^2u^n), \tag{5.47}$$

where in this case

$$\boldsymbol{N}(\mathrm{D}^2u^n) := \begin{bmatrix} 2\partial_{11}u^n + \frac{5}{2}\partial_{22}u^n & 0 \\ 0 & \frac{5}{2}\partial_{11}u^n + 2\partial_{22}u^n \end{bmatrix} \tag{5.48}$$

$$g(\mathrm{D}^2u^n) := \frac{1}{2} - (\partial_{11}u^n)^2 - (\partial_{22}u^n)^2 - \frac{5}{2}\partial_{11}u^n\partial_{22}u^n. \tag{5.49}$$

We discretise problem (5.50) under the same framework as the previous examples, that is, given U^0 we seek the sequence $(U^n)_{n\in\mathbb{N}_0}$ such that

$$\left\langle \boldsymbol{N}(\boldsymbol{H}[U^n]){:}\boldsymbol{H}\left[U^{n+1} - U^n\right], \mathring{\Phi}\right\rangle = \left\langle g(\boldsymbol{H}[U^n]), \mathring{\Phi}\right\rangle \quad \forall\, \mathring{\Phi} \in \mathring{\mathbb{V}}. \tag{5.50}$$

Since the problem (5.50) resembles the equation

$$(\Delta u)^2 = 1/2 \tag{5.51}$$

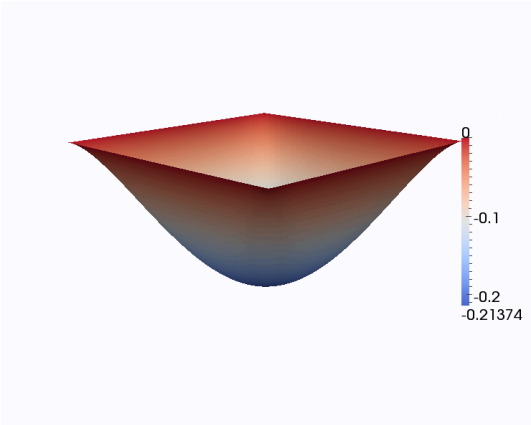
we take the initial guess U^0 to be the finite element approximation to $\Delta u = 1/\sqrt{2}$. We fix $h = \sqrt{2}/64$ and show the NLFE solution, $(U^n)_{n\in[0:3]}$ at each Newton step. The surface plots of U^n with $n = 0, \ldots, 3$ are given in Figure 5.3.

5.4 Constrained fully nonlinear PDEs - the Monge–Ampère equation

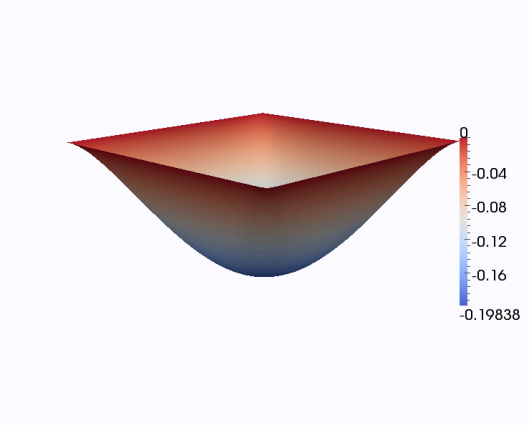
The Monge–Ampère equation is an important example of a fully nonlinear elliptic PDE since it is used as a model for other fully nonlinear PDEs. It is derived from differential geometry. The problem is

$$\begin{aligned} \mathscr{N}[u] := \det \mathrm{D}^2u - f = 0 \text{ in } \Omega \\ u = 0 \text{ on } \partial\Omega. \end{aligned} \tag{5.52}$$

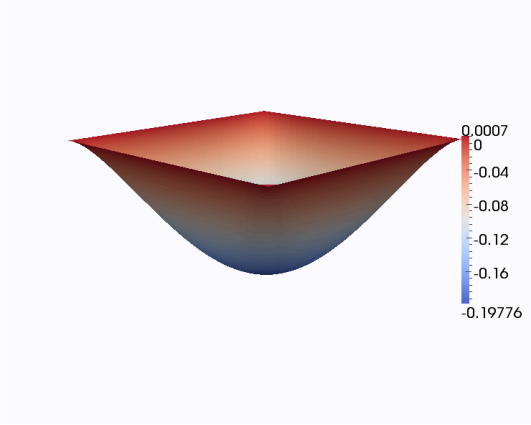
Figure 5.3: The solution of problem 5.46 under a Newton linearisation. We show the initial guess and the first three Newton iterates.



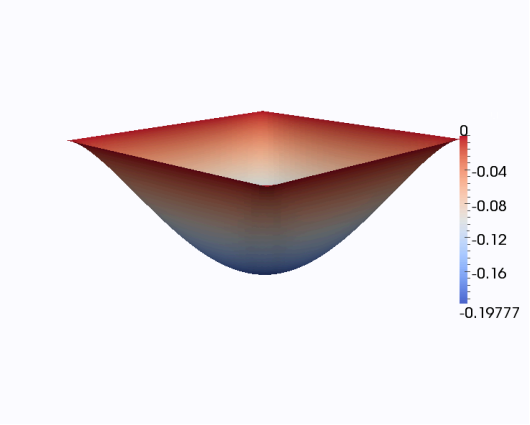
(a) The initial guess U^0



(b) The first Newton step U^1



(c) The second Newton step U^2



(d) The third Newton step U^3

This equation is clearly fully nonlinear for $d \geq 2$, in fact it is multi-linear with respect to columns (or rows) of the Hessian. This makes for simpler computations when it comes to linearising the problem.

There have been numerical studies on this problem in the context of finite differences by Oliker and Prussner [OP88]. Feng and Neilan [FN07, FN08b, FN08a] propose a mixed finite element method using sequences of fourth order quasi-linear equations. Despite these works this is still a tricky problem to formulate correctly. There are restrictions that must be put in place in order to guarantee ellipticity for example.

We require Ω to be convex and $f > 0$ for a classical solution to even exist. Monge–Ampère (5.52) will be uniformly elliptic if $\mathrm{D}^2 u$ is positive definite. If these restrictions are then satisfied then (5.52) admits a unique convex viscosity solution.

Indeed without the constraint on the Hessian of the solution the problem admits two solutions, one convex and one concave, in $d = 2$ for example with homogeneous Dirichlet boundary both u and $-u$ solve (5.52).

5.4.1 Newton's method applied to Monge–Ampère

In view of the characteristic expansion of determinant if $\boldsymbol{X}, \boldsymbol{Y} \in \mathrm{Sym}^+(\mathbb{R}^{d\times d})$ then denoting $\mathrm{Cof}\, \boldsymbol{X}$ to be the matrix of cofactors of $\boldsymbol{X}$

$$\det(\boldsymbol{X} + \epsilon \boldsymbol{Y}) = \det \boldsymbol{X} + \epsilon \,\mathrm{trace}\,(\mathrm{Cof}(\boldsymbol{X})\boldsymbol{Y}) + \cdots + \epsilon^d \det \boldsymbol{Y} \tag{5.53}$$

and thus

$$\mathscr{N}'[u]\, v = \mathrm{Cof}\, \mathrm{D}^2 u{:}\mathrm{D}^2 v. \tag{5.54}$$

We set

$$\boldsymbol{N}(\mathrm{D}^2 u^n) = \mathrm{Cof}\, \mathrm{D}^2 u^n, \tag{5.55}$$

$$g(\mathrm{D}^2 u^n) = \delta(f - \det \mathrm{D}^2 u^n) + \mathrm{Cof}\, \mathrm{D}^2 u^n{:}\mathrm{D}^2 u^n, \tag{5.56}$$

where $\delta \in (0, 1]$ is a Newton damping factor which although practically is always taken as $\delta = 1$ is useful in the proof of Lemma 5.5.0.4.

5.4.1.1 Remark (relating cofactors to determinants)**.** Note that for a generic function v it holds that

$$d \det \mathrm{D}^2 v = \mathrm{Cof}\, \mathrm{D}^2 v{:}\mathrm{D}^2 v. \tag{5.57}$$

Using this formulation we could construct a simple fixed point method for the Monge–Ampère equation.

Note that in view of Remark 5.4.1.1 g can be further simplified

$$\begin{aligned} g(\mathrm{D}^2u^n) &= \delta(f - \det \mathrm{D}^2u^n) + \operatorname{Cof} \mathrm{D}^2u^n{:}\mathrm{D}^2u^n \\ &= \delta(f - \det \mathrm{D}^2u^n) + d \det \mathrm{D}^2u^n \\ &= \delta\left(f + \left(\frac{d}{\delta} - 1\right) \det \mathrm{D}^2u^n\right). \end{aligned} \tag{5.58}$$

Newton's method reads: Given u^0 for each $n \in \mathbb{N}$ find u^{n+1} such that

$$\boldsymbol{N}(\mathrm{D}^2u^n){:}\mathrm{D}^2u^{n+1} = g(\mathrm{D}^2u^n). \tag{5.59}$$

We are going to study this problem and demonstrate the difficulties in passing the preservation of convexity to the discrete level.

The first step to this end is to study the properties of the linearised Monge–Ampère equation at the continuous level.

5.5 Monge–Ampère at the continuous level

5.5.0.2 Theorem (regularity estimate for the Monge–Ampère equation [Caf90, Theorem 2])**.** *Let $u : \Omega \to \mathbb{R}$ be the classical solution of the Monge–Ampère equation (5.52). Assume further that the data function f satisfies $\int_\Omega f = 1$ and that $f > 0$ and bounded then there exists a $C > 0$ such that*

$$\|u\|_{\mathrm{C}^{2,\alpha}(\Omega)} \leq C \|f\|_{\mathrm{C}^{\alpha}(\Omega)}. \tag{5.60}$$

5.5.0.3 Theorem ([LR05])**.** *Let the assumptions of Theorem 5.5.0.2 hold. In addition assume $(u^n)_{n\in\mathbb{N}}$ is the sequence of functions obtained by solving the linearised Monge–Ampère equation (5.59). Then for any initial guess u^0 there exists a $\delta > 0$ such that*

$$\lim_{n\to\infty} \|u^n - u\|_{\mathrm{C}^{2,\alpha'}(\Omega)} = 0 \quad \forall\, \alpha' < \alpha. \tag{5.61}$$

To prove this Theorem we make use of the following Lemmas.

5.5.0.4 Lemma. *Let the assumptions of Theorem 5.5.0.3 hold. Then for each $n \in \mathbb{N}_0$ there exist constants $C_1 > 1$ and $C_2 > 0$ such that*

$$\frac{1}{C_1} f \leq \det \mathrm{D}^2 u^n \leq C_1 f \tag{5.62}$$

$$\text{and } \left\| f - \det \mathrm{D}^2 u^n \right\|_{\mathrm{C}^\alpha(\Omega)} \leq C_2. \tag{5.63}$$

Proof We restrict ourselves to $d = 2$ for simplicity. The statements (5.62), (5.63) are proved by induction.

base case.

Assuming the initial guess is smooth ($u^0 \in C^2(\overline{\Omega})$), then there are always constants C_1, C_2 such that (5.62), (5.63) are satisfied.

inductive step.

Assume (5.62), (5.63) hold for all $k \leq n$. Let us denote $\theta^{n+1} = u^{n+1} - u^n$ then it is clear that θ^{n+1} solves the elliptic problem

$$\operatorname{Cof} \mathrm{D}^2 u^n : \mathrm{D}^2 \theta^{n+1} = \delta (f - \det \mathrm{D}^2 u^n). \tag{5.64}$$

Now

$$\begin{aligned} \det \mathrm{D}^2 u^{n+1} &= \det(\mathrm{D}^2 u^n + \mathrm{D}^2 \theta^{n+1}) \\ &= \det \mathrm{D}^2 u^n + \operatorname{Cof} \mathrm{D}^2 u^n : \mathrm{D}^2 \theta^{n+1} + \det \mathrm{D}^2 \theta^{n+1} \\ &= \det \mathrm{D}^2 u^n + \delta \left(f - \det \mathrm{D}^2 u^n \right) + \det \mathrm{D}^2 \theta^{n+1}. \end{aligned} \tag{5.65}$$

Since θ^{n+1} solves (5.64) from the Schauder estimate given in Theorem 4.1.2.4 there exists a constant $C_3 = C_3(C_2, C_1)$ such that

$$\left\| \mathrm{D}^2 \theta^{n+1} \right\|_{\mathrm{C}^\alpha(\Omega)} \leq C_3 \delta \left\| f - \det \mathrm{D}^2 u^n \right\|_{\mathrm{C}^\alpha(\Omega)} \tag{5.66}$$

and (5.66) shows there exists a constant $C_4 = C_4(C_2, C_1) = 2C_3^2$ such that

$$\left\| \det \mathrm{D}^2 \theta^{n+1} \right\|_{\mathrm{C}^\alpha(\Omega)} \leq C_4 \delta^2 \left\| f - \det \mathrm{D}^2 u^n \right\|^2_{\mathrm{C}^\alpha(\Omega)}. \tag{5.67}$$

From (5.65) we see

$$\det \mathrm{D}^2 u^{n+1} - f = (1 - \delta)(\det \mathrm{D}^2 u^n - f) + \det \mathrm{D}^2 \theta^{n+1} \tag{5.68}$$

which after combining with (5.67) gives us

$$\left\|f - \det \mathrm{D}^2 u^{n+1}\right\|_{\mathrm{C}^\alpha(\Omega)} \leq (1-\delta)\left\|f - \det \mathrm{D}^2 u^n\right\|_{\mathrm{C}^\alpha(\Omega)} + C_4\delta^2\left\|f - \det \mathrm{D}^2 u^n\right\|^2_{\mathrm{C}^\alpha(\Omega)}. \tag{5.69}$$

By the induction assumption

$$\left\|f - \det(\mathrm{D}^2 u^n)\right\|_{\mathrm{C}^\alpha(\Omega)} \leq C_2, \tag{5.70}$$

which substituting into (5.69) gives

$$\begin{aligned}\left\|f - \det \mathrm{D}^2 u^{n+1}\right\|_{\mathrm{C}^\alpha(\Omega)} &\leq \left(1 - \delta + C_2C_4\delta^2\right)\left\|f - \det \mathrm{D}^2 u^n\right\|_{\mathrm{C}^\alpha(\Omega)} \\ &\leq \kappa\left\|f - \det \mathrm{D}^2 u^n\right\|_{\mathrm{C}^\alpha(\Omega)} \\ &\leq \kappa^n\left\|f - \det \mathrm{D}^2 u^0\right\|_{\mathrm{C}^\alpha(\Omega)} \\ &\leq \left\|f - \det \mathrm{D}^2 u^0\right\|_{\mathrm{C}^\alpha(\Omega)} \iff \delta \leq \frac{1}{C_2C_4}.\end{aligned} \tag{5.71}$$

This proves the 2nd part of the Lemma (5.63) if we choose δ small enough. For the lower bound of the first part (5.62), note

$$\begin{aligned}\frac{1}{C_1}f &\leq \det \mathrm{D}^2 u^n \\ f - \det \mathrm{D}^2 u^n &\leq f(1 - \frac{1}{C_1}).\end{aligned} \tag{5.72}$$

Now, from equation (5.67), together with the fact $f \in \mathrm{L}_\infty(\Omega)$ we see there exists a constant $C_5 = C_5(C_1, C_2)$

$$\left\|\det \mathrm{D}^2\theta^{n+1}\right\|_{\mathrm{L}_\infty(\Omega)} \leq C_5\delta^2. \tag{5.73}$$

Combining (5.68) and (5.73) gives

$$\begin{aligned}f - \det \mathrm{D}^2 u^{n+1} &\leq (1-\delta)(f - \det \mathrm{D}^2 u^n) + C_5\delta^2 \\ &\leq (1-\delta)(1 - \frac{1}{C_1})f + C_5\delta^2 \\ &\leq (1 - \frac{1}{C_1})f \iff \delta < \frac{\sup_{x\in\Omega} f(x)(1 - \frac{1}{C_1})}{C_5}.\end{aligned} \tag{5.74}$$

For the upper bound of (5.62) note

$$\begin{aligned}\det \mathrm{D}^2 u^n &\leq C_1 f \\ \det \mathrm{D}^2 u^n - f &\leq f(C_1 - 1).\end{aligned} \tag{5.75}$$

Now using the same argument as in (5.74)

$$\begin{aligned} f - \det \mathrm{D}^2 u^{n+1} &\leq (1-\delta)(f - \det \mathrm{D}^2 u^n) + C_5\delta^2 \\ &\leq (1-\delta)(C_1 - 1)f + C_5\delta^2 \\ &\leq (C_1 - 1)f \iff \delta < \frac{\sup_{x\in\Omega} f(x)(C_1-1)}{C_5}. \end{aligned} \tag{5.76}$$

Hence there exists a $\delta = \delta(\sup_{x\in\Omega} f(x), C_1, C_2)$ such that (5.62) and (5.63) are satisfied for all $n \in \mathbb{N}_0$. □

Proof of Theorem 5.5.0.3 [LR05]. Recall from the proof of Lemma 5.5.0.4

$$\left\|f - \det \mathrm{D}^2 u^{n+1}\right\|_{\mathrm{C}^\alpha(\Omega)} = (1-\delta)\left\|f - \det \mathrm{D}^2 u^n\right\|_{\mathrm{C}^\alpha(\Omega)} + C_4\delta^2 \left\|f - \det \mathrm{D}^2 u^n\right\|^2_{\mathrm{C}^\alpha(\Omega)}. \tag{5.77}$$

If

$$\left\|f - \det \mathrm{D}^2 u^n\right\|_{\mathrm{C}^\alpha(\Omega)} \leq \frac{1}{2C_4\delta}, \tag{5.78}$$

then we see that

$$\begin{aligned} \left\|f - \det \mathrm{D}^2 u^{n+1}\right\|_{\mathrm{C}^\alpha(\Omega)} &\leq \frac{1-\delta}{2C_4\delta} + \frac{1}{4C_4} \\ &\leq \frac{2-\delta}{4C_4\delta}. \end{aligned} \tag{5.79}$$

Hence

$$\lim_{n\to\infty} \left\|f - \det \mathrm{D}^2 u^n\right\|_{\mathrm{C}^\alpha(\Omega)} = 0. \tag{5.80}$$

From Theorem 5.5.0.2 the sequence $\{u^n\}_{n\in\mathbb{N}}$ is bounded in $\mathrm{C}^{2,\alpha}(\Omega)$ and hence by Theorem A.1.0.9 $\{u^n\}_{n\in\mathbb{N}}$ is precompact in $\mathrm{C}^{2,\alpha'}(\Omega)$ for all $\alpha' < \alpha$. Since the solution u is unique in the set of convex functions, $\{u^n\}_{n\in\mathbb{N}} \to u$ in $\mathrm{C}^{2,\alpha}(\Omega)$. □

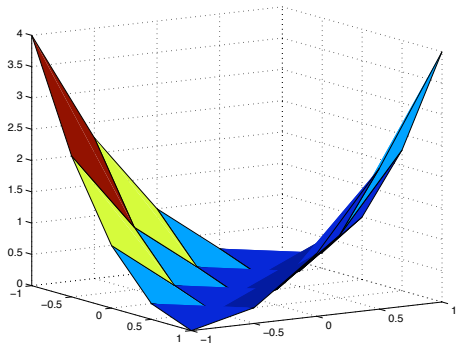
5.5.0.5 Remark. Theorem 5.5.0.2 is a generalisation of Schauder estimates to the Monge–Ampère equation. Caffarelli has also proved generalisations of Caldrón–Zygmund estimates which are of the form:

$$\|u\|_{\mathrm{H}^2(\Omega)} \leq C\|f\|. \tag{5.81}$$

With these estimates it should also be possible to prove convergence for strong solutions.

5.5.0.6 Remark (on the initial guess to the linearised Monge–Ampère)**.** Since we restrict our solution to the space of convex functions, it is prudent for the initial guess to also be convex. Moreover we must rule out constant and linear functions over Ω, since the

Figure 5.4: The Lagrange interpolant of the function $(x_1 + x_2)^2$ over a regular diagonal mesh. Notice it is NOT convex by the classical definition.



Hessian of these objects would be identically zero, destroying ellipticity on the initial Newton step. Hence we specify that the initial guess to (5.59) must be strictly convex.

5.6 Passing the constraint to the discrete level

The discretisation we initially tested was essentially the one already given in Theorem 5.2.0.3. That is given $U^0 = \Lambda u^0$ for each $n \in [0 : M]$ find $U^{n+1} \in \mathring{\mathbb{V}}$ such that

$$\left\langle \operatorname{Cof} \boldsymbol{H}[U^n] : \boldsymbol{H}[U^{n+1} - U^n], \mathring{\Phi} \right\rangle = \left\langle f - \det \boldsymbol{H}[U^n], \mathring{\Phi} \right\rangle \quad \forall\, \mathring{\Phi} \in \mathring{\mathbb{V}}. \tag{5.82}$$

However, it soon become apparent that the matrix $\operatorname{Cof} \boldsymbol{H}[U^n]$ was not remaining positive definite on the entire domain Ω and hence the problem lost ellipticity.

To overcome this problem we must enforce some form of convexity on the solution U^{n+1}.

The biggest challenge, yet to be overcome, is in what sense should the constraint of convexity be enforced in the discrete scheme?

In fact it is not obvious what is meant by convexity of a discrete function. As motivated by Morin and Aguilera [AM09, AM08] a projection or interpolation of a convex object need not be convex by the classical definition.

Figure 5.4 gives an example of this. From this is becomes clear that the notion of convexity cannot be passed in a pointwise (DOF-wise) sense.

5.6.0.7 Remark (convexity of distributions [AM09])**.** In the paper by Dudley [Dud77], it is shown that given a distribution $v \in \mathcal{D}(\Omega)$, if it holds that

$$\left\langle \mathrm{D}^2 v \,|\, \phi \right\rangle = \left\langle v \,|\, \mathrm{D}^2 \phi \right\rangle > 0 \quad \forall\, \phi \in \mathrm{C}_0^\infty(\Omega) \tag{5.83}$$

where $\phi \geq 0$ in Ω, then v is an element of the equivalence class of continuous convex functions.

From Remark 5.6.0.7 we may define a finite element convex function in a similar framework.

5.6.0.8 Definition (finite element convexity [AM09])**.** A function, v, is said to be *finite element convex* if

$$\langle \boldsymbol{H}[v], \Phi \rangle \geq 0 \quad \forall\, \Phi \in \mathbb{V} \tag{5.84}$$

where $\Phi \geq 0$ on Ω.

5.6.0.9 Remark (on the restriction $\Phi \geq 0$ on Ω.)**.** In the case of standard Lagrange finite elements, this immediately restricts us to piecewise linear finite elements. All higher order Lagrange elements attain negative values somewhere in their support.

5.6.0.10 Remark (limits of finite element convex functions)**.** In fact in [AM09] it is proven that given an indexed sequence of finite element convex functions that has a limit, said limit is convex in the classical sense as given in the Theorem 5.6.0.11.

5.6.0.11 Theorem ([AM09] Theorem 3.3)**.** *Suppose $(U_n, \mathring{\mathbb{V}}_n)_{n\in\mathbb{N}}$ is a sequence of finite element functions and spaces such that*

- *(U_n) converges weakly to a function u in $\mathrm{H}^1(\Omega)$*

$$U_i \rightharpoonup u \tag{5.85}$$

- *every finite element function is finite element convex, that is, for each $U_n \in \mathring{\mathbb{V}}_n$*

$$\langle \boldsymbol{H}[U_n], \Phi \rangle \geq 0 \quad \forall\, \Phi \in \mathbb{V}_n \tag{5.86}$$

then u is convex.

In view of Remark 5.6.0.9 we restrict $\mathbb{V}$ to be the space of piecewise linear functions on Ω. We then set

$$\mathfrak{R}[U^{n+1}] = \operatorname{Cof} \boldsymbol{H}[U^n] : \boldsymbol{H}[U^{n+1} - U^n] - f + \det \boldsymbol{H}[U^n] \tag{5.87}$$

to be the residual of the problem. The discrete formulation we become interested in is to

$$\text{minimise } \left\langle \mathfrak{R}[U^{n+1}], \mathring{\Phi} \right\rangle \quad \forall\, \mathring{\Phi} \in \mathring{\mathbb{V}}, \tag{5.88}$$

$$\text{subject to } \left\langle \boldsymbol{H}[U^{n+1}], \Phi \right\rangle \geq 0 \quad \forall\, \Phi \in \mathbb{V}. \tag{5.89}$$

This is a semidefinite programming problem [VB96]. In this case (5.88) is the objective function and (5.89) is the constraint (see §A.4).

5.7 Implementation

The implementation of the numerical scheme (5.88) and (5.89) is done in Matlab ®. The finite element component of the code is based on that already developed for the linear problems described in §4. As for the semidefinite program, we made use of the Matlab implemented SeDuMi (Self-Dual-Minimization) [Stu99].

In order to make use of SeDuMi we must first pose the semidefinite program in the form of (A.29), that is, a linear algebra problem. The equivalent system is just

$$\text{minimise } |\mathbf{E}\mathbf{v} - \mathbf{b}| \tag{5.90}$$

$$\text{subject to } [\mathbf{M}\mathbf{h}_{\alpha,\beta}]_{\alpha,\beta} > 0 \tag{5.91}$$

where recall from §4

$$\mathbf{E}\mathbf{v} - \mathbf{b} = \begin{bmatrix} \mathbf{M} & \mathbf{0} & \cdots & \mathbf{0} & \mathbf{0} & -\mathring{\mathbf{C}}_{11} \\ \mathbf{0} & \mathbf{M} & \cdots & \mathbf{0} & \mathbf{0} & -\mathring{\mathbf{C}}_{12} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \mathbf{M} & \mathbf{0} & -\mathring{\mathbf{C}}_{dd-1} \\ \mathbf{0} & \mathbf{0} & \ldots & \mathbf{0} & \mathbf{M} & -\mathring{\mathbf{C}}_{dd} \\ \mathring{\mathbf{B}}_n^{11} & \mathring{\mathbf{B}}_n^{12} & \ldots & \mathring{\mathbf{B}}_n^{dd-1} & \mathring{\mathbf{B}}_n^{dd} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \mathbf{h}_{1,1} \\ \mathbf{h}_{1,2} \\ \vdots \\ \mathbf{h}_{d,d-1} \\ \mathbf{h}_{d,d} \\ \mathring{\mathbf{u}} \end{bmatrix} - \begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \vdots \\ \mathbf{0} \\ \mathbf{0} \\ \mathring{\mathbf{f}} \end{bmatrix} \tag{5.92}$$

and we are using

$$[\mathbf{Mh}_{\alpha,\beta}]_{\alpha,\beta} = \begin{bmatrix} \mathbf{Mh}_{1,1} & \mathbf{Mh}_{1,2} & \dots & \mathbf{Mh}_{1,d} \\ \mathbf{Mh}_{2,1} & \mathbf{Mh}_{2,2} & \dots & \mathbf{Mh}_{2,d} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{Mh}_{d,1} & \mathbf{Mh}_{d,2} & \dots & \mathbf{Mh}_{d,d} \end{bmatrix} > 0, \tag{5.93}$$

as our convexity constraint of the NVFE-solution.

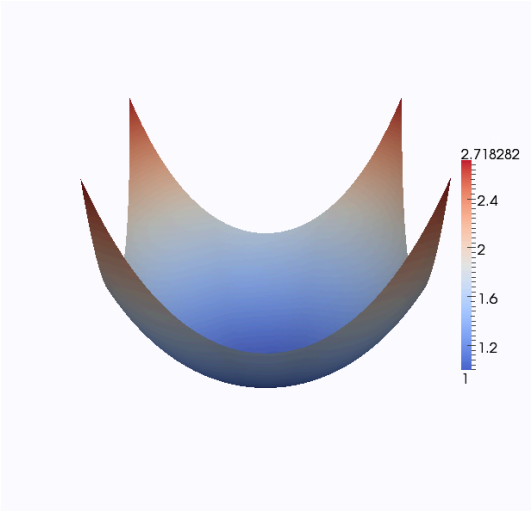
5.8 Numerical experiments

In this section we study the numerical behaviour of the scheme presented above.

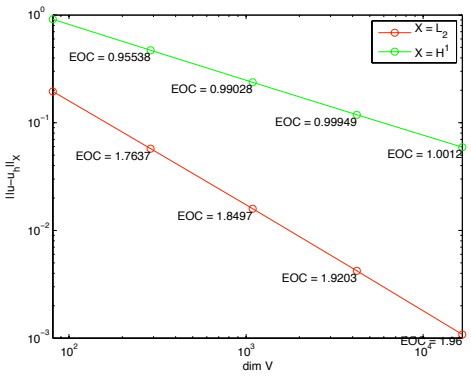
We present a set of benchmark problems constructed from the problem data such that the solution to the Monge–Ampère equation is known and construct the method presented above. We fix Ω to be the square $S = [-1, 1]^2$ or $[0, 1]^2$ and test convergence rates of the discrete solution to the exact solution.

Figures 5.5–5.8 details the various experiments and shows numerical convergence results for each of the problems studied. In each of these cases the Dirichlet boundary values are not zero. We make use of Method 1 proposed in §4.4.1 for the inhomogeneous boundaries.

Figure 5.5: Numerical results for Monge–Ampère on the square $\Omega = [-1, 1]^2$. Choosing f and g appropriately such that the solution is the radially symmetric function $u(\boldsymbol{x}) = e^{\frac{|\boldsymbol{x}|^2}{2}}$.

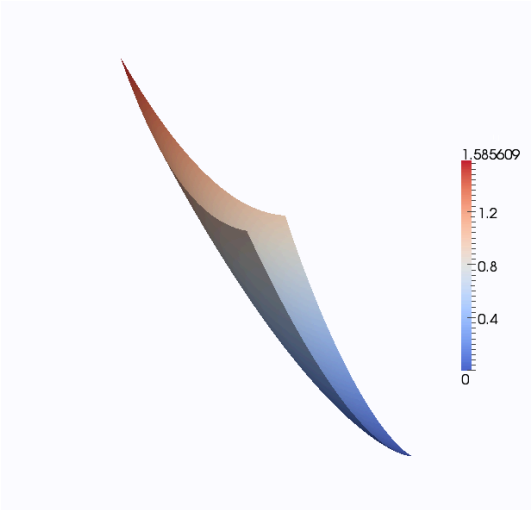


(a) The FE approximation to the function $u(\boldsymbol{x}) = e^{\frac{|\boldsymbol{x}|^2}{2}}$

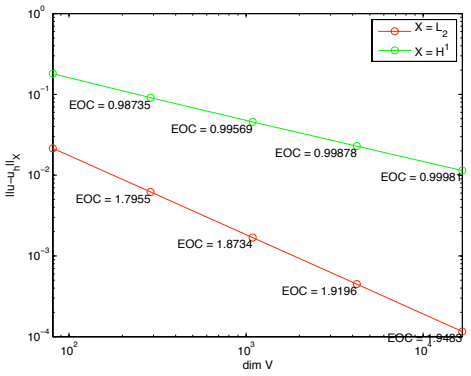


(b) Log–log error plot for $\mathbb{P}^1$ Lagrange FEs

Figure 5.6: Numerical results for Monge–Ampère on the square $\Omega = [0, 1]^2$. Choosing an $f = |\boldsymbol{x}|^{-1}$ and g. Notice that it blows up at the boundary, the solution is the function function $u(\boldsymbol{x}) = \frac{2\sqrt{2}}{3} |\boldsymbol{x}|^{3/2}$.

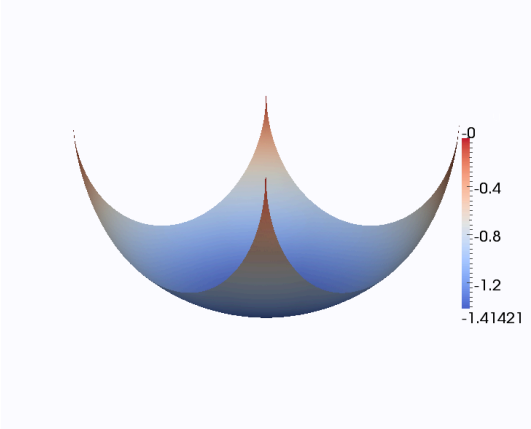


(a) The FE approximation to the function $u(\boldsymbol{x}) = \frac{2\sqrt{2}}{3} |\boldsymbol{x}|^{3/2}$

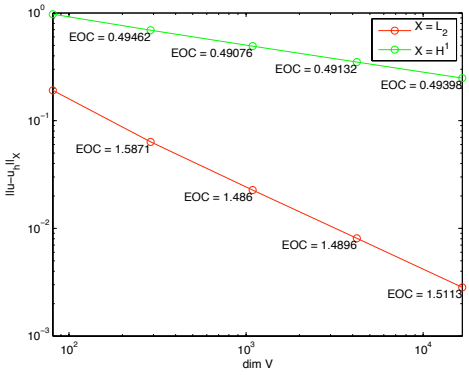


(b) Log–log error plot for $\mathbb{P}^1$ Lagrange FEs

Figure 5.7: Numerical results for Monge–Ampère on the square $\Omega = [-1, 1]^2$. Choosing f and g appropriately such that the solution is $u(\boldsymbol{x}) = -\sqrt{2 - x_1^2 - x_2^2}$.

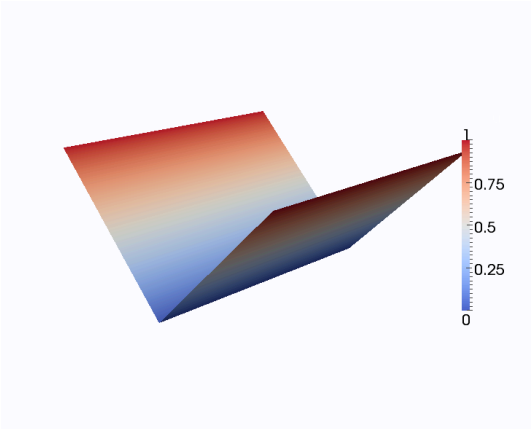


(a) The FE approximation to the function $u(\boldsymbol{x}) = -\sqrt{2 - x_1^2 - x_2^2}$.

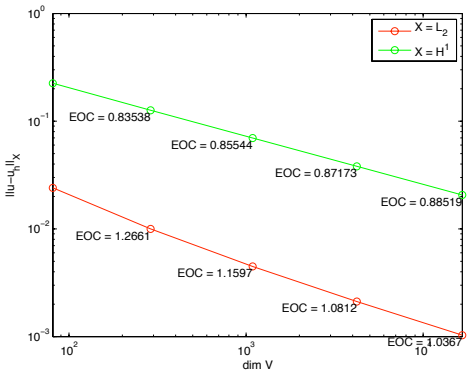


(b) Log–log error plot for $\mathbb{P}^1$ Lagrange FEs.

Figure 5.8: Numerical results for Monge–Ampère on the square $\Omega = [-1, 1]^2$. Choosing f and g appropriately such that the solution is $u(\boldsymbol{x}) = |x_1|$. This is enforced with a "discrete Dirac function".



(a) The FE approximation to the function $u(\boldsymbol{x}) = |x_1|$.



(b) Log–log error plot for $\mathbb{P}^1$ Lagrange FEs.

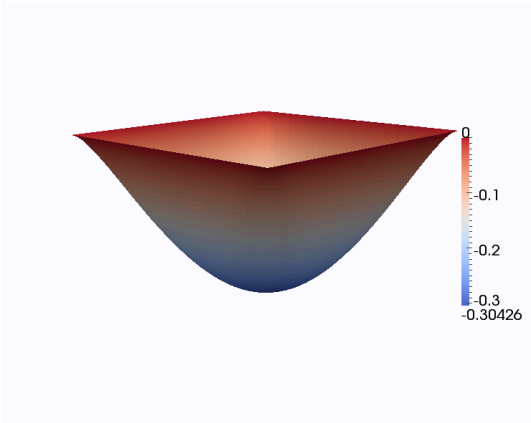
We also test the ability of the method to approximate functions which can only satisfy the Monge–Ampère equation in the viscosity sense. To that end consider the Monge–Ampère equation with problem data as follows:

$$\begin{aligned} \det \mathrm{D}^2 u &= 1 \text{ in } \Omega \\ u &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{5.94}$$

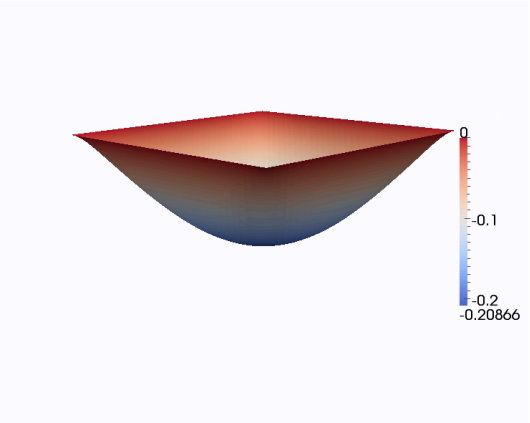
This equation can have no classical solution. It is found in most of the previous numerical studies on the equation. We include it here for comparitive purposes.

We fix $h = \sqrt{2}/64$ and show the NLFE solution at each Newton step. Here we take the initial guess to be the finite element solution to $\Delta u = 1/\sqrt{2}$. The surface plots of U^i with $i = 0, \ldots, 3$ are given in Figure 5.9. We also show contour plots of U^i with $i = 0 \ldots 3$ in Figure 5.10. We complete the study of this problem by looking at the convergence rate of the final residual in the Newton scheme $\mathfrak{R}[U^N]$ as we refine the mesh, this is given in Figure 5.11.

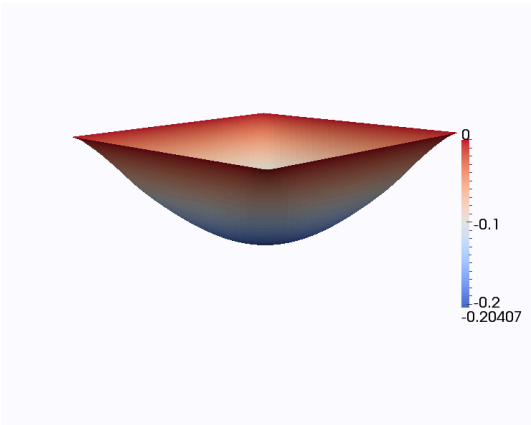
Figure 5.9: The numerical approximation of the solution of the Monge–Ampère equation (5.94) under a Newton linearisation. We show the initial guess and the first three Newton iterates. Note convexity (in the classical sense) is violated near the corners of the mesh.



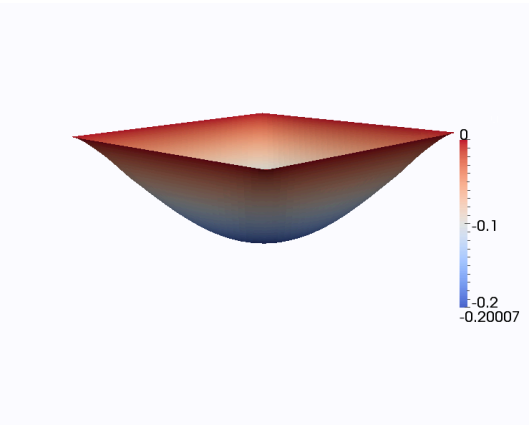
(a) The initial guess U^0



(b) The first Newton step U^1

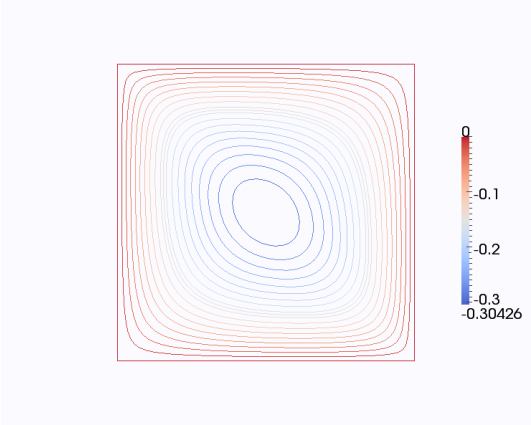


(c) The second Newton step U^2

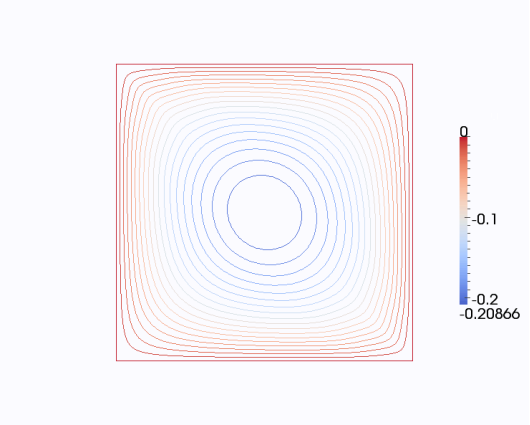


(d) The third Newton step U^3

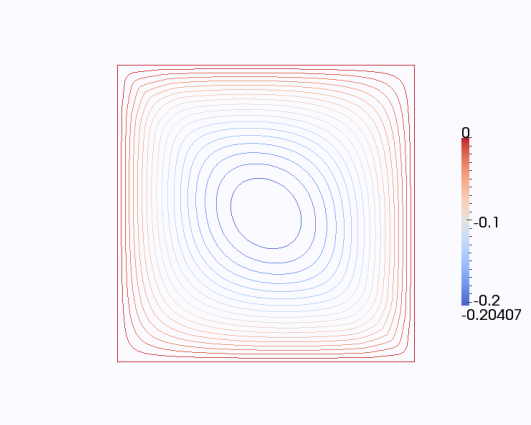
Figure 5.10: Contour plots of the numerical approximation of the solution of the Monge–Ampère equation under a Newton linearisation. We show the initial guess and the first three Newton iterates. Note convexity (in the classical sense) is violated near the corners of the mesh.



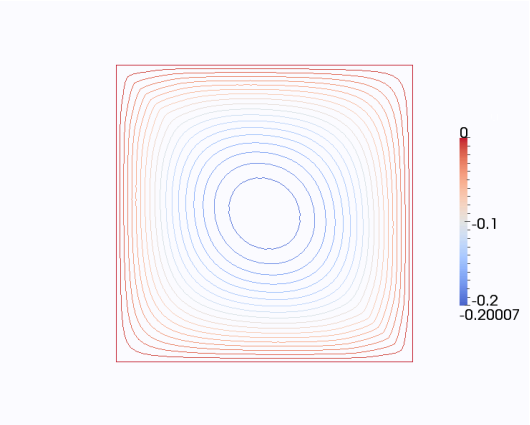
(a) The initial guess U^0



(b) The first Newton step U^1

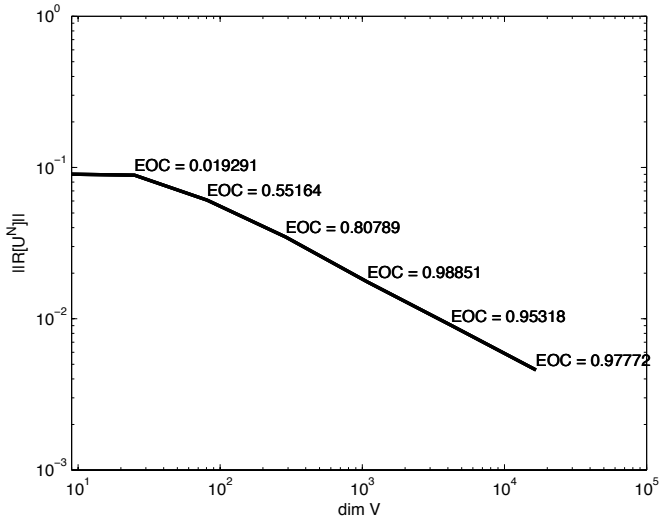


(c) The second Newton step U^2



(d) The third Newton step U^3

Figure 5.11: We numerically study the convergence rates of the residual $\mathfrak{R}[U^N]$, of the final Newton step defined in (5.87). Notice it has a linear convergence, i.e., $\left\|\mathfrak{R}[U^N]\right\| = \mathrm{O}(h)$.



5.9 Towards a rigorous error analysis of the nonlinear finite element method

In this section we give a possible starting point for analysing the method proposed for general unconstrained fully nonlinear PDEs §5.2.

5.9.0.12 Remark (on the error of the NLFEM)**.** The "full" error of the proposed method (Definition 5.2.0.3) can readily be split into its linearisation and discretisation errors. Suppose u solves the continuous nonlinear problem (5.21). Given u^0, let $(u^n)_{n\in\mathbb{N}_0}$ be the sequence of solutions to the continuous linear problems

$$\boldsymbol{N}(\mathrm{D}^2u^n):\mathrm{D}^2u^{n+1} = g(\mathrm{D}^2u^n). \tag{5.95}$$

Set $U^0 = \Lambda u^0$, then let $(U^n)_{n\in\mathbb{N}_0}$ solve the discrete linear problems

$$\left\langle \boldsymbol{N}(\boldsymbol{H}[U^n]):\boldsymbol{H}[U^{n+1}], \mathring{\Phi}\right\rangle = \left\langle g(\boldsymbol{H}[U^n]), \mathring{\Phi}\right\rangle \quad \forall\, \mathring{\Phi} \in \mathring{\mathbb{V}}. \tag{5.96}$$

Then we may write

$$\|u - U^n\|_{\mathscr{X}} \leq \|u - u^n\|_{\mathscr{X}} + \|u^n - U^n\|_{\mathscr{X}}\,. \tag{5.97}$$

We denote the term $\|u - u^n\|_{\mathscr{X}}$ to be the linearisation error and $\|u^n - U^n\|_{\mathscr{X}}$ to be the discretisation error. The linearisation error is well known to be quadratically convergent in n given a sufficiently "good" initial guess [Kel95]. We will only study the discretisation error further here.

5.9.0.13 Remark (a "nonvariational" crime)**.** Applying a direct discretisation to

$$\text{find } u^{n+1} \text{ such that } \boldsymbol{N}(\mathrm{D}^2 u^n){:}\mathrm{D}^2 u^{n+1} = g(\mathrm{D}^2 u^n), \tag{5.98}$$

gives

$$\text{find } U^{n+1} \text{ such that } \boldsymbol{N}(\mathrm{D}^2 u^n){:}\boldsymbol{H}[U^{n+1}] = g(\mathrm{D}^2 u^n). \tag{5.99}$$

However $\mathrm{D}^2 u^n$ is not available to us so we approximate it in the problem coefficients with the finite element Hessian $\boldsymbol{H}[U^n]$ and hence commit a *nonvariational crime*. The extent of the nonvariational crime is quantified in Theorem 5.9.0.14.

5.9.0.14 Theorem (convergence of the finite element Hessian)**.** *Let $v \in \mathrm{H}^{j+2}(\Omega) \cap \mathrm{H}^1_0(\Omega)$, for some $0 \leq j \leq p+1$, be a generic function and $V \in \mathring{\mathbb{V}}$. Then there exist constants C_1 and C_2 such that the following bound holds*

$$\left\|\boldsymbol{H}[V] - \mathrm{D}^2 v\right\|_{\mathrm{H}^{-1}(\Omega)} \leq (1 + C_2)\,|V - v|_1 + C_1 h^{j+1}\,|v|_{j+2}\,. \tag{5.100}$$

Proof Testing $\boldsymbol{H}[V] - \mathrm{D}^2 v$ with a generic $\phi \in \mathrm{H}^1_0(\Omega)$ Theorem **??** (the existence of $\boldsymbol{H}[V]$ over $\mathbb{V}$) gives

$$\begin{aligned}
\left\langle \boldsymbol{H}[V] - \mathrm{D}^2 v, \phi \right\rangle &= \langle \boldsymbol{H}[V], \mathrm{P}_{\mathbb{V}}\,\phi\rangle - \left\langle \mathrm{D}^2 v, \phi \right\rangle \\
&= \left\langle \mathrm{D}^2 V \mid \mathrm{P}_{\mathbb{V}}\,\phi \right\rangle + \langle \nabla v \otimes \nabla \phi \rangle \\
&= -\left\langle \nabla V \otimes \nabla\, \mathrm{P}_{\mathbb{V}}\,\phi \right\rangle + \langle \nabla v \otimes \nabla \phi \rangle\,.
\end{aligned} \tag{5.101}$$

Adding and subtracting $\langle \nabla V \otimes \nabla \phi \rangle$ and $\langle \nabla v \otimes \nabla \left(\mathrm{P}_{\mathbb{V}} \, \phi - \phi \right) \rangle$

$$\begin{aligned}
\left\langle \boldsymbol{H}[V] - \mathrm{D}^2 v, \phi \right\rangle &= \langle \nabla V \otimes \nabla \left(\mathrm{P}_{\mathbb{V}} \, \phi - \phi \right) \rangle + \langle \nabla \left(v - V \right) \otimes \nabla \phi \rangle \\
&= \langle \nabla \left(V - v \right) \otimes \nabla \left(\mathrm{P}_{\mathbb{V}} \, \phi - \phi \right) \rangle + \langle \nabla \left(v - V \right) \otimes \nabla \phi \rangle \\
&\qquad + \langle \nabla v \otimes \nabla \left(\mathrm{P}_{\mathbb{V}} \, \phi - \phi \right) \rangle \\
&= \langle \nabla \left(V - v \right) \otimes \nabla \left(\mathrm{P}_{\mathbb{V}} \, \phi - \phi \right) \rangle + \langle \nabla \left(v - V \right) \otimes \nabla \phi \rangle \\
&\qquad - \left\langle \mathrm{D}^2 v, \mathrm{P}_{\mathbb{V}} \, \phi - \phi \right\rangle \\
&= \langle \nabla \left(V - v \right) \otimes \nabla \left(\mathrm{P}_{\mathbb{V}} \, \phi - \phi \right) \rangle + \langle \nabla \left(v - V \right) \otimes \nabla \phi \rangle \\
&\qquad - \left\langle (I - \mathrm{P}_{\mathbb{V}}) \mathrm{D}^2 v, \phi \right\rangle \\
&\leq |v - V|_1 \left(|\mathrm{P}_{\mathbb{V}} \, \phi - \phi|_1 + |\phi|_1 \right) + \left\| (I - \mathrm{P}_{\mathbb{V}}) \mathrm{D}^2 v \right\|_{\mathrm{H}^{-1}(\Omega)} |\phi|_1 \\
&\leq |\phi|_1 \left((1 + C_2) \, |V - v|_1 + C_1 h^{j+1} \left| \mathrm{D}^2 v \right|_j \right),
\end{aligned} \tag{5.102}$$

where we have made use of the self adjoint property of the $\mathrm{L}_2(\Omega)$ projection, together with an $\mathrm{H}^1(\Omega)$ stability bound and the convergence result from Lemma 4.6.0.3. □

5.9.0.15 Theorem (discretisation error)**.** *Let $(u^n)_{n \in \mathbb{N}_0}$ be the sequence of solutions to the sequence of linear equations*

$$\boldsymbol{N}(\mathrm{D}^2 u^n) : \mathrm{D}^2 u^{n+1} = g(\mathrm{D}^2 u^n). \tag{5.103}$$

Let $(U^n)_{n \in \mathbb{N}_0}$ be the sequence of finite element solutions to the discretisation of 5.103 as given in Theorem 5.2.0.3. Then the following bound holds

$$\begin{aligned}
\left\| \boldsymbol{N}(\mathrm{D}^2 u^n) : \mathrm{D}^2 u^{n+1} - \boldsymbol{N}(\boldsymbol{H}[U^n]) : \boldsymbol{H}[U^{n+1}] \right\|_{-1} \leq C \Bigg(& h^{j+1} \left(\left| u^{n+1} \right|_{j+1} + |u^n|_{j+1} \right) \\
& + |u^n - U^n|_1 \Bigg).
\end{aligned} \tag{5.104}$$

Proof We begin by splitting the error into a "pure" discretisation error (as appearing in Remark 5.9.0.13) and the "nonvariational" error

$$\begin{aligned}
\left\| \boldsymbol{N}(\mathrm{D}^2 u^n) : \mathrm{D}^2 u^{n+1} - \boldsymbol{N}(\boldsymbol{H}[U^n]) : \boldsymbol{H}[U^{n+1}] \right\|_{-1} &\leq \left\| \boldsymbol{N}(\mathrm{D}^2 u^n) : \left(\mathrm{D}^2 u^{n+1} - \boldsymbol{H}[U^{n+1}] \right) \right\|_{-1} \\
&\quad + \left\| \left(\boldsymbol{N}(\mathrm{D}^2 u^n) - \boldsymbol{N}(\boldsymbol{H}[U^n]) \right) : \boldsymbol{H}[U^{n+1}] \right\|_{-1} \\
&=: \delta + \nu.
\end{aligned} \tag{5.105}$$

Where we are using δ to denote the "pure" discretisation error and ν the "nonvariational". Now δ can be bounded using the theory of the linear case (see Theorem 4.6.0.4) as follows. Assume that the continuous solution of the $(n+1)$-th linearised problem u^{n+1} belongs to $\mathrm{H}^{j+2}(\Omega)$ for some $j = 0, \ldots, p+1$ then

$$\begin{aligned} \delta &= \left\| \boldsymbol{N}(\mathrm{D}^2 u^n) : \left(\mathrm{D}^2 u^{n+1} - \boldsymbol{H}[U^{n+1}]\right) \right\|_{-1} \\ &\leq C h^{j+1} \left| g(\mathrm{D}^2 u^n) \right|_j \\ &\leq C h^{j+1} \left| u^{n+1} \right|_{j+2}. \end{aligned} \tag{5.106}$$

By Theorem 5.9.0.14 we have

$$\begin{aligned} \nu &= \left\| \left(\boldsymbol{N}(\mathrm{D}^2 u^n) - \boldsymbol{N}(\boldsymbol{H}[U^n])\right) : \boldsymbol{H}[U^{n+1}] \right\|_{-1} \\ &\leq C \left\| \mathrm{D}^2 u^n - \boldsymbol{H}[U^n] \right\|_{-1} \\ &\leq C \left(\left| u^n - U^n \right|_1 + h^{j+1} \left| u^n \right|_{j+2} \right), \end{aligned} \tag{5.107}$$

where C will depend on the coefficient matrix $\boldsymbol{N}$. Combining the bounds for δ and ν yields the desired result. □

5.9.0.16 Remark (the presence of $|u^n - U^n|_1$)**.** Currently we have no analytical bounds for the gradient error of the NDFE solution. Although numerically it may be inferred (see §4.5) that $|u^n - U^n|_1 = \mathrm{O}(h^p)$. This would then yield optimal convergence rates of the residual under the assumption $(u^n)_{n \in \mathbb{N}_0} \in \mathrm{H}^2(\Omega)$.

Chapter 6

Summary and Open Problems

6.1 Part 1

In the first half of this work we rigorously analysed the backward Euler finite element approximation of linear parabolic equations in an aposteriori sense. In this case we used the heat equation as a prototype although the analysis can easily be extended to general linear parabolic operators.

We extensively tested the resultant aposteriori estimate numerically, showing under the step size condition $\tau \ll h^p$ the resultant estimator is asymptotically exact. We also used the estimator to drive a heuristic adaptive algorithm.

As is always the case this work has generated more questions than answers, we propose new directions for the research below.

6.1.1 Lower bounds

In §3 we only derive upper bounds of the semidiscrete and fully discrete error. It is possible to derive lower bounds for the error. In [CJ04] Chen and Feng bound the spatial error indicators based on the *bubble functions* introduced by Verfürth for elliptic problems [Ver94b]. Work of note is that of Bergam, Bernardi and Mghazli [BBM05] who give lower bounds to their estimators. This is interesting since they provide space and time estimators that are fully decoupled.

6.1.2 Rates of convergence for the adaptive scheme

In §3.7 we gave a heuristic adaptive algorithm. We provide no analytical convergence proof for this algorithm. Convergence of adaptive finite element methods for elliptic problems is reasonably well understood, [BDD04, MNS02b, MNS00]. For evolution problems some work has been done on the convergence of the adaptive scheme by Chen and Feng [CF04] however no rates are given.

6.1.3 Higher order approximation of the time derivative

We analysed the backward Euler approximation of the time derivative in the heat equation. We may use a higher order approximation. Work has been done on Crank–Nicholson and general Runge–Kutta schemes approximation of general parabolic problems by Akrivis, Makridakis and Nochetto [AMN06, AMN09] for residual estimators.

6.2 Part 2

In the second half of this work we derived a finite element method for linear elliptic problems in nonvariational form. We studied the method numerically and showed the error converged "optimally", by which we mean $\|u - U\| = \mathrm{O}(h^{p+1})$ and $|u - U|_1 = \mathrm{O}(h^p)$.

6.2.1 Nonconforming finite element approximation

We believe a discontinuous Galërkin finite element method may lend itself to the framework laid out in §4 even better than the conforming method presented. The reason for this is the mass matrix appearing in the linear system can be decomposed into elementwise components. This means the linear system itself could be solved locally, resulting in extremely fast solution time. The added complication in this case would be defining the Hessian of an object that was still piecewise smooth however no longer continuous.

6.2.2 Numerical apriori analysis

Although apriori analysis is given for the method, we only obtain rates of convergence for the residual in a dual norm. A more thorough apriori analysis is required for the method. In particular to show convergence of the fully nonlinear scheme we must first

derive bounds for the linear case, as noted in Remark 5.9.0.16, we must bound the term $|u - U|_1$.

6.2.3 Stochastic processes

Stochastic analysis yields many nonvariational form equations. In particular, the forward Kolmogorov equation (also known as Fokker–Plank equation) is of this form. This area of research is lacking numerical methods and we believe ours to be a valuable addition to the field.

6.2.4 Condition number of the block matrix E

We numerically studied E and shown it to have a condition number of $\mathrm{O}(h^{-2})$. No analytic proof for this observation exists and would be useful. In fact this is a major advantage to the method we propose in §5 over that proposed in [FN07, FN08b, FN08a] since they must discretise the biharmonic equation and hence the resultant linear system has at best a condition number of $\mathrm{O}(h^{-4})$.

6.2.5 Aposteriori analysis

The aposteriori error analysis given in §4.6 is of residual type and as such it should be possible to prove lower bounds using the bubble functions proposed by Verfürth [Ver94b].

6.2.6 Termination of linearisations

Moving onto the nonlinear cases we studied, as noted in Remark 4.8.1.3 the stopping criterion we used for the linearisation of both the quasilinear and the fully nonlinear equations is not ideal. A more thorough investigation is needed into efficient termination of the iterative solver.

6.2.7 Analysis of the fully nonlinear scheme

A very simple groundwork to this end has already been set out in §5.9. A thorough analysis of this method hinges on the estimates obtained in the linear case.

6.2.8 Dealing with constraints generally

Although we have proposed a numerical method for general unconstrained fully nonlinear PDEs, in the general constrained case it is difficult to see how to pass constraints down to the discrete level. In the case of the Monge–Ampère equation we may formulate the discrete problem as a semidefinite program. However if the constraint is any "more nonlinear" than the quadratic there is no obvious way to deal with it.

6.2.9 Approximating every convex function

Since we are restricted in the theory to using test functions which are nonnegative throughout the domain (see Remark 5.6.0.9), the method proposed for the Monge–Ampère equation is with piecewise linear finite elements, $p = 1$. Note, however, with certain problems the method does not converge to the solution but to some other function (albeit one not far from the solution). This is due to the observation in [AM09] that the finite element space consisting of piecewise linear functions is not "rich enough to approximate every convex function". A solution may be to create extra degrees of freedom on the elements and use the quadratic bubble functions defined as standard for $\mathbb{P}^2$ elements. These functions again remain nonnegative through the domain and so this method would fit the theory set out in Remark 5.6.0.7.

6.2.10 Parabolic fully nonlinear equations

We have provided a novel numerical method for the elliptic Monge–Ampère equation. An obvious extension is to the parabolic Monge–Ampère equation. In fact to the author's knowledge there has been no work done on numerical methods for the parabolic Monge–Ampère equation and this would make for an interesting field of research.

Appendix A

Interesting things

A.1 Useful theorems and inequalities

A.1.0.1 Definition (dual space)**.** The dual of of a Hilbert space, H, is defined as the space consisting of all continuous linear functionals from H into $\mathbb{R}$ (or $\mathbb{C}$).

A.1.0.2 Theorem (Reisz Representation Theorem)**.** *Let $(\cdot,\cdot)$ be the inner product of a Hilbert space H. Let $\Lambda : H \to \mathbb{R}$ be a continuous, linear functional then there exists a $u \in H$ such that*

$$\Lambda(v) = (v,u) \quad \forall\, v \in H. \tag{A.1}$$

A.1.0.3 Lemma (Lax-Milgram [Cia78])**.** *Let $a(\cdot,\cdot) : \mathrm{H}_0^1(\Omega) \times \mathrm{H}_0^1(\Omega) \to \mathbb{R}$ be a bilinear form that is: continuous*

$$a(u,v) \leq \alpha \left\|u\right\|_1 \left\|v\right\|_1 \tag{A.2}$$

and coercive

$$a(u,u) \geq \beta \left\|u\right\|_1^2 . \tag{A.3}$$

Let $l(\cdot) : \mathrm{H}_0^1(\Omega) \to \mathbb{R}$ be a linear form that is continuous

$$l(v) \leq C \left\|v\right\|_1 . \tag{A.4}$$

Then the weak formulation

$$\text{find } u \in \mathrm{H}_0^1(\Omega) \text{ such that } a(u,v) = l(v) \quad \forall\, v \in \mathrm{H}_0^1(\Omega) \tag{A.5}$$

has a unique weak solution.

A.1.0.4 Lemma (Poincaré–Friedrichs Inequality)**.** *Let $\Omega \subset \mathbb{R}^d$ be open, bounded and have Lipschitz boundary. Then there exists a $C = C(\Omega)$ such that for any $v \in \mathrm{H}_0^1(\Omega)$*

$$\|v\| \leq C \|\nabla v\| \tag{A.6}$$

or for $v \in \mathrm{H}^1(\Omega)$, denoting $\bar{v} = \frac{1}{|\Omega|} \int_\Omega v$

$$\|v - \bar{v}\| \leq C \|\nabla v\| . \tag{A.7}$$

A.1.0.5 Lemma (Young's Inequality)**.** *Given $a, b \in \mathbb{R}^+$ then for any $\epsilon > 0$ the following inequality holds*

$$ab \leq \frac{a^2}{2\epsilon} + \frac{\epsilon b^2}{2}. \tag{A.8}$$

A.1.0.6 Lemma (Cauchy–Bunyakovskǐ–Schwarz)**.** *Given a normed space N equipped with an inner product $(\cdot, \cdot)_N$ and norm $\|\cdot\|_N$. For each $x, y \in N$ it holds that*

$$(x, y)_N \leq \|x\|_N \|y\|_N . \tag{A.9}$$

A.1.0.7 Lemma (Green's Identity)**.** *Let $u, v \in \mathrm{C}^1(\Omega)$ then*

$$\int_\Omega \nabla uv = -\int_\Omega u \nabla v + \int_{\partial\Omega} uv\boldsymbol{n}, \tag{A.10}$$

where $\boldsymbol{n}$ is the outward pointing normal to Ω.

A.1.0.8 Definition (Precompact)**.** A set S is precompact if its closure $\bar{S}$ is compact.

A.1.0.9 Theorem (Ascoli–Arzelá)**.** *Let $S \subset \mathcal{C}(\Omega)$ where $\Omega \subset \mathbb{R}^d$. Suppose the functions in S are uniformly bounded*

$$\sup_{f \in S} \|f\|_{L^\infty} < \infty \tag{A.11}$$

and equicontinuous over Ω

$$|f(\boldsymbol{x}) - f(\boldsymbol{y})| \leq \epsilon_S(\delta) \text{ whenever } \|\boldsymbol{x} - \boldsymbol{y}\| \leq \delta \tag{A.12}$$

where $\lim_{\delta \to 0} \epsilon_s(\delta) = 0$. Then S is precompact in $\mathcal{C}(\Omega)$.

A.1.0.10 Definition (Gram matrix)**.** The Gram matrix of a set of vectors $\mathbf{v} = (v_1, \ldots, v_n)$ in an inner product space, $\mathscr{X}$ is the matrix of inner products, whose entries are given by

$$\mathbf{X} = \langle \mathbf{v}, \mathbf{v}^{\mathsf{T}} \rangle_{\mathscr{X}} . \tag{A.13}$$

A.1.0.11 Theorem (regularity of solutions for second order elliptic PDEs in divergence form)**.** *Let u be the solution to*

$$-\operatorname{div} \boldsymbol{A} \nabla u = f. \tag{A.14}$$

Assume $\partial\Omega \in \mathrm{C}^{1,1}(\Omega)$ and the coefficients of problem (A.14) satisfy the following

$$\begin{aligned} &\boldsymbol{A} \in \mathrm{L}_\infty(\Omega)^{d\times d}, \\ &\boldsymbol{A}(\boldsymbol{x}) \in \mathrm{Sym}^+(\mathbb{R}^{d\times d}) \quad \forall\, \boldsymbol{x} \in \Omega, \\ &\exists \lambda > 0 : \lambda^{-1} |\xi|^2 \leq \xi^T \boldsymbol{A} \xi \leq \lambda |\xi|^2 \quad \forall\, \xi \in \mathbb{R}^d. \end{aligned} \tag{A.15}$$

Then if $f \in \mathrm{L}_2(\Omega)$ it follows that $u \in \mathrm{H}^2(\Omega)$ and there exists a constant $C = C(d, \Omega, \lambda, \mathbb{V})$ such that

$$|u|_2 \leq C \|f\| . \tag{A.16}$$

A.1.0.12 Theorem ([Gri85] trace theorem for polygonal domains)**.** *Let Ω be a Lipschitz domain, then the trace operator, T,*

$$\begin{aligned} T: \quad \mathrm{H}^1(\Omega) &\to \mathrm{H}^{1/2}(\partial\Omega) \\ u &\mapsto Tu := u|_{\partial\Omega} , \end{aligned} \tag{A.17}$$

is bounded, linear and injective, that is there exists a $C > 0$ such that

$$\|Tu\|_{\mathrm{H}^{1/2}(\partial\Omega)} \leq C \|u\|_1 . \tag{A.18}$$

Moreover there exists an operator T^{-1} such that given a function $g \in \mathrm{H}^{1/2}(\partial\Omega)$

$$TT^{-1}g = g. \tag{A.19}$$

Hence the operator T^{-1} is a right inverse of T.

A.2 Fractional order Sobolev spaces

In addition to integer ordered Sobolev spaces like those defined in §2 it is also possible to define fractional order spaces. Assume $k \geq 0$ [1] and the relation $k = l + m$, with $l \in \mathbb{N}_0$ such that $m \in (0, 1)$ holds. Then the space $\mathrm{H}^k(\Omega)$ can be defined using a integral version of Hölder continuity as follows

$$\mathrm{H}^k(\Omega) = \mathrm{H}^l(\Omega) \cap \left\{ \phi : \sum_{\alpha=l} \int_\Omega \int_\Omega \frac{\phi^{(\alpha)}(\boldsymbol{x}) - \phi^{(\alpha)}(\boldsymbol{y})}{|\boldsymbol{x} - \boldsymbol{y}|^{d-2m}} \right\} \leq \infty. \tag{A.20}$$

This is also known as a Sobolev–Slobodeckij space. Its corresponding norm is

$$\|v\|_k^2 := \|v\|_l^2 + \sum_{\alpha=l} \int_\Omega \int_\Omega \frac{\phi^{(\alpha)}(\boldsymbol{x}) - \phi^{(\alpha)}(\boldsymbol{y})}{|\boldsymbol{x} - \boldsymbol{y}|^{d-2m}}. \tag{A.21}$$

A.3 Classically convex functions

A.3.0.13 Definition (convexity, strict convexity and uniform convexity)**.** A function $\phi : \Omega \to \mathbb{R}$ is *convex* on Ω if for all $\boldsymbol{x}, \boldsymbol{y} \in \Omega$ and $\alpha \in \mathbb{R}^+$

$$\phi(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \leq \alpha\phi(\boldsymbol{x}) + (1-\alpha)\phi(\boldsymbol{y}). \tag{A.22}$$

It is *strictly convex* if

$$\phi(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}) < \alpha\phi(\boldsymbol{x}) + (1-\alpha)\phi(\boldsymbol{y}). \tag{A.23}$$

It is *uniformly convex* if there exists a $C > 0$ such that

$$\phi(\alpha\boldsymbol{x} + (1-\alpha)\boldsymbol{y}) \leq \alpha\phi(\boldsymbol{x}) + (1-\alpha)\phi(\boldsymbol{y}) - C\alpha(1-\alpha)\left|\boldsymbol{x} - \boldsymbol{y}\right|^2. \tag{A.24}$$

If f is a convex, stricly convex or uniformly convex function then $-f$ is a concave, strictly concave or uniformly concave function respectively.

A.3.0.14 Proposition (linear combination of convex functions)**.** *Let $\{\phi_i\}_i$ be a finite set of convex functions on Ω. Let $\{\alpha_i\}_i$ denote a finite set of nonnegative real numbers. Then*

$$\sum_i \alpha_i\phi_i \text{ is a convex function on } \Omega. \tag{A.25}$$

[1] but may no longer be in $\mathbb{N}_0$.

A.3.0.15 Proposition (gradient condition of convexity)**.** *Let $\phi : \Omega \to \mathbb{R}$ be differentiable then*

- *ϕ is convex if and only if*

$$\phi(\boldsymbol{y}) - \phi(\boldsymbol{x}) \geq \mathrm{D}f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) \tag{A.26}$$

- *ϕ is strictly convex if and only if*

$$\phi(\boldsymbol{y}) - \phi(\boldsymbol{x}) > \mathrm{D}f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) \tag{A.27}$$

- *ϕ is uniformly convex if and only if*

$$\phi(\boldsymbol{y}) - \phi(\boldsymbol{x}) \geq \mathrm{D}f(\boldsymbol{x})(\boldsymbol{y} - \boldsymbol{x}) + C\left|\boldsymbol{y} - \boldsymbol{x}\right|^2. \tag{A.28}$$

A.3.0.16 Proposition (Hessian condition of convexity)**.** *Let $\phi : \Omega \to \mathbb{R}$ be twice differentiable then*

- *ϕ is convex if and only if $\mathrm{D}^2\phi(\boldsymbol{x}) \geq 0$ for each $\boldsymbol{x} \in \Omega$.*
- *ϕ is strictly convex if and only if $\mathrm{D}^2\phi(\boldsymbol{x}) > 0$ for each $\boldsymbol{x} \in \Omega$.*
- *ϕ is uniformly convex if and only if there exists $C > 0$ such that $\mathrm{D}^2\phi(\boldsymbol{x}) \geq C\left|\boldsymbol{x}\right|^2$ for each $\boldsymbol{x} \in \Omega$.*

A.4 Semidefinite programming

A semidefinite programming problem is an optimisation problem of the form : Given $\boldsymbol{a}, \boldsymbol{x} \in \mathbb{R}^N, \{\boldsymbol{A}_i\}_{i=1}^M \in \mathrm{Sym}(\mathbb{R}^{M\times M})$

$$\begin{aligned} &\text{minimise } \boldsymbol{a}^\intercal \boldsymbol{x} \\ &\text{subject to } \sum_{i=1}^{N} x_i \boldsymbol{A}_i \geq 0. \end{aligned} \tag{A.29}$$

This is a generalisation of linear programming to include positive semidefinite constraints. These can be solved efficiently using interior point methods (barrier methods) for example. We will not discuss this further, instead we direct interested readers to [BGLS06].

A.5 Viscosity solutions

This section will provide a brief introduction to viscocity solutions of uniformly elliptic fully nonlinear PDEs of second order of the form

$$\mathscr{N}[u] := F(\mathrm{D}^2 u) - f = 0. \tag{A.30}$$

Since the weak solution framework cannot be applied in the context of fully nonlinear equations a different notion of weak solutions must be considered. The paper by Crandall and Lions [CL83] gives a notion of viscosity solutions for first order Hamilton–Jacobi type equations. This was later extended to second order PDEs in [CIL92]. The viscosity solution gives a very useful theory in proving existence of solutions to such equations.

Before introducing explicitly what a viscocity solution is we first give the following variant of the maximum principle

A.5.0.17 Theorem (maximum principle)**.** *A given function $u \in \mathrm{C}^2(\overline{\Omega})$ is said to be a classical solution of (5.2) if and only if both the following two conditions hold:*

- *For each $\phi \in \mathrm{C}^2(\overline{\Omega})$, if x_0 is a local maximum of $u - \phi$ then*

$$F(\mathrm{D}^2\phi(x_0)) \leq f(x_0) \tag{A.31}$$

- *For each $\phi \in \mathrm{C}^2(\overline{\Omega})$, if x_0 is a local minimum of $u - \phi$ then*

$$F(\mathrm{D}^2\phi(x_0)) \geq f(x_0) \tag{A.32}$$

Note we are assuming no regularity on u, in fact all the smoothness requirements are on ϕ.

The heuristic idea of a viscocity solution is to use the two properties given in Theorem A.5.0.17 and only *after* do we study the properties (existence, uniqueness, etc.).

A.5.0.18 Definition (viscoscity solution)**.** A *viscoscity solution* is a notion of weak solution for nonlinear elliptic (and parabolic) equations. In particular it is used for equations in non-divergence form. A continuous function $u \in \mathrm{C}^0(\Omega)$ is a viscosity supersolution (resp. viscosity subsolution) when the following holds. Suppose $x_0 \in \Omega$, $\phi \in \mathrm{C}^2(\Omega)$ and $u - \phi$ has a local min at x_0 then

$$F(D^2\phi) \leq f(x_0) \tag{A.33}$$

(resp. $u - \phi$ has a local max at x_0 then

$$F(D^2\phi) \geq f(x_0) \quad). \tag{A.34}$$

If u is both a supersolution and a subsolution it is then called a viscocity solution.

Bibliography

[AK01] Mark Ainsworth and Donald W. Kelly, *A posteriori error estimators and adaptivity for finite element approximation of the non-homogeneous Dirichlet problem*, Adv. Comput. Math. **15** (2001), no. 1-4, 3–23 (2002), A posteriori error estimation and adaptive computational methods. MR MR1887727 (2003a:65096)

[AM08] Nestor Aguilera and Pedro Morin, *Approximating optimization problems over convex functions*, Arxiv (2008).

[AM09] Néstor E. Aguilera and Pedro Morin, *On convex functions and the finite element method*, SIAM J. Numer. Anal. **47** (2009), no. 4, 3139–3157. MR MR2551161

[AMN06] Georgios Akrivis, Charalambos Makridakis, and Ricardo H. Nochetto, *A posteriori error estimates for the Crank-Nicolson method for parabolic equations for parabolic equations*, Math. Comp. **75** (2006), no. 254, 511–531 (electronic). MR MR2196979 (2007a:65114)

[AMN09] ———, *Optimal order a posteriori error estimates for a class of Runge-Kutta and Galerkin methods*, Numer. Math. **114** (2009), no. 1, 133–160. MR MR2557872

[AO00] Mark Ainsworth and J. Tinsley Oden, *A posteriori error estimation in finite element analysis*, Pure and Applied Mathematics (New York), Wiley-Interscience [John Wiley & Sons], New York, 2000. MR MR1885308 (2003b:65001)

[Arg01] Ioannis K. Argyros, *Local and semilocal convergence theorems for Newton's method based on continuously Fréchet differentiable operators*, Southwest J. Pure Appl. Math. (2001), 22–28 (electronic). MR MR1841317 (2002c:65072)

[AV02] A. Agouzal and Yu. Vassilevski, *On a discrete Hessian recovery for P_1 finite elements*, J. Numer. Math. **10** (2002), no. 1, 1–12. MR MR1905846 (2003c:65137)

[BBM05] A. Bergam, C. Bernardi, and Z. Mghazli, *A posteriori analysis of the finite element discretization of some parabolic equations*, Math. Comp. **74** (2005), no. 251, 1117–1138 (electronic). MR MR2136996 (2007c:65072)

[BC02] Sören Bartels and Carsten Carstensen, *Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids. II. Higher order FEM*, Math. Comp. **71** (2002), no. 239, 971–994 (electronic). MR MR1898742 (2003e:65207)

[BDD04] Peter Binev, Wolfgang Dahmen, and Ron DeVore, *Adaptive finite element methods with convergence rates*, Numer. Math. **97** (2004), no. 2, 219–268. MR MR2050077 (2005d:65222)

[BGLS06] J. Frédéric Bonnans, J. Charles Gilbert, Claude Lemaréchal, and Claudia A. Sagastizábal, *Numerical optimization*, second ed., Universitext, Springer-Verlag, Berlin, 2006, Theoretical and practical aspects. MR MR2265882 (2007e:90001)

[Boh08] Klaus Bohmer, *On finite element methods for fully nonlinear elliptic equations of second order*, preprint (2008).

[BPS02] James H. Bramble, Joseph E. Pasciak, and Olaf Steinbach, *On the stability of the L^2 projection in $H^1(\Omega)$*, Math. Comp. **71** (2002), no. 237, 147–156 (electronic). MR MR1862992 (2002h:65175)

[BR78] Ivo Babuška and Werner C. Rheinboldt, *Error estimates for adaptive finite element computations*, SIAM J. Numer. Anal. **15** (1978), no. 4, 736–754. MR 58 #3400

[Bra01] Dietrich Braess, *Finite elements*, second ed., Cambridge University Press, Cambridge, 2001, Theory, fast solvers, and applications in solid mechanics, Translated from the 1992 German edition by Larry L. Schumaker. MR 2001k:65002

[BRR81] F. Brezzi, J. Rappaz, and P.-A. Raviart, *Finite-dimensional approximation of nonlinear problems. II. Limit points*, Numer. Math. **37** (1981), no. 1, 1–28. MR MR615889 (83f:65089b)

[BRR81] ———, *Finite-dimensional approximation of nonlinear problems. I. Branches of nonsingular solutions*, Numer. Math. **36** (1980/81), no. 1, 1–25. MR MR595803 (83f:65089a)

[BRR82] ———, *Finite-dimensional approximation of nonlinear problems. III. Simple bifurcation points*, Numer. Math. **38** (1981/82), no. 1, 1–30. MR MR634749 (83f:65090)

[BS91] G. Barles and P. E. Souganidis, *Convergence of approximation schemes for fully nonlinear second order equations*, Asymptotic Anal. **4** (1991), no. 3, 271–283. MR MR1115933 (92d:35137)

[BS94] Susanne C. Brenner and L. Ridgway Scott, *The mathematical theory of finite element methods*, Springer-Verlag, New York, 1994. MR 95f:65001

[BX91] James H. Bramble and Jinchao Xu, *Some estimates for a weighted L^2 projection*, Math. Comp. **56** (1991), no. 194, 463–476. MR MR1066830 (91k:65140)

[BX03a] Randolph E. Bank and Jinchao Xu, *Asymptotically exact a posteriori error estimators. I. Grids with superconvergence*, SIAM J. Numer. Anal. **41** (2003), no. 6, 2294–2312 (electronic). MR MR2034616 (2004k:65194)

[BX03b] ———, *Asymptotically exact a posteriori error estimators. II. General unstructured grids*, SIAM J. Numer. Anal. **41** (2003), no. 6, 2313–2332 (electronic). MR MR2034617 (2004m:65212)

[Caf90] Luis A. Caffarelli, *Interior $W^{2,p}$ estimates for solutions of the Monge-Ampère equation*, Ann. of Math. (2) **131** (1990), no. 1, 135–150. MR MR1038360 (91f:35059)

[Car02] Carsten Carstensen, *Merging the Bramble-Pasciak-Steinbach and the Crouzeix-Thomée criterion for H^1-stability of the L^2-projection onto finite element spaces*, Math. Comp. **71** (2002), no. 237, 157–163 (electronic). MR MR1862993 (2002i:65125)

[Car04a] ———, *All first-order averaging techniques for a posteriori finite element error control on unstructured grids are efficient and reliable*, Math. Comp. **73** (2004), no. 247, 1153–1165 (electronic). MR MR2047082 (2005e:65173)

[Car04b] ———, *Some remarks on the history and future of averaging techniques in a posteriori finite element error analysis*, ZAMM Z. Angew. Math. Mech. **84** (2004), no. 1, 3–21. MR MR2031241 (2005d:65204)

[Car04c] ———, *Some remarks on the history and future of averaging techniques in a posteriori finite element error analysis*, ZAMM Z. Angew. Math. Mech. **84** (2004), no. 1, 3–21. MR MR2031241 (2005d:65204)

[CB02] Carsten Carstensen and Sören Bartels, *Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids. I. Low order conforming, nonconforming, and mixed FEM*, Math. Comp. **71** (2002), no. 239, 945–969 (electronic). MR MR1898741 (2003e:65212)

[CC95] Luis A. Caffarelli and Xavier Cabré, *Fully nonlinear elliptic equations*, American Mathematical Society Colloquium Publications, vol. 43, American Mathematical Society, Providence, RI, 1995. MR MR1351007 (96h:35046)

[CF01a] Carsten Carstensen and Stefan A. Funken, *Averaging technique for a posteriori error control in elasticity. III. Locking-free nonconforming FEM*, Comput. Methods Appl. Mech. Engrg. **191** (2001), no. 8-10, 861–877. MR MR1870519 (2002j:65106)

[CF01b] ———, *Averaging technique for FE—a posteriori error control in elasticity. I. Conforming FEM*, Comput. Methods Appl. Mech. Engrg. **190** (2001), no. 18-19, 2483–2498. MR MR1815651 (2002a:74114)

[CF01c] ———, *Averaging technique for FE—a posteriori error control in elasticity. II. λ-independent estimates*, Comput. Methods Appl. Mech. Engrg. **190** (2001), no. 35-36, 4663–4675. MR MR1840795 (2002d:65140)

[CF04] Zhiming Chen and Jia Feng, *An adaptive finite element algorithm with reliable and efficient error control for linear parabolic problems*, Math. Comp. **73** (2004), no. 247, 1167–1193 (electronic). MR MR2047083 (2005e:65131)

[Che02] Zhiming Chen, *A posteriori error analysis and adaptive methods for parabolic problems*, Recent progress in computational and applied PDEs (Zhangjiajie, 2001), Kluwer/Plenum, New York, 2002, pp. 145–156. MR MR2039563 (2004k:65156)

[Cia78] Philippe G. Ciarlet, *The finite element method for elliptic problems*, North-Holland Publishing Co., Amsterdam, 1978, Studies in Mathematics and its Applications, Vol. 4. MR 58 #25001

[CIL92] Michael G. Crandall, Hitoshi Ishii, and Pierre-Louis Lions, *User's guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. (N.S.) **27** (1992), no. 1, 1–67. MR MR1118699 (92j:35050)

[CJ04] Zhiming Chen and Feng Jia, *An adaptive finite element algorithm with reliable and efficient error control for linear parabolic problems*, Math. Comp. **73** (2004), no. 247, 1167–1193 (electronic). MR MR2047083 (2005e:65131)

[CKNS08] J. Manuel Cascon, Christian Kreuzer, Ricardo H. Nochetto, and Kunibert G. Siebert, *Quasi-optimal convergence rate for an adaptive finite element method*, SIAM J. Numer. Anal. **46** (2008), no. 5, 2524–2550. MR MR2421046

[CL83] Michael G. Crandall and Pierre-Louis Lions, *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc. **277** (1983), no. 1, 1–42. MR MR690039 (85g:35029)

[CL06] Jinhai Chen and Weiguo Li, *Convergence behaviour of inexact Newton methods under weak Lipschitz condition*, J. Comput. Appl. Math. **191** (2006), no. 1, 143–164. MR MR2217790 (2006m:65096)

[Clé75] Philippe Clément, *Approximation by finite element functions using local regularization*, RAIRO, Rouge, Anal. Numér. **9** (1975), no. R-2, 77–84. MR 53 #4569

[CLY06] Carsten Carstensen, W. Liu, and N. Yan, *A posteriori FE error control for p-Laplacian by gradient recovery in quasi-norm*, Math. Comp. **75** (2006), no. 256, 1599–1616 (electronic). MR MR2240626 (2007g:65103)

[CS08] Luis A. Caffarelli and Panagiotis E. Souganidis, *A rate of convergence for monotone finite difference approximations to fully nonlinear, uniformly elliptic PDEs*, Comm. Pure Appl. Math. **61** (2008), no. 1, 1–17. MR MR2361302

[CSX07] Long Chen, Pengtao Sun, and Jinchao Xu, *Optimal anisotropic meshes for minimizing interpolation errors in L^p-norm*, Math. Comp. **76** (2007), no. 257, 179–204 (electronic). MR MR2261017 (2008e:65385)

[CT87] M. Crouzeix and V. Thomée, *The stability in L_p and W_p^1 of the L_2-projection onto finite element function spaces*, Math. Comp. **48** (1987), no. 178, 521–532. MR MR878688 (88f:41016)

[Deu04] Peter Deuflhard, *Newton methods for nonlinear problems*, Springer Series in Computational Mathematics, vol. 35, Springer-Verlag, Berlin, 2004, Affine invariance and adaptive algorithms. MR MR2063044 (2005h:65002)

[DG03] Edward J. Dean and Roland Glowinski, *Numerical solution of the two-dimensional elliptic Monge-Ampère equation with Dirichlet boundary conditions: an augmented Lagrangian approach*, C. R. Math. Acad. Sci. Paris **336** (2003), no. 9, 779–784. MR MR1989280

[DG04] ———, *Numerical solution of the two-dimensional elliptic Monge-Ampère equation with Dirichlet boundary conditions: a least-squares approach*, C. R. Math. Acad. Sci. Paris **339** (2004), no. 12, 887–892. MR MR2111728

[DG05] ———, *On the numerical solution of a two-dimensional Pucci's equation with Dirichlet boundary conditions: a least-squares approach*, C. R. Math. Acad. Sci. Paris **341** (2005), no. 6, 375–380. MR MR2169156

[DG06] E. J. Dean and R. Glowinski, *Numerical methods for fully nonlinear elliptic equations of the Monge-Ampère type*, Comput. Methods Appl. Mech. Engrg. **195** (2006), no. 13-16, 1344–1386. MR MR2203972 (2006i:65191)

[DLM09] Alan Demlow, Omar Lakkis, and Charalambos Makridakis, *A posteriori error estimates in the maximum norm for parabolic problems*, SIAM J. Numer. Anal. **arXiv 0711-3928** (to appear 2009).

[Dud77] R. M. Dudley, *On second derivatives of convex functions*, Math. Scand. **41** (1977), no. 1, 159–174. MR MR0482164 (58 #2250)

[Dup82] Todd Dupont, *Mesh modification for evolution equations*, Math. Comp. **39** (1982), no. 159, 85–107. MR 84g:65131

[DW00] W. Dörfler and O. Wilderotter, *An adaptive finite element method for a linear elliptic equation with variable coefficients*, ZAMM Z. Angew. Math. Mech. **80** (2000), no. 7, 481–491. MR MR1774078 (2001e:65179)

[EJ91] Kenneth Eriksson and Claes Johnson, *Adaptive finite element methods for parabolic problems. I. A linear model problem*, SIAM J. Numer. Anal. **28** (1991), no. 1, 43–77. MR 91m:65274

[EJ95a] ———, *Adaptive finite element methods for parabolic problems. II. Optimal error estimates in $L_\infty L_2$ and $L_\infty L_\infty$*, SIAM J. Numer. Anal. **32** (1995), no. 3, 706–740. MR 96c:65162

[EJ95b] ———, *Adaptive finite element methods for parabolic problems. IV. Nonlinear problems*, SIAM J. Numer. Anal. **32** (1995), no. 6, 1729–1749. MR 96i:65081

[EJ95c] ———, *Adaptive finite element methods for parabolic problems. V. Long-time integration*, SIAM J. Numer. Anal. **32** (1995), no. 6, 1750–1763. MR MR1360458 (96i:65082)

[EJL98] Kenneth Eriksson, Claes Johnson, and Stig Larsson, *Adaptive finite element methods for parabolic problems. VI. Analytic semigroups*, SIAM J. Numer. Anal. **35** (1998), no. 4, 1315–1325 (electronic). MR MR1620144 (99d:65281)

[Eva98] Lawrence C. Evans, *Partial differential equations*, Graduate Studies in Mathematics, vol. 19, American Mathematical Society, Providence, RI, 1998. MR MR1625845 (99e:35001)

[Fae00] Birgit Faermann, *Localization of the Aronszajn-Slobodeckij norm and application to adaptive boundary element methods. I. The two-dimensional case*, IMA J. Numer. Anal. **20** (2000), no. 2, 203–234. MR 1752263 (2001e:65192)

[FN07] Xiaobing Feng and Michael Neilan, *Analysis of galerkin methods for the fully nonlinear monge ampere equation*, Arxiv (2007).

[FN08a] ———, *Mixed finite element methods for the fully nonlinear monge-ampere equation based on vanishing moment method*, Arxiv (2008).

[FN08b] ———, *Vanishing moment method and moment solution for second order fully nonlinear partial differential equations*, Arxiv (2008).

[FV03] Francesca Fierro and Andreas Veeser, *On the a posteriori error analysis for equations of prescribed mean curvature*, Math. Comp. (2003), Posted March 26.

[FV06] Francesca Fierro and Andreas Veeser, *A posteriori error estimators, gradient recovery by averaging, and superconvergence*, Numer. Math. **103** (2006), no. 2, 267–298. MR MR2222811 (2007a:65178)

[GL09] Emmanuil Georgoulis and Omar Lakkis, *A posteriori error control for discontinuous Galerkin methods for parabolic problems*, SIAM J. Numer. Anal. **arXiv 0804.4262** (to appear 2009).

[Gri85] P. Grisvard, *Elliptic problems in nonsmooth domains*, Monographs and Studies in Mathematics, vol. 24, Pitman (Advanced Publishing Program), Boston, MA, 1985. MR MR775683 (86m:35044)

[GT83] David Gilbarg and Neil S. Trudinger, *Elliptic partial differential equations of second order*, second ed., Springer-Verlag, Berlin, 1983. MR 86c:35035

[HSWW01] W. Hoffmann, A. H. Schatz, L. B. Wahlbin, and G. Wittum, *Asymptotically exact a posteriori estimators for the pointwise gradient error on each element in irregular meshes. I. A smooth problem and globally quasi-uniform meshes*, Math. Comp. **70** (2001), no. 235, 897–909 (electronic). MR MR1826572 (2002a:65178)

[HTW02] Bjørn-Ove Heimsund, Xue-Cheng Tai, and Junping Wang, *Superconvergence for the gradient of finite element approximations by L^2 projections*, SIAM J. Numer. Anal. **40** (2002), no. 4, 1263–1280 (electronic). MR MR1951894 (2004a:65153)

[IL90] H. Ishii and P.-L. Lions, *Viscosity solutions of fully nonlinear second-order elliptic partial differential equations*, J. Differential Equations **83** (1990), no. 1, 26–78. MR MR1031377 (90m:35015)

[Jen88] Robert Jensen, *The maximum principle for viscosity solutions of fully nonlinear second order partial differential equations*, Arch. Rational Mech. Anal. **101** (1988), no. 1, 1–27. MR MR920674 (89a:35038)

[Joh87] Claes Johnson, *Numerical solution of partial differential equations by the finite element method*, Cambridge University Press, Cambridge, 1987. MR MR925005 (89b:65003a)

[Joh00] Volker John, *A numerical study of a posteriori error estimators for convection-diffusion equations*, Comput. Methods Appl. Mech. Engrg. **190** (2000), no. 5-7, 757–781. MR MR1800575 (2001i:65118)

[Kel95] C. T. Kelley, *Iterative methods for linear and nonlinear equations*, Frontiers in Applied Mathematics, vol. 16, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1995, With separately available software. MR MR1344684 (96d:65002)

[KN87] Michal Křížek and Pekka Neittaanmäki, *On superconvergence techniques*, Acta Appl. Math. **9** (1987), no. 3, 175–198. MR MR900263 (88h:65208)

[Kry95] N. V. Krylov, *On the general notion of fully nonlinear second-order elliptic equations*, Trans. Amer. Math. Soc. **347** (1995), no. 3, 857–895. MR MR1284912 (95f:35075)

[KT92] Hung Ju Kuo and Neil S. Trudinger, *Discrete methods for fully nonlinear elliptic equations*, SIAM J. Numer. Anal. **29** (1992), no. 1, 123–135. MR MR1149088 (93e:65129)

[Ley04] Dmitriy Leykekhman, *Pointwise localized error estimates for parabolic finite element equations*, Numer. Math. **96** (2004), no. 3, 583–600. MR MR2028727 (2004k:65175)

[LM06] Omar Lakkis and Charalambos Makridakis, *Elliptic reconstruction and a posteriori error estimates for fully discrete linear parabolic problems*, Math. Comp. **75** (2006), no. 256, 1627–1658 (electronic). MR MR2240628 (2007e:65122)

[LMP10] Omar Lakkis, Charalambos Makridakis, and Tristan Pryer, *A comparison of duality and energy aposteriori error analysis using elliptic reconstruction*, Submitted (2010).

[LP10a] Omar Lakkis and Tristan Pryer, *Analysis of a finite element method for second order nonvariational elliptic problems*, in preparation (2010).

[LP10b] ———, *A finite element method for fully nonlinear elliptic problems*, in preparation (2010).

[LP10c] ———, *A finite element method for second order nonvariational elliptic problems*, submitted - tech report available on ArXiV http://arxiv.org/abs/1003.0292 (2010).

[LP10d] ———, *Gradient recovery in adaptive methods for parabolic problems*, accepted - tech report available on ArXiV http://arxiv.org/abs/0905.2764 (2010).

[LR05] Grégoire Loeper and Francesca Rapetti, *Numerical solution of the Monge-Ampère equation by a Newton's algorithm*, C. R. Math. Acad. Sci. Paris **340** (2005), no. 4, 319–324. MR MR2121899

[LW06] Dmitriy Leykekhman and Lars Wahlbin, *A posteriori error estimates by recovered gradients in parabolic finite element equations*, Tech. report, University of Texas, Austin, 2006, Preprint (submitted to Math. Comp.).

[LZ99] Bo Li and Zhimin Zhang, *Analysis of a class of superconvergence patch recovery techniques for linear and bilinear finite elements*, Numer. Methods Partial Differential Equations **15** (1999), no. 2, 151–167. MR MR1674357 (99m:65201)

[MN03] Charalambos Makridakis and Ricardo H. Nochetto, *Elliptic reconstruction and a posteriori error estimates for parabolic problems*, SIAM J. Numer. Anal. **41** (2003), no. 4, 1585–1594 (electronic). MR MR2034895 (2004k:65157)

[MNS00] Pedro Morin, Ricardo H. Nochetto, and Kunibert G. Siebert, *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal. **38** (2000), no. 2, 466–488 (electronic). MR 2001g:65157

[MNS02a] ———, *Convergence of adaptive finite element methods*, SIAM Rev. **44** (2002), no. 4, 631–658 (electronic) (2003), Revised reprint of "Data oscillation and convergence of adaptive FEM" [SIAM J. Numer. Anal. **38** (2000), no. 2, 466–488 (electronic); MR1770058 (2001g:65157)]. MR MR1980447

[MNS02b] ———, *Convergence of adaptive finite element methods*, SIAM Rev. **44** (2002), no. 4, 631–658 (electronic) (2003), Revised reprint of "Data oscillation and convergence of adaptive FEM" [SIAM J. Numer. Anal. **38** (2000), no. 2, 466–488 (electronic); MR1770058 (2001g:65157)]. MR MR1980447

[Mor99] Benedetta Morini, *Convergence behaviour of inexact Newton methods*, Math. Comp. **68** (1999), no. 228, 1605–1613. MR MR1653970 (99m:65114)

[MSV08] Pedro Morin, Kunibert G. Siebert, and Andreas Veeser, *A basic convergence result for conforming adaptive finite elements*, Math. Models Methods Appl. Sci. **18** (2008), no. 5, 707–737. MR MR2413035

[Nad08] Vladut Nadirashvili, *Singular solutions of hessian fully nonlinear elliptic equations*, Arxiv (2008).

[Obe05] Adam M. Oberman, *A convergent difference scheme for the infinity Laplacian: construction of absolutely minimizing Lipschitz extensions*, Math. Comp. **74** (2005), no. 251, 1217–1230 (electronic). MR MR2137000 (2006h:65165)

[Obe06] ———, *Convergent difference schemes for degenerate elliptic and parabolic equations: Hamilton-Jacobi equations and free boundary problems*, SIAM J. Numer. Anal. **44** (2006), no. 2, 879–895 (electronic). MR MR2218974 (2007a:65173)

[Obe08] ———, *Wide stencil finite difference schemes for the elliptic Monge-Ampère equation and functions of the eigenvalues of the Hessian*, Discrete Contin. Dyn. Syst. Ser. B **10** (2008), no. 1, 221–238. MR MR2399429

[OP88] V. I. Oliker and L. D. Prussner, *On the numerical solution of the equation* $(\partial^2 z/\partial x^2)(\partial^2 z/\partial y^2) - ((\partial^2 z/\partial x \partial y))^2 = f$ *and its discretizations. I*, Numer. Math. **54** (1988), no. 3, 271–293. MR MR971703 (90h:65164)

[OR66] James M. Ortega and Werner C. Rheinboldt, *On discretization and differentiation of operators with application to Newton's method*, SIAM J. Numer. Anal. **3** (1966), no. 1, 143–156. MR MR0205450 (34 #5278)

[Ova07] Jeffrey Ovall, *Function, gradient and hessian recovery using quadratic edge-bump functions.*, J. Sci. Comput. **45** (2007), no. 3, 1064–1080.

[Par95] Eun-Jae Park, *Mixed finite element methods for nonlinear second-order elliptic problems*, SIAM J. Numer. Anal. **32** (1995), no. 3, 865–885. MR MR1335659 (96d:65187)

[Pic98] Marco Picasso, *Adaptive finite elements for a linear parabolic problem*, Comput. Methods Appl. Mech. Engrg. **167** (1998), no. 3-4, 223–237. MR 2000b:65188

[Pic03] M. Picasso, *An anisotropic error indicator based on Zienkiewicz-Zhu error estimator: application to elliptic and parabolic problems*, SIAM J. Sci. Comput. **24** (2003), no. 4, 1328–1355 (electronic). MR MR1976219 (2004e:65124)

[Sch06] Alfred H. Schatz, *Some new local error estimates in negative norms with an application to local a posteriori error estimation*, Int. J. Numer. Anal. Model. **3** (2006), no. 3, 371–376. MR MR2237890

[SS05] Alfred Schmidt and Kunibert G. Siebert, *Design of adaptive finite element software*, Lecture Notes in Computational Science and Engineering, vol. 42, Springer-Verlag, Berlin, 2005, The finite element toolbox ALBERTA, With 1 CD-ROM (Unix/Linux). MR MR2127659 (2005i:65003)

[SST08] Christoph Schwab, Endre Süli, and Radu Alexandru Todor, *Sparse finite element approximation of high-dimensional transport-dominated diffusion problems*, M2AN Math. Model. Numer. Anal. **42** (2008), no. 5, 777–819. MR MR2454623 (2009g:65130)

[Ste73] Hans J. Stetter, *Analysis of discretization methods for ordinary differential equations*, Springer-Verlag, New York, 1973, Springer Tracts in Natural Philosophy, Vol. 23. MR MR0426438 (54 #14381)

[Stu99] J.F. Sturm, *Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones*, Optimization Methods and Software **11–12** (1999), 625–653, Version 1.05 available from `http://fewcal.kub.nl/sturm`.

[SW04] Alfred H. Schatz and Lars B. Wahlbin, *Asymptotically exact a posteriori estimators for the pointwise gradient error on each element in irregular meshes. II. The piecewise linear case*, Math. Comp. **73** (2004), no. 246, 517–523 (electronic). MR MR2028417 (2004i:65127)

[SZ90] L. Ridgway Scott and Shangyou Zhang, *Finite element interpolation of non-smooth functions satisfying boundary conditions*, Math. Comp. **54** (1990), no. 190, 483–493. MR MR1011446 (90j:65021)

[Tho06] Vidar Thomée, *Galerkin finite element methods for parabolic problems*, second ed., Springer Series in Computational Mathematics, vol. 25, Springer-Verlag, Berlin, 2006. MR MR2249024 (2007b:65003)

[VB96] Lieven Vandenberghe and Stephen Boyd, *Semidefinite programming*, SIAM Rev. **38** (1996), no. 1, 49–95. MR MR1379041 (96m:90005)

[Ver94a] R. Verfürth, *A posteriori error estimates for nonlinear problems. Finite element discretizations of elliptic equations*, Math. Comp. **62** (1994), no. 206, 445–475. MR MR1213837 (94j:65136)

[Ver94b] ———, *A posteriori error estimation and adaptive mesh-refinement techniques*, Proceedings of the Fifth International Congress on Computational and Applied Mathematics (Leuven, 1992), vol. 50, 1994, pp. 67–83. MR MR1284252 (95c:65171)

[Ver96] Rüdiger Verfürth, *A review of* a posteriori *error estimation and adaptive mesh-refinement techniques*, Wiley-Teubner, Chichester-Stuttgart, 1996.

[Ver03] R. Verfürth, *A posteriori error estimates for finite element discretizations of the heat equation*, Calcolo **40** (2003), no. 3, 195–212. MR MR2025602 (2005f:65131)

[VMD+07] M.-G. Vallet, C.-M. Manole, J. Dompierre, S. Dufour, and F. Guibault, *Numerical comparison of some Hessian recovery techniques*, Internat. J. Numer. Methods Engrg. **72** (2007), no. 8, 987–1007. MR MR2360556 (2008k:65187)

[Whe73] Mary Fanett Wheeler, *A priori L_2 error estimates for Galerkin approximations to parabolic partial differential equations*, SIAM J. Numer. Anal. **10** (1973), 723–759. MR 50 #3613

[XZ04] Jinchao Xu and Zhimin Zhang, *Analysis of recovery type a posteriori error estimators for mildly structured grids*, Math. Comp. **73** (2004), no. 247, 1139–1152 (electronic). MR MR2047081 (2005f:65141)

[Zha01] Zhimin Zhang, *A posteriori error estimates on irregular grids based on gradient recovery*, Adv. Comput. Math. **15** (2001), no. 1-4, 363–374 (2002), A posteriori error estimation and adaptive computational methods. MR MR1887740 (2003a:65092)

[Zlá77] Miloš Zlámal, *Some superconvergence results in the finite element method*, Mathematical aspects of finite element methods (Proc. Conf., Consiglio Naz. delle Ricerche (C.N.R.), Rome, 1975), Springer, Berlin, 1977, pp. 353–362. Lecture Notes in Math., Vol. 606. MR MR0488863 (58 #8365)

[ZW98] S. Ziukas and N.-E. Wiberg, *Adaptive procedure with superconvergent patch recovery for linear parabolic problems*, Finite element methods (Jyväskylä, 1997), Lecture Notes in Pure and Appl. Math., vol. 196, Dekker, New York, 1998, pp. 303–314. MR MR1602726

[ZZ87] O. C. Zienkiewicz and J. Z. Zhu, *A simple error estimator and adaptive procedure for practical engineering analysis*, Internat. J. Numer. Methods Engrg. **24** (1987), no. 2, 337–357. MR MR875306 (87m:73055)

[ZZ95] Zhimin Zhang and J. Z. Zhu, *Superconvergence of the derivative patch recovery technique and a posteriori error estimation*, Modeling, mesh generation, and adaptive numerical methods for partial differential equations (Minneapolis, MN, 1993), IMA Vol. Math. Appl., vol. 75, Springer, New York, 1995, pp. 431–450. MR MR1370261 (97i:65170)

Index