University of Reading

Adaptive numerical techniques for problems related to flow in porous media

Ben Ashby

Department of Mathematics and Statistics

Thesis submitted for the degree of
Doctor of Philosophy

June 2022

Abstract

The solution of partial differential equations modelling water infiltration into soil poses many challenges. The multi-scale and nonlinear nature of soil makes the design of robust and accurate numerical schemes particularly difficult. In addition, error estimation is complicated by low solution regularity. In this thesis, we investigate the mathematical and numerical aspects of the approximation of problems related to subsurface flow by the finite element method. We begin with a variational inequality as a simplified model (albeit of significant interest and complexity in its own right) of a seepage problem. The so-called Signorini problem includes many of the key difficulties, namely nonlinear boundary conditions and lack of dual regularity. We derive rigorous and computable a posteriori error estimates using duality arguments that require careful analysis of primal and dual problems. Crucial in this argument is the design of a novel nonlinear bound-preserving interpolant that respects various inequalities related to the weak form of the problem. These estimates are used to implement a mesh adaptive routine. We then study a physically realistic seepage problem complete with nonlinear coefficients and mixed boundary conditions and inequality constraints. This time, we apply the dual-weighted residual framework of a posteriori error estimation and derive error estimates that are used to optimise the computational mesh for a quantity of interest. The estimates are tested on realistic groundwater scenarios that utilise field data. We conclude with a numerical study of a time-dependent and nonlinear model of two-dimensional subsurface flow. We introduce a method to regularise the nonlinearity in the soil porosity function and derive a posteriori error estimates that account for this approximation in linear elliptic and parabolic cases. We show that in the nonlinear parabolic case, this regularisation mitigates the commonly observed failure of nonlinear solvers for Richards' equation.

Declaration

I confirm that this is my own work and the use of all material from other sources has been properly and fully acknowledged.

Ben Ashby

Acknowledgements

I would most of all like to thank Tristan Pryer for going above and beyond as my PhD supervisor, always giving his time for mathematical support and prudent advice on all aspects of being a mathematician. His endless patience and constructive criticism made this work possible. I would also like to thank my additional supervisors: Alex Lukyanov, this work benefits from his expertise in many different fields, and Cassiano Bortolozo, whose input on geophysics and provision of data were extremely valuable. I am grateful to all students, staff and alumni of the MPECDT, who maintained a support network through nearly two years of lockdowns and uncertainty. Thank you to my parents. Their continuing support is the reason I am where I am today. Finally, to my wife Hannah, thank you for everything.

Contents

Chapter 1

Introduction

The work in this thesis is broadly concerned with the adaptive solution of nonlinear partial differential equations by the finite element method. In each case, the problem to be solved shares key characteristics with models of fluid flow in porous media. Several rather different aims and approaches are considered in the chapters that make up this thesis. For this reason, some of the later chapters will contain a detailed literature review specific for the material studied within.

1.1 A posteriori error estimation and adaptivity

To begin, we give a broad overview of a posteriori estimation and adaptive strategies for finite element methods.

Finite element methods are well suited to adaptive strategies due to their ability to handle unstructured meshes without difficulty, allowing meshes to be updated dynamically during a calculation. In addition, the mathematical framework of finite element methods provides theoretical basis for a posteriori error estimation. Finite element methods are used to approximate solutions to weak problems posed in Hilbert spaces using numerically tractable subspaces such as piecewise polynomial spaces. This allows the use of results from functional analysis and PDE theory to analyse the full problem and to be combined with results on polynomial approximation to quantify error. In keeping with the finite element philosophy of approximating solution spaces rather than PDE operators, we can thus view an adaptive finite element method as a sequence of subspaces which approximate the Hilbert space containing the true solution more and more closely.

The theory of a posteriori error estimation and adaptive methods is now rather well developed for elliptic problems but there are still many gaps for related problems such as variational inequalities based upon elliptic operators. A broad overview of the many different approaches is given in [110], and a rather comprehensive collection on results for elliptic problems is given in [2]. We discuss a brief selection here.

There are several ways that adaptivity can be incorporated into computations. Broadly speaking, one can use a priori knowledge in mesh design, heuristic 'smoothness' indicators, or error bounds obtained by a posteriori analysis. One must use one of the latter methods if automatic mesh refinement is required with no input from the user.

Broadly speaking, in finite element computations, automatic adaptivity means either altering the local approximation order (p-adaptivity) typically done by enriching approximation spaces with higher order polynomial, moving mesh nodes to alter the local resolution and capture solution features (r-adaptivity) to optimise for a specified error metric, or adding and removing degrees of freedom to change the local resolution (h-adaptivity). In this thesis we will only consider h-adaptivity.

As an example of a priori mesh design, in aerofoil problems meshes are often designed with high resolution around the wing to resolve the boundary layer and coarse resolution in the laminar flow away from the wing. Examples of heuristic error indicators include the popular Kelly indicator [63] which uses the size of the jump of the flux across element boundaries as a measure of smoothness, norms of vorticity or pressure gradient in computations with fluids [14]. Estimates of the solution gradient obtained via patch recovery can be used to inform refinement around sharp interfaces, and perform well in numerical simulations [56]. Many of these heuristic indicators are prompted by a posteriori analysis on some level, for example the Kelly indicator gives an upper bound on the error for Laplace's equation up to higher order pertubation terms [29].

A common way to test numerical solutions to PDEs for accuracy when no exact solution is available is to compare to a 'better' solution. Here, 'better' usually means approximation on a finer mesh or with a higher order method. This leads to the idea of *hierarchical error estimation* (see [42, 2]). Using the standard error equation in terms of the residual, a Galerkin approximation of the error is sought in an enriched (i.e. larger) function space. Since this would involve solving a larger problem than the original calculation, the finer solution is not actually calculated but approximated to reduce the amount of computation needed. The original finite element space is extended via a direct sum, and the error is estimated in the orthogonal component to the original space to simplify calculations, ignoring any coupling between the two spaces. For this to be a reasonable approximation, assumptions must be made on the degree of orthogonality on the spaces. The reader is referred to [10] for more details on this form of error estimation, and also to [99] for an approach based more upon applicability to a wide variety of problems without having to derive PDE-specific estimates.

More recently, goal-oriented a posteriori error estimation (also known as the dual-weighted residual technique) has become a popular and successful technique, especially in CFD [87]. Here, the focus turns away from estimating global measures of error (such as norms) and towards a specific quantity of interest (or 'goal'). The technique uses the solution to an adjoint PDE to set up an error representation consisting of residuals of the finite element solution weighted by the dual solution. This is then evaluated by approximately solving the dual problem and inserting this numerical solution into the error representation to obtain a computable estimate. The key difference in this case is that the result is not necessarily an upper bound. The method of approximation does not guarantee anything, however the additional information given by values of the dual solution can often provide more accurate information on where to refine, and in addition is capable of targeting mesh refinement in areas that specifically increase accuracy in the target quantity. It is for this reason that goal-oriented adaptivity has proven effective in a variety of computations. For the theory and some applications of this method, see for example [54, 15, 14, 8, 37].

For time dependent problems the difficulties are greater, both analytically and computationally. While in stationary problems one can refine until a tolerance is met, this becomes more difficult when solving for many time steps, as the amount of data that must be stored can become prohibitive. Singularities that are relatively easily resolved in the stationary case can move in space. In addition, errors can arise from changing the mesh, since when evaluating difference quotients needed for the discretisation of unsteady problems, we must compare approximate solutions defined upon different meshes. One must therefore be transferred to the newer mesh, and if the newer mesh has been coarsened, information is lost as degrees of freedom are removed.

An important work in the literature of a posteriori techniques for parabolic problems is the series of papers beginning with [46], where a dual problem is used to quantify error propagation in simulations. In simple cases (linear, constant coefficients) it was shown that analytical stability bounds can be obtained for the dual problem which are then used to derive sharp bounds on the error. However, for more complex problems it is not always possible, and the authors acknowledge that a certain degree of computation may be needed to fill in the theoretical gaps, such as computational estimates of stability constants.

Estimates for parabolic problems can sometimes be obtained by utilising results from the elliptic theory and exploiting the fact that many parabolic problems are of the form

$$u_t - \mathcal{A}u = f \tag{1.1}$$

where $\mathcal{A}$ is an elliptic operator, f is problem data and u is the solution, chosen to be in a suitable function space. This is known as the *elliptic reconstruction* technique of [70], see also [104]. The technique was introduced to circumvent the shortcoming of energy methods for parabolic problems, which provide a posteriori error estimates of suboptimal order in $\mathrm{L}^\infty - \mathrm{L}^2$. The error bound one obtains from an elliptic reconstruction argument involves a spatial elliptic estimator which accumulates in time and a contribution from the initial condition. These temporal accumulations will inevitably grow rendering it impossible to control error in general, and so a common approach is to work in an L^∞ norm in time [79, 104, 46].

The goal-oriented approach has also been applied to parabolic problems, but in this case the analysis and implementation are more difficult. The adjoint (backward in time) problem in this case needs to be solved over the entire time domain before error estimates can be calculated, which is a huge computational burden. In addition, it is not known how to prove well-posedness of the dual problem in many cases [46] (nor indeed if it is even true). Nevertheless, very good results can be obtained for relatively short simulations [96].

The computational costs may be amplified again for nonlinear problems, where many linear problems may have to be solved at each time step as part of a nonlinear iteration scheme. There are a posteriori results on nonlinear elliptic problems, but these often have to make assumptions on the problem that are violated in practice, see the discussion of regular solutions in §5.1.2 of [110]. The dual-weighted residual (DWR) approach is applicable to nonlinear problems, albeit with the same issues with well-posedness of the dual problem [8].

1.2 Aims of the thesis

The motivation for the work in this thesis stems from collaboration with CEMADEN (National Centre for Natural Disaster Monitoring and Alerts), Brazil. Their broad aim is to develop a landslide prediction mechnism based upon models of soil infiltration combined with data measured in the field. The broader physics and design principles are fully described in the forthcoming book [6]. Much of the underlying mathematical theory is developed here. In particular, we investigate the application of adaptive techniques to problems relevant to groundwater flow. Much of the the work on a posteriori error estimation outlined above is not directly applicable to such problems. For the variational inequlities of chapters 3 and 4, the loss of Galerkin orthogonality and dual regularity complicate matters. Richards' equation with constitutive relations that are highly nonlinear lead to practical problems such as convergence failure or spurious oscillations.

1.3 Structure of thesis

The rest of the thesis is structured as follows. In chapter 2, we present material that is required for later chapters. We fix notation and introduce elliptic boundary value problems and variational inequalities, with the fundamental theoretical results for each.

In chapter 3, we study a relatively simple variational inequality for which the nonlinearity of the problem arises from inequality constraints at the boundary. Our motivation is that similar constraints arise from the interaction of subsurface water with the atmosphere and bodies of water. We use recent work on a priori analysis of finite element solutions to this problem in low order norms given in [34] to derive a rigorous a posteriori error estimate which is, to the best of our knowledge, new.

The standard machinery for deriving bounds in low order norms is the so called Aubin-Nitsche duality argument, which is not immediately applicable to the Signorini problem, as the dual solution is not sufficiently regular. In addition, the usual Galerkin orthogonality does not hold for variational inequalities, and an analogous result must be derived to quantify the interaction of the error $u - u_h$ and the finite element space - in this case an inequality. Recent regularity results for the Signorini problem and an appropriate dual problem provide just enough regularity to eliminate the dual solution from the error representation, giving a computable estimate in L^4.

Numerical experiments verify the validity of this error estimate, which is then used as a criterion for adaptive mesh refinement. The resulting adaptive algorithm is shown to give better accuracy per degree of freedom than solving on uniform grids. We also investigate the performance or the error estimate in cases where the theory is not applicable due to decreased regularity, namely in three dimensions and on non-convex domains. Although optimal rates of convergence are lost in this case (as expected) the error estimate is still shown to be useful at reducing computational effort.

In chapter 4 we introduce a more complete version of the practical application that motivates the previous theoretical study of variational inequalities in chapter 3. The problem is variably saturated subsurface flow of water around a well. This is modelled by a nonlinear and possibly degenerate elliptic problem with boundary constraints analogous to those in the Signorini problem. In this case the constraints represent the interaction of the pore water with that in the well. The problem specification includes surface and bedrock conditions, which necessitates the ability to include a mixture of Neumann, Dirichlet and Signorini boundary conditions. Another key difficulty is the permeability function of the subsurface. This is commonly modelled with a strongly nonlinear and possible non-Lipschitz diffusion coefficient. This function may also vary by several orders of magnitude across the domain due to heterogeneity in the subsurface. Due to all of these factors we do not have the same regularity results as before. It is however crucial to resolve features of the solution such as contact sets (which are known as seepage faces for this application) and areas of strong permeability gradients such as layers in the subsurface and regions which are approaching saturation.

We utilise the dual-weighted residual framework for a posteriori error estimation to derive an error bound on a key quantity of interest. We make use of a pseudo-linearised dual problem that is able to capture the problem behaviour around the contact set, and use this to set up an error inequality that replaces the usual dual error representation in duality-based error analysis. We then make use of an intermediate function which satisfies an elliptic problem without the constraints on the boundary, allowing the introduction of the problem data into the error bound.

Our numerical experiments demonstrate the need for adaptivity by computing a quantity of interest from the discrete solution in a range of realistic test cases including heterogeneous soil structures. In this case it is particularly important to balance mesh refinement around the seepage face with resolving the heterogeneity in the soil.

In the final part of the thesis, chapters 5 and 6, we move on to study time-dependent subsurface flow modelled with Richards' equation. The equation is one of the most widely used to model the flow of a water-air mixture in a porous medium and has a large range of practical applications. However, solving it numerically in practical situations still has many difficulties associated with it. Even with modern computing, the nonlinear nature of the equation can be vastly expensive or even impossible to solve, due to the resolution required to represent singularities in the nonlinear terms, slow nonlinear solver convergence or failure to converge at all. The purpose of the work in this part of the thesis is to try to isolate the reasons for these problems and propose tools to mitigate against them. The hydraulic conductivity function, which in certain cases of practical interest is not even Lipschitz continous, is observed to be a principal factor in slow convergence or failure. A regularisation of this function is proposed, controlled by a parameter that allows it to be applied locally based on an indicator.

In chapter 6, error analysis is conducted for simple model problems to investigate the effect of this regularisation. Error indicators are derived which combine the usual finite element error with 'model' error that arises from the regularisation of the coefficient.

This motivates an indicator can be combined with more standard error indicators for Richards' equation, commonly referred to spatial and temporal indicators, to implement a space-time adaptive algorithm that increases computational efficiency and protects against convergence failure.

Chapter 2

Elliptic partial differential equations and variational inequalities

2.1 Abstract

In this chapter we present the necessary material for later chapters. We use the opportunity to introduce elliptic boundary value problems and related variational inequalities. Since it will be important for later analysis to be able to quantify the regularity of solutions, we discuss elliptic regularity and analogous results for variational inequalities. We also include some technical results on finite element approximation here for convenience. We finally present a simple case of a posteriori error estimation via duality to illustrate the key ideas of this approach and introduce the reader to the difficulties in the case of a variational inequality.

2.2 Function spaces

In this section we introduce Lebesgue and Sobolev spaces. Sobolev spaces are the appropriate choice in which to seek weak solutions of the partial differential equations studied in this thesis under minimal regularity requirements (see e.g. 5.1 of [49] or the discussion in 1.1 of [52]). We also take the opportunity to introduce notation that will be used throughout.

Throughout this work, we let $\Omega \subseteq \mathbb{R}^N$, $N = 2$ or 3 be a bounded, convex domain with boundary $\partial\Omega$ which we assume to be polygonal.

Definition 2.2.1 (Lebesgue spaces)**.** *Let* $1 \leqslant p < \infty$*. We denote by* $\mathrm{L}^p(\Omega)$ *the Lebesgue space consisting of measurable functions* v *such that*

$$\|v\|_{\mathrm{L}^p(\Omega)} := \left(\int_\Omega |v|^p \,\mathrm{d}x\right)^{1/p} < \infty. \tag{2.1}$$

The space $\mathrm{L}^p(\Omega)$ *is a normed vector space with norm* $\|\cdot\|_{\mathrm{L}^p(\Omega)}$ *defined in* (2.1).

We also define the space $\mathrm{L}^\infty(\Omega)$ *to be the space of measurable functions* v *such that*

$$\|v\|_{\mathrm{L}^\infty(\Omega)} := \inf\{M \geqslant 0, \textit{ s.t. } |v| \leqslant M \textit{ almost everywhere in } \Omega\}. \tag{2.2}$$

For all p, $\mathrm{L}^p(\Omega)$ is a Banach space, and $\mathrm{L}^p(\Omega)$ is a Hilbert space if and only if $p = 2$ with inner product

$$\langle u, v\rangle = \int_\Omega uv \,\mathrm{d}x$$

for $u, v \in \mathrm{L}^2(\Omega)$. Let A be a measurable subset of Ω. Then we define

$$\langle u, v\rangle_A = \int_A uv \,\mathrm{d}x.$$

If A is a subset of the boundary $\partial\Omega$ we interpret $\langle u, v\rangle_A$ as a line integral if $N = 2$ and a surface integral if $N = 3$.

Remark 2.2.2 (Dual spaces of Lebesgue spaces)**.** *For $1 < p < \infty$, we define the Hölder conjugate of p to be the number q such that $1/p+1/q = 1$. Then the dual space of $\mathrm{L}^p(\Omega)$ is isomorphic to $\mathrm{L}^q(\Omega)$. Via the natural isomorphism, an element $v \in \mathrm{L}^q(\Omega)$ maps to the functional on $\mathrm{L}^p(\Omega)$ defined by $w \mapsto \int_\Omega w\, v \,\mathrm{d}x$. This integral represents a* duality pairing *between $\mathrm{L}^p(\Omega)$ and $\mathrm{L}^q(\Omega)$ which will be exploited in our error estimates. The situation is more subtle with the pair $p = 1$, $q = \infty$, and will not be discussed here.*

Definition 2.2.3 (Sobolev spaces)**.** *Let $1 \leqslant p \leqslant \infty$ and let k be a non-negative integer. Then we introduce the Sobolev space*

$$\mathrm{W}^{k,p}(\Omega) = \{v \in \mathrm{L}^p(\Omega) \,|\, \partial^{\boldsymbol{\alpha}} v \in \mathrm{L}^p(\Omega) \text{ for } |\boldsymbol{\alpha}| \leqslant k\}, \tag{2.3}$$

where $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_N)$ is a multi-index, $|\boldsymbol{\alpha}| = \sum_i \alpha_i$ and $\partial^{\boldsymbol{\alpha}} = \partial_{x_1}^{\alpha_1} ... \partial_{x_N}^{\alpha_N}$ is a (weak) derivative of order $|\boldsymbol{\alpha}|$.

We equip the Sobolev spaces with the norm

$$\|u\|_{\mathrm{W}^{s,p}(\Omega)} = \left(\sum_{|\alpha| \leqslant s} \|\partial^\alpha u\|^p_{\mathrm{L}^p(\Omega)} \right)^{1/p} \tag{2.4}$$

and the seminorm

$$|u|_{\mathrm{W}^{s,p}(\Omega)} = \left(\sum_{|\alpha| = s} \|\partial^\alpha u\|^p_{\mathrm{L}^p(\Omega)} \right)^{1/p} \tag{2.5}$$

With the norm above, $\mathrm{W}^{s,p}(\Omega)$ is a Banach space. In the case $p = 2$, $\mathrm{W}^{s,p}(\Omega)$ is a Hilbert space, and it is customary to define special notation, namely

$$\mathrm{H}^s(\Omega) := \mathrm{W}^{s,2}(\Omega). \tag{2.6}$$

Let $C^0(\Omega)$ denote the space of continuous functions, and for integer $k \geqslant 1$, $C^k(\Omega)$ is the space of functions that are continuously differentiable k times. We also let $C_c^\infty(\Omega)$ denote the space of infinitely differentiable functions which are compactly supported in Ω.

Definition 2.2.4 (Hölder spaces)**.** *We define $C^{0,\alpha}(\Omega)$ for $0 \leqslant \alpha \leqslant 1$ to be the space of functions v that are Hölder continuous with exponent α, which means that v is continuous and that*

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

Being a member of a Sobolev space and therefore having integrable derivatives is a strong condition, and guarantees membership in certain other Sobolev, Lebesgue or Hölder spaces depending on the order of derivatives, exponent, spatial dimension and sometimes properties of the domain. The fundamental theorems of Sobolev and Morrey give the various embeddings of Sobolev and Lebesgue spaces needed for later work, and are summarised in the following theorem. Proofs of the theorems can be found in [98] and [22] respectively.

Theorem 2.2.5 (Sobolev embeddings)**.** *Let p, $q \geqslant 1$, and let p^* be the number such that $\frac{1}{p*} = \frac{1}{p} - \frac{1}{N}$. If either*

- *$p < N$ and $p \leqslant q \leqslant p^*$;*
- *$p = N$ and $p \leqslant q < \infty$;*

then

$$\mathrm{W}^{1,p}(\mathbb{R}^N) \subset \mathrm{L}^q(\mathbb{R}^N)$$

with continuous embedding. That is, there exists a constant $C_{Sob} > 0$ such that

$$\|v\|_{\mathrm{W}^{1,p}(\mathbb{R}^N)} \leqslant C_{Sob} \|v\|_{\mathrm{L}^q(\mathbb{R}^N)}. \tag{2.7}$$

If Ω is a bounded domain with Lipschitz boundary, $s \geqslant 1$ is an integer and if $p > \frac{s}{N}$ then

$$\mathrm{W}^{s,p}(\Omega) \subset \mathrm{L}^{\infty}(\Omega) \cap C^{0,\alpha}(\bar{\Omega})$$

with continuous embedding, where $\alpha = 1 - \frac{N}{sp}$. This latter result is known as Morrey's inequality, while (2.7) is a special case of Sobolev's embedding theorem.

Remark 2.2.6. *The final statement gives sufficient conditions for members of Sobolev spaces to have a continuous representative. In this case, we can consider these functions to have point values, so that it makes sense to talk of level sets of a function. In addition, we will need well-defined point values later in this chapter to define the Lagrange interpolant.*

Any function that is continuous on $\bar{\Omega}$ has well-defined boundary values. The operator that maps a function to its boundary values in a classical sense can be extended continuously to $\mathrm{W}^{1,p}(\Omega)$, giving meaning to the boundary values of a function (the trace) even though $\mathrm{W}^{1,p}(\Omega)$ may not embed into the continuous functions for $N > 1$. The trace operator construction is demonstrated in [49], section 5.5 for smooth domains.

Definition 2.2.7 (Sobolev space of functions with compact support.)**.** $\mathrm{W}^{s,p}_0(\Omega)$ *is defined to be the closure of $C^{\infty}_c(\Omega)$ in $\mathrm{W}^{s,p}(\Omega)$. The trace theorem as presented in [48] (Theorem B.52) characterises $\mathrm{W}^{1,p}_0(\Omega)$ as the space of functions in $\mathrm{W}^{1,p}(\Omega)$ whose trace is zero. As an important special case, we mention $\mathrm{H}^1_0(\Omega)$, whose dual with respect to the $\mathrm{L}^2(\Omega)$ inner product is denoted $H^{-1}(\Omega)$.*

2.3 Elliptic boundary value problems

With the necessary function spaces outlined above, we may now introduce an elliptic Dirichlet problem and its weak form.

We first state the strong form of an elliptic partial differential equation. We seek a function $u \in \mathrm{H}^2(\Omega) \cap \mathrm{H}^1_0(\Omega)$ such that

$$\begin{aligned} -\Delta u + u &= f \quad \text{in } \Omega \\ u &= 0 \quad \text{on } \partial\Omega. \end{aligned} \tag{2.8}$$

The regularity requirement $u \in \mathrm{H}^2(\Omega)$ is not always appropriate, and in many cases there is no solution in this space, for example if the domain Ω has a reentrant corner. We obtain the *weak* form of (2.8) by testing with $v \in \mathrm{H}^1_0(\Omega)$ and integrating by parts. The problem becomes: seek $u \in \mathrm{H}^1_0(\Omega)$ such that

$$\int_\Omega (\nabla u \cdot \nabla v + uv) \, \mathrm{d}x = \int_\Omega fv \, \mathrm{d}x \quad \forall v \in \mathrm{H}^1_0(\Omega). \tag{2.9}$$

We now note that the minimum regularity requirements on problem data for the above equation to make sense are $f \in \mathrm{H}^{-1}(\Omega)$. We write the problem in a more compact form by defining

$$a(u, v) := \int_\Omega (\nabla u \cdot \nabla v + uv) \, \mathrm{d}x. \tag{2.10}$$

Hence the problem becomes: find $u \in \mathrm{H}^1_0(\Omega)$ such that

$$a(u, v) = \langle f, v \rangle \quad \forall v \in \mathrm{H}^1_0(\Omega). \tag{2.11}$$

Remark 2.3.1 (Equivalence of formulations)**.** *It can be shown that if u solves (2.11), and is sufficiently regular then it satisfies (2.8) almost everywhere. The regularity can be guaranteed by making assumptions on the data f and the domain Ω. Indeed, for $N = 2$, $f \in \mathrm{L}^2(\Omega)$ and Ω a convex polygon, we have $u \in \mathrm{H}^2(\Omega)$.*

2.3.1 Well-posedness and regularity

Scalar elliptic problems such as (2.9) have a coercivity property, which is a key component of their solubility and stability analysis, as well as numerical analysis.

Definition 2.3.2 (Coercivity and boundedness)**.** *Let $\mathcal{V}$ be a Hilbert space, and $a(\cdot,\cdot)$ a bilinear form on $\mathcal{V}$. We say that $a(\cdot,\cdot)$ is bounded if there exists a constant $C_1 > 0$ such that for all $u, v \in \mathcal{V}$*

$$a(u,v) \leqslant C_1 \left\|u\right\|_{\mathcal{V}} \left\|v\right\|_{\mathcal{V}},$$

and that a is coercive if there exists $C_0 > 0$ such that for all $v \in \mathcal{V}$

$$a(v,v) \geqslant C_0 \left\|v\right\|_{\mathcal{V}}^2.$$

Remark 2.3.3. *It is easy to see that the choice*

$$a(u,v) := \int_\Omega (\nabla u \cdot \nabla v + uv) \,\mathrm{d}x \tag{2.12}$$

is bounded and coercive on $\mathrm{H}_0^1(\Omega)$. Indeed, for boundedness, we calculate

$$\begin{aligned} a(u,v) &= \int_\Omega uv + \nabla u \cdot \nabla v \,\mathrm{d}x \\ &\leqslant \left\|u\right\|_{\mathrm{L}^2(\Omega)} \left\|v\right\|_{\mathrm{L}^2(\Omega)} + \left\|\nabla u\right\|_{\mathrm{L}^2(\Omega)} \left\|\nabla v\right\|_{\mathrm{L}^2(\Omega)} \\ &\leqslant \left(\left\|u\right\|_{\mathrm{L}^2(\Omega)}^2 + \left\|\nabla u\right\|_{\mathrm{L}^2(\Omega)}^2\right)^{1/2} \left(\left\|v\right\|_{\mathrm{L}^2(\Omega)}^2 + \left\|\nabla v\right\|_{\mathrm{L}^2(\Omega)}^2\right)^{1/2} \\ &= \left\|u\right\|_{\mathrm{H}^1(\Omega)} \left\|v\right\|_{\mathrm{H}^1(\Omega)} \end{aligned} \tag{2.13}$$

so that $a(\cdot,\cdot)$ is bounded with $C_1 = 1$. We also have

$$\begin{aligned} \left\|u\right\|_{\mathrm{H}^1(\Omega)}^2 &= \left(\left\|u\right\|_{\mathrm{L}^2(\Omega)}^2 + \left\|\nabla u\right\|_{\mathrm{L}^2(\Omega)}^2\right) \\ &= a(u,u) \end{aligned} \tag{2.14}$$

and therefore $a(\cdot,\cdot)$ is coercive with $C_0 = 1$.

The coercivity and boundedness properties of $a(\cdot,\cdot)$, as well as the regularity on f allows the application of the Lax-Milgram lemma to problem (2.9) (see for example Theorem 1.1.3 of [35]). This ensures that problem (2.9) has a unique solution.

Proposition 2.3.4 (Elliptic regularity)**.** *Let u be the solution of* (2.11) *and let m be a non-negative integer. Suppose that $f \in \mathrm{H}^m(\Omega)$. Then $u \in \mathrm{H}^{m+2}(\Omega)$ (see [49, §6.3.1] for smooth domains, with many results extended to nonsmooth domains in [52, eq (4.1.2)]). That is, there exists a constant $C_{reg} = C_{reg}(m,\Omega)$ such that*

$$\|u\|_{\mathrm{H}^{m+2}(\Omega)} \leqslant C_{reg} \|f\|_{\mathrm{H}^m(\Omega)} . \tag{2.15}$$

2.4 Elliptic variational inequalities

Variational inequalities are a generalisation of the variational problems of the previous section. In this section, we give a brief discussion of where these problems arise from physics, then present a mathematical formulation.

The classical Signorini problem arises in elasticity. In this case the solution variable is vector displacement of an elastic body. Let us consider a linearly elastic solid body resting on a foundation that we assume to be rigid. The rigid foundation limits the deformation of the body (since it cannot pass through). At the contact boundary (which is unknown a priori) the stress in the body must satisfy conditions of equilibrium. These conditions are linearised under the assumption that any displacements of the body are small to arrive at inequality constraints at the boundary.

In this section, we will present a simplified version of the Signorini problem which encapsulates the key analytical difficulties compared to the analysis and approximation of boundary value problems. The problem is sometimes referred to as the scalar Signorini problem [34] or the unilateral problem

[24], however these terms are rather flexible in the literature and can refer to scalar obstacle problems where inequalities must be satisfied on the whole domain, or on the boundary only.

We are motivated by the case where the constraints are present on the boundary, not only because it is a well-studied prototype in the mathematical literature, but also because such problems can arise in porous media flow problems which will be the focus of §4.

2.4.1 The scalar Signorini problem

We choose to first present an abstract variational form of the Signorini problem, then show equivalence with a concrete strong form.

Set $\mathcal{V} = \mathrm{H}^1(\Omega)$, and as above let $a(\cdot,\cdot)$ be the coercive and bounded bilinear form on $\mathcal{V}$ defined by (2.12).

Let $\mathcal{K}$ be a closed, convex subset of $\mathrm{H}^1(\Omega)$ and $f \in \mathrm{L}^2(\Omega)$. We consider the variational inequality for $u \in \mathcal{K}$,

$$a(u, v-u) \geqslant \langle f, v-u \rangle \quad \forall v \in \mathcal{K}. \tag{2.16}$$

It is well known that this problem has a unique solution (see [68]).

Remark 2.4.1. *Variational inequalities generalise variational problems in the sense that if $\mathcal{K} = \mathcal{V}$, the problem reduces to that of finding $u \in \mathcal{V}$ such that*

$$a(u, v) = \langle f, v \rangle \quad \forall v \in \mathcal{V}.$$

In [68] the proof is given under minimal assumptions, and the bilinear form a is not required to be symmetric. In our applications, though it will be.

We now specialise to a particular variational inequality. Let $\mathcal{K}$ be the convex set

$$\mathcal{K} := \{v \in \mathrm{H}^1(\Omega) \mid v \geqslant 0 \text{ on } \partial\Omega\}.$$

Then (2.16) is the weak form of the following problem known as the scalar Signorini problem - a partial differential equation with inequality constraints on the boundary.

$$-\Delta u + u = f \text{ in } \Omega, \tag{2.17}$$

coupled with boundary conditions

$$u \geqslant 0, \quad \nabla u \cdot \boldsymbol{n} \geqslant 0, \quad u \nabla u \cdot \boldsymbol{n} = 0 \text{ on } \partial\Omega. \tag{2.18}$$

Proposition 2.4.2 (Equivalence of weak and strong forms)**.** *If u solves* (2.17) - (2.18) *almost everywhere, then it solves* (2.16)*. Conversely, if u solves* (2.16) *and is sufficiently smooth, then it solves* (2.17) - (2.18)*.*

Proof. We begin by supposing that u solves the strong form (2.17) - (2.18). Testing (2.17) with arbitrary v and subtracting the result of testing with u, one obtains

$$a(u, v-u) = \langle f, v-u \rangle + \int_{\partial\Omega} (v-u)\partial_n u \, \mathrm{d}S. \tag{2.19}$$

Since $v \in \mathcal{K}$ and u satisfies (2.18), the integral term on $\partial\Omega$ must be non-negative, and therefore for any $v \in \mathcal{K}$ we have

$$a(u, v-u) \geqslant \langle f, v-u \rangle .$$

Conversely, suppose now that u is the solution of the weak problem (2.16). Let w be a smooth function compactly supported in Ω. Then $u + w$ is a member of $\mathcal{K}$ and we may choose $v = u + w$ in (2.16), and since $u \in \mathrm{H}^2(\Omega)$ we may integrate $a(u, w)$ by parts to see that

$$\int_{\Omega} w\,(-\Delta u + u - f)\,\mathrm{d}x \geqslant 0, \tag{2.20}$$

where we note that the boundary terms vanish due to the properties of w.

After taking $u - w$ as test function, we also have

$$\int_{\Omega} w\left(-\Delta u + u - f\right) \mathrm{d}x \leqslant 0, \tag{2.21}$$

and therefore equality must hold. Since this is true for all smooth w compactly supported on Ω, $-\Delta u + u = f$ almost everywhere in Ω.

We now observe that the constraint $u \geqslant 0$ at the boundary is automatically satisfied due to the choice of function space in the weak formulation. Further, if a smooth function w satisfies $w \geqslant 0$ then $u + w \in \mathcal{K}$ and we can make this choice as test funtion in (2.16). After integrating by parts and using the fact that $-\Delta u + u - f = 0$ almost everywhere, we are left with

$$\int_{\partial\Omega} w \partial_n u \, \mathrm{d}S \geqslant 0,$$

which is true for any smooth $w \geqslant 0$. We therefore must have $\partial_n u \geqslant 0$.

We finally suppose that at some point $x_0 \in \partial\Omega$ we have $u > 0$. We let w be a smooth, non-positive function such that $w(x_0) < 0$ but $u + w \geqslant 0$ on $\partial\Omega$ and $(u + w)(x_0) > 0$, so that we can again choose $v = u + w$ in (2.16) to see that

$$\int_{\partial\Omega} w \partial_n u \, \mathrm{d}S \leqslant 0.$$

Thus, $\partial_n u = 0$, and we have shown that $u \partial_n u = 0$ on $\partial\Omega$. □

It will be important in our analysis to quantify the regularity of the solution of this problem. To state the key result, we must make the following definitions, and reformulate the nonlinear boundary condition.

Definition 2.4.3 (Subdifferential)**.** *Suppose that φ is a convex function, $\varphi : \mathbb{R} \to (-\infty, \infty]$, and let $D(\varphi) = \{x \in \mathbb{R} \mid \varphi(x) < \infty\}$. Then the subdifferential $\partial\varphi(x)$ at $x \in D(\varphi)$ is the set*

$$\partial\varphi(x) = \{y \in \mathbb{R} \mid \varphi(z) - \varphi(x) \geqslant y(z - x) \, \forall z \in D(\varphi)\}. \tag{2.22}$$

Definition 2.4.4 (Monotone operator on the real line.)**.** *Let E be a mapping from $\mathbb{R}$ to $2^{\mathbb{R}}$, the set of all subsets of $\mathbb{R}$. E is said to be monotone if $(y_1 - y_2)(x_1 - x_2) \geqslant 0$ for all $x_1, x_2 \in \mathbb{R}$ such that Ex_1, $Ex_2 \neq \emptyset$ and any $y_1 \in Ex_1$, $y_2 \in Ex_2$. Moreover, E is said to be maximal monotone if there does not exist a monotone operator whose graph properly contains the graph of E. Any maximal monotone operator on $\mathbb{R}$ is the subdifferential of a convex function, but this statement does not hold over more general Hilbert spaces.*

We are now ready to state a regularity result for the solution u when Ω is a bounded convex domain is provided by Theorem 3.2.3.1 in [52]. The result is rather general, but we state it in full to clarify when it is applicable. Indeed, we shall encounter variational inequalities in later chapters for which this result does not hold.

Theorem 2.4.5 (Regularity)**.** *With Ω a bounded convex domain and $f \in \mathrm{L}^2(\Omega)$, consider the following problem.*

$$\begin{aligned} -\Delta u + u &= f \ \textit{ in } \Omega, \\ -\nabla u \cdot \boldsymbol{n} &\in \beta(u) \ \textit{ a.e. on } \partial\Omega \end{aligned} \tag{2.23}$$

where $\boldsymbol{n}$ is the unit outward normal to $\partial\Omega$. Suppose that β is a maximal monotone operator on $\mathbb{R}$ with $0 \in \beta(0)$. Then (2.23) has a unique solution $u \in H^2(\Omega)$.

Proof. For a detailed proof of this theorem, the interested reader is referred to chapter 3 of [52]. □

Remark 2.4.6. *Theorem 2.4.5 is valid for problems whose boundary conditions can be represented as in (2.23). We give here further details on theorem 2.4.5, and formulate the Signorini problem in a form suitable for its application. We note that this material is covered in detail in section 3.3.2 in [52]. First, let $j : \mathbb{R} \to (-\infty, \infty]$ be an arbitrary convex function, which we now use to construct a convex functional φ to reformulate problem (2.23).*

$$\varphi(v) = \begin{cases} \frac{1}{2}\int_\Omega |\nabla v|^2 \,\mathrm{d}x + \int_{\partial\Omega} j(v)\,\mathrm{d}S \\ \qquad \textit{if } v \in \mathrm{H}^1(\Omega) \textit{ and } j(v) \in \mathrm{L}^1(\Omega) \\ \infty \ \textit{otherwise.} \end{cases} \tag{2.24}$$

It is then shown that solving problem (2.23) is equivalent to minimising $v \mapsto \varphi(v) + \frac{1}{2}\int_\Omega |v|^2 \,\mathrm{d}x - \int_\Omega fv\,\mathrm{d}x$, where j is a convex function such that $\beta = \partial j$, the subdifferential of j. The existence of such a function is guaranteed for maximal monotone operators on the real line (see section 3.2.2 in [52]).

This theorem is sufficiently general to include certain linear and nonlinear boundary conditions. If we set

$$j(x) = \begin{cases} \infty & x \neq 0 \\ 0 & x = 0, \end{cases} \tag{2.25}$$

which gives

$$\varphi(v) = \begin{cases} \frac{1}{2}\int_\Omega |\nabla v|^2 \,\mathrm{d}x & \textit{if } v \in \mathrm{H}^1_0(\Omega) \\ \infty & \textit{otherwise.} \end{cases} \tag{2.26}$$

so that the boundary condition is that $(u, -\nabla u \cdot \boldsymbol{n})$ lies on $\{x = 0\} \subseteq \mathbb{R}^2$, giving precisely a homogeneous Dirichlet boundary condition. One can think of the function j penalising boundary values that deviate from 0.

Finally, to see that problem (2.17)-(2.18) fits into this framework, we instead start from the choice of j, selecting

$$j(x) = \begin{cases} \infty & x < 0 \\ 0 & x \geqslant 0. \end{cases} \tag{2.27}$$

This choice of j in equation (2.24) gives

$$\varphi(v) = \begin{cases} \frac{1}{2}\int_\Omega |\nabla v|^2 \, \mathrm{d}x \\ \quad \textit{if } v \in \mathrm{H}^1(\Omega) \textit{ and } v \geqslant 0 \textit{ a.e. on } \partial\Omega \\ \infty \qquad \textit{otherwise.} \end{cases} \tag{2.28}$$

For $\beta = \partial j$, the condition $-\nabla u \cdot \boldsymbol{n} \in \beta(u)$ a.e. on $\partial\Omega$ is that almost everywhere on $\partial\Omega$, we have $(u, -\nabla u \cdot \boldsymbol{n})$ must be in the graph of β. This gives precisely the boundary conditions (2.18).

2.5 Finite element methods

The key idea of finite element methods is to use an approximation of a function space that contains the solution of the problem at hand. This is as opposed to finite difference methods which approximate the differential operator itself. See [20, 48] for detailed introductions to the fundamental concepts. In this section we describe the fundamental principles of the finite element method. We begin with definitions of meshes and mesh regularity. Finite element spaces on meshes, the afforementioned approximation spaces, are then introduced and their approximation properties are stated. Finally, we state some standard results from the analysis of finite element methods for elliptic problems.

2.5.1 Triangulations

Suppose that $\Omega \subseteq \mathbb{R}^2$ has polygonal boundary, and let $\mathscr{T}$ to be a conforming triangulation of Ω, namely, $\mathscr{T}$ is a finite family of sets such that

1. $K \in \mathscr{T}$ implies K is an open simplex or box,

2. for any $K, J \in \mathscr{T}$ we have that $\overline{K} \cup \overline{J}$ is a full lower-dimensional simplex (i.e., it is either $\emptyset$, a vertex, an edge or the whole of $\overline{K}$ and $\overline{J}$) of both $\overline{K}$ and $\overline{J}$ and

3. $\bigcup_{K\in\mathscr{T}} \overline{K} = \overline{\Omega}$.

Further, we let h_K be the diameter of K, and define $h : \Omega \to \mathbb{R}$ to be the piecewise constant *meshsize function* of $\mathscr{T}$ given by

$$h(\boldsymbol{x}) := \max_{\overline{K}\ni\boldsymbol{x}} h_K. \tag{2.29}$$

We let $\mathscr{E}$ be the skeleton (set of common interfaces) of the triangulation $\mathscr{T}$ and say $e \in \mathscr{E}$ if e is on the interior of Ω and $e \in \partial\Omega$ if e lies on the boundary $\partial\Omega$ and set h_e to be the diameter of e.

The *patch* of an element K is denoted by $\widetilde{K}$, and is defined to be the union of all elements sharing at least one vertex with K. For example, in a uniform mesh consisting of squares, $\widetilde{K}$ can consist of up to 9 elements. We refer to figures 2.1 and 2.2 for illustrations of element patches on triangular and quadrilateral meshes respectively.

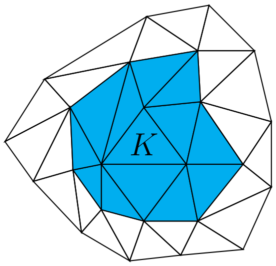


Figure 2.1: The patch of the element K is highlighted in blue on a nonuniform triangular mesh.

From this point on, we will use the terms mesh, partition and triangulation interchangeably. A mesh is said to be quasi-uniform if there exists a constant $C_{qu} > 0$ such that

$$\frac{\max_{K\in\mathscr{T}} h_K}{\min_{K\in\mathscr{T}} h_K} \leqslant C_{qu}. \tag{2.30}$$

We will make the assumption throughout this work that meshes are *shape*

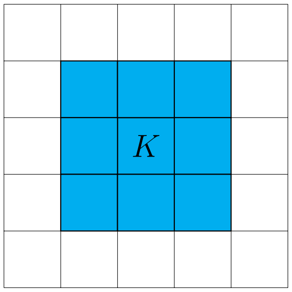


Figure 2.2: The patch of the element K is highlighted in blue on a uniform quadrilateral mesh.

regular. This notion is defined as follows. For triangular meshes, the regularity of an element is defined as

$$\sigma_K := \frac{h_K}{\rho_K}$$

where ρ_K is the diameter of the largest circle that can be inscribed in K (see figure 2.3).

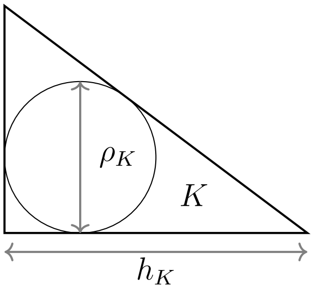


Figure 2.3: Shape regularity parameters of a triangular element. σ_K is minimised when K is an equilateral triangle and can become arbitrarily large if poor refinement decisions are made.

For quadrilateral elements we consider the four triangles that can be formed by choosing three of the four vertices of the square, as shown in figure 2.4. Each triangle T_i has ρ_K^i and h_K^i as defined above. For the quadrilateral element we set

$$\sigma_K = \frac{\max_i h_K^i}{\min_i \rho_K^i}. \tag{2.31}$$

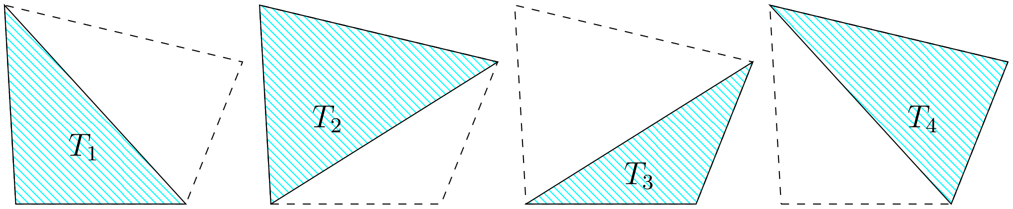


Figure 2.4: For a quadrilateral element, the four triangles shown above are used to define its shape regularity. Note that in this case, large aspect ratio can lead to large σ_K, but also the case where one side is much shorter than the other three.

A family of triangulations $\{\mathscr{T}_i\}_{i \in I}$ is said to be shape regular if there is a constant $\sigma > 0$ such that

$$\sup_{K \in \mathscr{T}_i, i \in I} \sigma_K \leqslant \sigma \tag{2.32}$$

Remark 2.5.1. *Since interpolation estimates and therefore error analysis of finite element methods depend on shape regularity of elements, it is important that as meshes are refined, the shape regularity does not degenerate. This issue will be discussed in §2.7 where the practicalities of mesh refinement are addressed.*

2.5.2 Finite element spaces

Let $P(K)$ be a vector space of polynomials defined on the element K. We will set $P(K) = \mathbb{P}^1(K)$ or $P(K) = \mathbb{Q}^1(K)$, which are the spaces of piecewise linear polynomials over a triangle or quadrilateral respectively, and introduce the *finite element space*

$$\mathcal{V}_h := \{\phi \in C^0(\overline{\Omega}) : \phi|_K \in P(K)\} \tag{2.33}$$

to be the usual space of continuous piecewise affine polynomial functions. We make the following observations about $\mathcal{V}_h$.

Remark 2.5.2. *$\mathcal{V}_h$ is a subspace of* $\mathrm{H}^1(\Omega)$. *We say such a space is a* conforming *finite element space. Functions in $\mathcal{V}_h$ are fully determined by their values at the vertices of the elements $K \in \mathscr{T}$. One can define different finite element spaces by choosing $P(K)$ to be a space of higher order polynomials, $P(K) = \mathbb{P}^k(K)$ or $P(K) = \mathbb{Q}^k(K)$ for triangular and quadrilateral meshes respectively, which in some cases can give improved approximation properties for sufficiently smooth functions. Occasionally in what follows we will refer to k as the* degree *of the finite element space. The approximation order of the space depends on the degree, and is quantified for the spaces $\mathcal{V}_h$ in theorem 2.5.6.*

Finite element functions on a triangulation $\mathscr{T}$ may have discontinuous gradient. We will therefore need to consider jumps of functions that are discontinuous over the edges of a finite element space, and we introduce the following general notation below.

Definition 2.5.3 (Jump operator)**.** *Consider an interface $e \in \mathscr{E}$ bordering elements K_1 and K_2 with outward normal vectors $\boldsymbol{n}_{K_1}$ and $\boldsymbol{n}_{K_2}$ respectively. We define jump operators as $[\![v]\!] = v|_{K_1}\boldsymbol{n}_{K_1} + v|_{K_2}\boldsymbol{n}_{K_2}$ for scalar valued functions and $[\![\boldsymbol{v}]\!] = \boldsymbol{v}|_{K_1} \cdot \boldsymbol{n}_{K_1} + \boldsymbol{v}|_{K_2} \cdot \boldsymbol{n}_{K_2}$ for vector valued functions.*

Proposition 2.5.4 (Trace estimates in L^p, [1, 110].)**.** *Suppose that K is an element of a shape regular partition $\mathscr{T}$ of Ω, and let $v \in \mathrm{W}^{1,p}(K)$ with $p \in (1, \infty)$. Then, there exists a constant C_{tr} depending only on p and K*

such that

$$\|v\|_{\mathrm{L}^p(\partial K)} \leqslant C_{tr}\left(h_K^{-\frac{1}{p}} \|v\|_{\mathrm{L}^p(K)} + h_K^{1-\frac{1}{p}} \|\nabla v\|_{\mathrm{L}^p(K)} \right). \tag{2.34}$$

2.5.3 Interpolation operators

The Lagrange interpolant is perhaps the most natural way to choose a polynomial to approximate a function in a Sobolev space. Assume for now that our Sobolev space is continuously embedded in $C^0(\Omega)$. This means that point values are well-defined and we can construct a Lagrange interpolant of functions v using its nodal values. We will only consider the piecewise linear case here but refer to any major finite element text for details of the higher order case (such as [20, 35, 48]).

Definition 2.5.5 (Lagrange interpolant onto piecewise linear finite element space)**.** *Let V be a sobolev space that is continuously embedded in $C^0(\overline{\Omega})$. Denote the vertices of the mesh $\mathscr{T}$ as $x_1, ..., x_{N_{vert}}$. For any $v \in V$, the lagrange interpolant $\mathcal{I}v$ is the unique element of $\mathcal{V}_h$ such that $v(x_i) = \mathcal{I}v(x_i)$ for all $i = 1, 2, ..., N_{vert}$.*

Function approximation properties of finite element spaces are a key ingredient required in the error analysis of finite element methods. The following is a quasi-optimality result for the Lagrange interpolant.

Theorem 2.5.6 (Quasi-optimal approximation)**.** *Let k be the degree of the finite element space and suppose that $0 \leqslant l \leqslant k$. For $v \in \mathrm{W}^{l+1,p}(\Omega)$, for all elements $K \in \mathscr{T}$, there exists a constant $C_{\mathcal{I}}$ depending only upon the shape regularity of the mesh such that for $m \in \{0, 1, ..., l+1\}$,*

$$|v - \mathcal{I}v|_{\mathrm{W}^{m,p}(K)} \leqslant C_{\mathcal{I}} h_K^{l+1-m} |v|_{\mathrm{W}^{l+1,p}(K)}. \tag{2.35}$$

Remark 2.5.7 (Regularity)**.** *The bound (2.35) tells us that we gain powers of h_K in the approximation order by using higher dimensional polynomials* *(i.e. increasing k). Note that interpolation onto higher order spaces requires higher regularity. This means that the higher approximation order that one expects when using higher order approximations will not be realised if the solution is not sufficiently regular, and therefore lower order finite element approximations are commonly used in such situations.*

Remark 2.5.8 (Lagrange interpolant is not well defined on H^1)**.** *We note that in some cases the weak solution of a PDE will not have point values, and this construction is not applicable. For example, the elliptic problem* (2.41) *with $f \in \mathrm{H}^{-1}$ has a solution which can only be expected to be in H^1. Domains with reentrant corners also result in solutions that lack regularity. In this case an interpolant based upon local averaging is required (see for example [36] or [93]). Analogous approximation properties can be proved.*

Definition 2.5.9 (Clément interpolant)**.** *Here we define the piecewise linear Clément interpolant into the finite element space $\mathcal{V}_h$, although it can be defined more generally [36]. Let $\{x_i\}_{i=1}^N$ denote the nodes of the triangulation $\mathscr{T}$ and let $\phi_i \in \mathcal{V}_h$ be the i-th canonical basis function, with $\phi_i(x_j) = \delta_{ij}$ for $i, j = 1, \dots, N$. Let*

$$\widehat{w}_j := supp(\phi_j). \tag{2.36}$$

Let $u \in \mathrm{L}^1(\Omega)$ and let p_i be the L^2-projection of u onto linear polynomials on $\widehat{w}_j$, that is, p_i is the unique linear polynomial such that

$$\langle u, \Phi \rangle_{\widehat{w}_j} = \langle p_i, \Phi \rangle_{\widehat{w}_j} \quad \forall \Phi \in \mathbb{P}(\widehat{w}_j) \tag{2.37}$$

Then the Clément interpolant of u, denoted $\Pi^C u$ is given by

$$\Pi^C u(x) = \sum_{i=1}^N p_i(x_i)\phi_i(x). \tag{2.38}$$

Theorem 2.5.10 (Approximation properties of the Clément interpolant)**.** *Let k be the degree of the finite element space, $0 \leqslant l \leqslant k$, and let $s \leqslant l + 1$.*

Then for $v \in \mathrm{W}^{l+1,p}(\Omega)$, for all elements $K \in \mathscr{T}$ and all edges $e \subset \partial K$, there exists a constant C_{clem} depending only upon the shape regularity of the mesh such that

$$|v - \mathcal{I}v|_{\mathrm{W}^{s,p}(K)} \leqslant C_{clem} h_K^{l+1-s} |v|_{\mathrm{W}^{l+1,p}(\widetilde{K})}. \tag{2.39}$$

$$|v - \mathcal{I}v|_{\mathrm{W}^{s,p}(e)} \leqslant C_{clem} h_K^{l+1-s-1/p} \left\| v \right\|_{\mathrm{W}^{l+1,p}(\widetilde{K})}. \tag{2.40}$$

Remark 2.5.11. *The Clément interpolant is constructed by local averaging to compensate for lack of regularity. In §3.5-§3.6, we have sufficient regularity to guarantee point values, but we require bilateral bound-preserving properties that standard interpolation operators do not have. To address this, as in the construction of the Clément interpolant, we will work on* node stars $\hat{w}_j$ *and locally adjust the Lagrange interpolant, resulting in a nonlinear interpolant that has the desired bound-preserving properties.*

We are now ready to define the finite element approximation to (2.11). We let $\mathcal{V}_h^0$ be the subspace of $\mathcal{V}_h$ consisting of functions that have zero trace on $\partial\Omega$. Then we seek $U \in \mathcal{V}_h^0$ such that

$$a(U, \Phi) = \langle f, \Phi \rangle \quad \forall \Phi \in \mathcal{V}_h^0. \tag{2.41}$$

A simple but important observation is *Galerkin orthogonality*. Due to the conforming nature of the finite element method, one has immediately from (2.11) and (2.41) that

$$a(u - U, \Phi) = 0 \quad \forall \Phi \in \mathcal{V}_h^0. \tag{2.42}$$

This property leads readily to a quasi-optimality result. Indeed, using Galerkin orthogonality and using the boundedness and coercivity of the bilinear form, we have for any $\Phi \in \mathcal{V}_h^0$,

$$\begin{aligned}\|u-U\|^2_{\mathrm{H}^1(\Omega)} &= a(u-U,u-U)\\ &= a(u-U,u-\Phi)\\ &\leqslant \|u-\Phi\|_{\mathrm{H}^1(\Omega)}\,\|u-U\|_{\mathrm{H}^1(\Omega)}\,.\end{aligned} \tag{2.43}$$

Dividing by $\|u-U\|_{\mathrm{H}^1(\Omega)}$ yields Céa's lemma

$$\|u-U\|_{\mathrm{H}^1(\Omega)} \leqslant \inf_{\Phi\in\mathcal{V}_h^0} \|u-\Phi\|_{\mathrm{H}^1(\Omega)}\,. \tag{2.44}$$

Setting Φ to be the Clément interpolant of u, we can use optimal approximation (2.39) to obtain an a priori error bound:

$$\|u-U\|_{\mathrm{H}^1(\Omega)} \leqslant C_{\mathcal{I}}h|u|_{\mathrm{H}^2(\Omega)}. \tag{2.45}$$

2.5.4 FE for variational inequalities

Now let $\mathcal{K}_h$ be a closed and convex subset of $\mathcal{K}$. For example, in chapters 3 and 4, we will set $\mathcal{K}_h = \mathcal{K}\cap\mathcal{V}_h$. However, it is not strictly required that $\mathcal{K}_h \subseteq \mathcal{K}$.

We can now write the discrete form of problem (2.16): find $U\in\mathcal{K}_h$ such that

$$a(U,\Phi-U) \geqslant \langle f,\Phi-U\rangle \quad \forall\Phi\in\mathcal{K}_h. \tag{2.46}$$

We remark that unlike the finite element approximation (2.41) to a boundary value problem, equation (2.46) does not reduce to a linear algebraic system. Due to the nonlinear nature of the problem, the finite element solution is found through an iterative procedure. One can apply quadratic programming to the discrete equations as is done in [34], projection methods as used in [83] or adjust boundary conditions within the nonlinear iteration procedure as in [94] (see also [38] for more details on this approach). Implementation will be discussed in more detail in the following two chapters.

Error estimates in $\mathrm{H}^1(\Omega)$ for obstacle problems of various types have been standard for some time (see [72, 73, 51, 24] for some early works). The optimal rate of convergence for finite element approximations does indeed hold as in the boundary value case. We summarise this result in the following theorem.

Theorem 2.5.12 (Error estimate for variational inequality, [51])**.** *Let u be the solution of (2.16), and U the finite element approximation, satisfying (2.46). Then there exists a constant C depending upon the regularity of the triangulation and the data f such that*

$$\|u - U\|_{\mathrm{H}^1(\Omega)} \leqslant C\left(\|u - V\|^2_{\mathrm{H}^1(\Omega)} + \|f - Au\|_{\mathrm{L}^2(\Omega)}\left[\,\|U - v\|_{\mathrm{L}^2(\Omega)} + \|u - V\|_{\mathrm{L}^2(\Omega)}\,\right]\right) \tag{2.47}$$

for all $v \in K$ and for all $V \in k_h$, where Au is defined by $\langle Au, v\rangle = a(u, v)$ for all $v \in \mathrm{H}^1(\Omega)$. Consequently, after introducing an interpolant and bounding these terms:

$$\|u - U\|_{\mathrm{H}^1(\Omega)} \leqslant Ch \tag{2.48}$$

2.6 A posteriori estimates via duality methods

In this section, we introduce duality techniques for a posteriori error estimation of finite element approximation of boundary value problems. The use of an appropriately defined dual problem allows error estimates to be derived in L^p-norms rather than the usual energy norms. We present a rigorous and fully computable upper bound on the approximation error.

The material in this section is known, but serves as a useful illustration of the Aubin-Nitsche duality argument, a modification of which will be an integral part of chapter 3.

Let $p \geqslant 2$. We recall that the *Hölder conjugate*, q, of p is the unique q such that

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Definition 2.6.1 (Dual problem)**.** *We now define the dual problem to* (2.11)*. For $p \geqslant 2$, find $z \in \mathrm{H}^1_0(\Omega)$ such that*

$$a(\varphi, z) = \left\langle |u - U|^{p-2}(u - U), \varphi \right\rangle \quad \forall \varphi \in \mathrm{H}^1_0(\Omega). \tag{2.49}$$

Theorem 2.6.2. *Let u solve* (2.11) *and let U be its finite element approximation. We assume either that $2 \leqslant p < \infty$ for $N = 2$ or $2 \leqslant p \leqslant 6$ for $N = 3$. Then there exists $C > 0$ depending on the Lebesgue exponent p and the shape regularity of the mesh such that*

$$\|u - U\|^p_{\mathrm{L}^p(\Omega)} \leqslant C \sum_{K \in \mathscr{T}} \left(\eta^p_K + \frac{1}{2}\eta^p_J \right), \tag{2.50}$$

where

$$\begin{aligned} \eta_K &= h^2_K \, \|f + \Delta U - U\|_{\mathrm{L}^p(K)} \\ \eta_J &= h^{2-1/q}_K \, \|[\![\nabla U]\!]\|_{\mathrm{L}^p(\partial K)} \,. \end{aligned} \tag{2.51}$$

Remark 2.6.3 (Restriction on p.)**.** *The key part of the Aubin-Nitsche duality argument is the use of dual stability to write the error representation in terms of the primal error in L^p. We require $z \in \mathrm{W}^{2,q}(\Omega)$, which is guaranteed by elliptic regularity (see proposition 2.3.4) if $|u - U|^{p-2}(u - U) \in \mathrm{L}^q(\Omega)$, noting the restrictions on p if N=3. Equivalently, we require $u - U \in \mathrm{L}^{q(p-1)}(\Omega) = \mathrm{L}^p(\Omega)$.*

Now, we have $u - U \in \mathrm{H}^1(\Omega)$, and we can use sobolev embeddings to determine values of p sufficient to meet the above requirement. In two spatial dimensions, $H^1(\Omega) \subset \mathrm{L}^p(\Omega)$ for all finite $p \geqslant 2$, and therefore $u - U$ satisties the above requirement. In higher dimensions the condition is more restrictive, for example when $N = 3$ we have $H^1(\Omega) \subset \mathrm{L}^p(\Omega)$ if $2 \leqslant p \leqslant 6$.

Proof. We first use the dual problem to set up an L^p error representation. We may take $\varphi = u - U$ in (2.49) to obtain

$$\|u - U\|^p_{\mathrm{L}^p(\Omega)} = a(u - U, z). \tag{2.52}$$

Using the Galerkin orthogonality property of the primal problem, we subtract the Lagrange interpolant of z:

$$\begin{aligned} a(u - U, z) &= a(u - U, z - \mathcal{I}z) \\ &= a(u, z - \mathcal{I}z) - a(U, z - \mathcal{I}z). \end{aligned} \tag{2.53}$$

We now use the weak form (2.9) on the first term, and integrate the second term by parts element-wise. This yields

$$\begin{aligned} \|u - U\|^p_{\mathrm{L}^p(\Omega)} &= \sum_{K \in \mathscr{T}} \Bigg(\int_K (f + \Delta U - U)\,(z - \mathcal{I}z)\,\mathrm{d}x \\ &\qquad - \int_{\partial K} \nabla U \cdot \boldsymbol{n}(z - \mathcal{I}z)\,\mathrm{d}S \Bigg) \\ &:= R_1 + R_2. \end{aligned} \tag{2.54}$$

We now proceed to bound the terms individually. We begin by applying Hölder's inequality to R_1:

$$R_1 \leqslant \sum_{K \in \mathscr{T}} \left\| h^2 (f + \Delta U - U) \right\|_{\mathrm{L}^p(K)} \left\| h^{-2}(z - \mathcal{I}z) \right\|_{\mathrm{L}^q(K)} \tag{2.55}$$

Applying proposition 2.5.6 on approximation properties of the Lagrange interpolant, we see that

$$|R_1| \leqslant C_{\mathcal{I}} \sum_{K \in \mathscr{T}} \left\| h^2 (f + \Delta U - U) \right\|_{\mathrm{L}^p(K)} |z|_{\mathrm{W}^{2,q}(K)} \tag{2.56}$$

Similarly for R_2,

$$|R_2| \leqslant \frac{1}{2} \sum_{K \in \mathscr{T}} \sum_{e \subseteq \partial K} \left\| h^{2-1/q} [\![\nabla U]\!] \right\|_{\mathrm{L}^p(e)} \left\| h^{-2+1/q} (z - \mathcal{I} z) \right\|_{\mathrm{L}^q(e)} . \tag{2.57}$$

Then using the interpolation estimate in proposition 2.5.6, we obtain

$$|R_2| \leqslant \frac{1}{2} C_{\mathcal{I}} \sum_{K \in \mathscr{T}} \eta_J \left\| z \right\|_{\mathrm{W}^{2,q}(\widetilde{K})} . \tag{2.58}$$

Combining with (2.56) we have

$$\left\| u - U \right\|_{\mathrm{L}^p(\Omega)}^p \leqslant C_{\mathcal{I}} \sum_{K \in \mathscr{T}} \left(\eta_K + \eta_J \right) \left\| z \right\|_{\mathrm{W}^{2,q}(\widetilde{K})} . \tag{2.59}$$

We now apply a discrete Hölder inequality to bound the above further:

$$\begin{aligned} \sum_{K \in \mathscr{T}} \left(\eta_K + \eta_J \right) \left\| z \right\|_{\mathrm{W}^{2,q}(\widetilde{K})} &\leqslant 2^{(p-1)/p} \left(\sum_{K \in \mathscr{T}} \left(\eta_K^p + \eta_J^p \right) \right)^{1/p} \left(\sum_{K \in \mathscr{T}} \left\| z \right\|_{\mathrm{W}^{2,q}(\widetilde{K})}^q \right)^{1/q} \\ &\leqslant 2^{(p-1)/p} C_{\mathscr{T}} \left(\sum_{K \in \mathscr{T}} \left(\eta_K^p + \eta_J^p \right) \right)^{1/p} \left\| z \right\|_{\mathrm{W}^{2,q}(\Omega)} , \end{aligned} \tag{2.60}$$

Where $C_{\mathscr{T}}$ is a constant to quantify the overlap between element patches. By elliptic regularity, we have

$$\left\| z \right\|_{\mathrm{W}^{2,q}(\Omega)} \leqslant C_{\mathrm{reg}} \left\| (u - U)^{p-1} \right\|_{\mathrm{L}^q(\Omega)} . \tag{2.61}$$

We can also exploit the relationship between p and q as follows

$$
\begin{aligned}
\left\|(u-U)^{p-1}\right\|_{\mathrm{L}^q(\Omega)} &= \left(\int_\Omega |u-U|^{(p-1)q}\,\mathrm{d}x\right)^{1/q} \\
&= \left(\int_\Omega |u-U|^{p}\,\mathrm{d}x\right)^{1/q} \\
&= \|u-U\|_{\mathrm{L}^p(\Omega)}^{p/q} \\
&= \|u-U\|_{\mathrm{L}^p(\Omega)}^{p-1}\,,
\end{aligned}
\tag{2.62}
$$

so that we finally have

$$
\|(u-U)\|_{\mathrm{L}^p(\Omega)}^{p} \leqslant 2^{(p-1)/p} C_{\mathcal{I}} C_{\mathscr{T}} C_{\mathrm{reg}} \left(\eta_K^p + \eta_J^p\right)^{1/p} \|(u-U)\|_{\mathrm{L}^p(\Omega)}^{p-1}\,, \tag{2.63}
$$

from which the result follows immediately. □

Remark 2.6.4 ($p=\infty$)**.** *We note that the constant in the bound of theorem 2.6.2 depends on p and blows up as $p \to \infty$ due to the factor of 2^{p-1}. Pointwise bounds on the error are available however, but require a different method of proof. In this case, the construction of an error representation requires Green's functions, which are used to express point values as integrals. Estimates and inequalities involving norms of the Green's functions are then used to bound the result. Since this is not relevant for the work in later chapters, we refer the interested reader to [41, 45, 80] for error estimates in* $\mathrm{L}^\infty(\Omega)$ *for elliptic PDEs and [82] for variational inequalities.*

2.6.1 Lower bound

To complete the analysis, we now show an a posteriori lower bound. This result takes a slightly different form to theorem 2.6.2 in that it bounds the *local* error indicator by the error localised to a patch. This is in contrast to theorem 2.6.2 which is an estimate of the global error, but gives no guarantee that the local error is bounded from above by the local indicator. Together, the two results show equivalence of error and estimate, which indicates that the error estimate is both reliable (i.e. provides an upper bound for the error) and optimal in the sense that the lower bound precludes heavy overestimation. The following theorem combines ideas from the presentation of chapter 10 in [48] and section 2.3 of [2].

Theorem 2.6.5 (A posteriori lower bound)**.** *Let u solve (2.11) and let U be its finite element approximation. Let η_K and η_J be as defined in theorem 2.6.2. Then for all elements K,*

$$C\left(\eta_K + \eta_J\right) \leqslant h^2 \left\|e\right\|_{\mathrm{L}^p(\widetilde{K})} + h \left\|e\right\|_{\mathrm{W}^{1,p}(\widetilde{K})} + h^2 \left\|f - f_h\right\|_{\mathrm{L}^p(\widetilde{K})}. \tag{2.64}$$

To prove theorem 2.6.5, we will need *bubble functions*. These functions are smooth, locally defined on an element or pair of elements, are zero outside the element (or pair) and preserve norms in a sense that will be made precise below. Since our computations are done using quadrilateral meshes, we define bubble functions on the reference quadrilateral, which are then mapped to the physical elements in the standard way.

Definition 2.6.6 (Bubble function)**.** *Let the reference element, $\widehat{K}$ be the unit square with corners at $(0,0)$ and $(1,1)$, and let $\widehat{x}, \widehat{y}$ denote the reference coordinates. Then the element bubble function is given by*

$$\widehat{b}_K := \frac{1}{2}(1 - \widehat{x}^2)(1 - \widehat{y}^2). \tag{2.65}$$

There are four edge bubble functions. For brevity, we just give one, and the others follow by symmetry. If e is the edge along the $\widehat{x}$-axis, then

$$\widehat{b}_e := \frac{1}{2}(1 - \widehat{x}^2)(1 + \widehat{y}) \tag{2.66}$$

Then we have $0 \leqslant \widehat{b}_K, \widehat{b}_e \leqslant 1$. We note also that when the reference cell is mapped to a mesh cell, b_K is supported on K, and b_e is supported on the *pair of elements K^+, K^-.*

Some useful properties follow from equivalence of norms on the reference element. We note here that analogous results hold for similarly defined bubble function on triangular meshes, see for example section 10.1 of [48].

Proposition 2.6.7 (properties of bubble functions, proposition 3.37 in [110])**.** *For all elements $K \in \mathscr{T}$ and for all edges $e \in K$, for all polynomials v and w defined respectively on K and e, there exist constants C^1_{Inv}, C^2_{Inv}, C_{inv}, $C^4_{Inv} > 0$ depending on the shape regularity of $\mathscr{T}$, the polynomial degree of the finite element method and the Lebesgue exponent p such that*

$$\|v\|_{\mathrm{L}^p(K)} \leqslant C^1_{Inv} \left\| b_K^{\frac{1}{p}} v \right\|_{\mathrm{L}^p(K)} \tag{2.67}$$

$$\left\| b_K^{1/p} v \right\|_{\mathrm{L}^p(K)} \leqslant C^2_{Inv} h_K^{\frac{1}{p}} \|v\|_{\mathrm{L}^p(e)} \tag{2.68}$$

$$\|\nabla(b_K v)\|_{\mathrm{L}^p(K)} \leqslant C^3_{Inv}\, h_K^{-1} \|v\|_{\mathrm{L}^p(K)} \tag{2.69}$$

$$\|\nabla(b_K v)\|_{\mathrm{L}^p(K)} \leqslant C^4_{Inv}\, h_K^{\frac{1}{p}-1} \|v\|_{\mathrm{L}^p(e)} \,. \tag{2.70}$$

In what follows, we shall denote all constants above by a generic C_{Inv}.

Proof of theorem 2.6.5. We write $r := f + \Delta U - U$ for the element residual, and for any f_h (which in practice will be an approximation of f), we write $r_h := f_h + \Delta U - U$. We begin by noting that by the triangle inequality,

$$\|r\|_{\mathrm{L}^p(K)} \leqslant \|r_h\|_{\mathrm{L}^p(K)} + \|f - f_h\|_{\mathrm{L}^p(K)} \,. \tag{2.71}$$

Now,

$$\|r_h\|_{\mathrm{L}^p(K)} \leqslant C^1_{\mathrm{Inv}} \left\| b_K^{1/p} r_h \right\|_{\mathrm{L}^p(K)} \tag{2.72}$$

where b_K is the element bubble function defined above. Now

$$\begin{aligned}\left\|b_K^{1/p} r_h\right\|_{\mathrm{L}^p(K)}^p &= \int_K b_K |r_h|^p \,\mathrm{d}x \\ &= \int_K r b_K |r_h|^{p-2} r_h \,\mathrm{d}x + \int_K (f - f_h) b_K |r_h|^{p-2} r_h \,\mathrm{d}x.\end{aligned} \tag{2.73}$$

Dealing with the terms separately, we see after applying Holder's inequality that

$$\begin{aligned}\int_K (f - f_h) b_K |r_h|^{p-2} r_h \,\mathrm{d}x &\leqslant \|b_K\|_{\mathrm{L}^\infty} \|f - f_h\|_{\mathrm{L}^p} \left\||r_h|^{p-1}\right\|_{\mathrm{L}^q} \\ &= \|f - f_h\|_{\mathrm{L}^p} \|r_h\|_{\mathrm{L}^p}^{p-1}.\end{aligned} \tag{2.74}$$

We also have from the weak form of the problem and integration by parts, noting that due to the bubble function vanishing on the cell boundary, edge terms are zero,

$$\begin{aligned}\int_K r b_K |r_h|^{p-2} r_h \,\mathrm{d}x = &\int_K (\nabla u - \nabla U) \cdot \nabla (b_K |r_h|^{p-2} r_h) \,\mathrm{d}x \\ &+ \int_K (u - U) b_K |r_h|^{p-2} r_h \,\mathrm{d}x.\end{aligned} \tag{2.75}$$

By the same steps as above,

$$\int_K (u - U) b_K |r_h r_h|^{p-2} r_h \,\mathrm{d}x \leqslant \|u - U\|_{L^p} \|r_h\|_{\mathrm{L}^p}^{p-1}. \tag{2.76}$$

We now use the inverse inequality (2.69) and Holder's inequality in the same way as above to obtain

$$\int_K (\nabla u - \nabla U)\cdot\nabla(b_K |r_h|^{p-2} r_h)\,\mathrm{d}x \\ \leqslant C_{\mathrm{Inv}}^2 h_K^{-1} |u - U|_{\mathrm{W}^{1,p}} \left\| r_h \right\|_{\mathrm{L}^p}^{p-1}. \tag{2.77}$$

We finally arrive at

$$C \left\| r_h \right\|_{\mathrm{L}^p(K)} \leqslant h_K^{-1} |u - U|_{\mathrm{W}^{1,p}} + \left\| u - U \right\|_{\mathrm{L}^p} + \left\| f - f_h \right\|_{\mathrm{L}^p}. \tag{2.78}$$

We therefore have the bound

$$C\eta_K \leqslant h_K |u - U|_{\mathrm{W}^{1,p}} + h_K^2 \left(\left\| u - U \right\|_{\mathrm{L}^p} + \left\| f - f_h \right\|_{\mathrm{L}^p} \right) \tag{2.79}$$

We now proceed to bound the edge residual. We introduce the function $R := -\left[\!\left[\nabla u \right]\!\right]$, defined on $\mathscr{E}$, the set of interior edges of the triangulation. We begin with the residual equation, i.e. for any v we have

$$a(e, v) = \sum_{K \in \mathscr{T}} \int_K rv + \int_{\partial K} Rv\,\mathrm{d}x. \tag{2.80}$$

Letting $e \in \mathscr{E}$, the choice $v = b_e |R|^{p-1}$, where v is understood to be zero outside the support of b_e, localises the residuals to $\tilde{e}$, and gives

$$a(e, b_e |R|^{p-1}) = \int_{\tilde{e}} b_e r |R|^{p-1}\,\mathrm{d}x + \int_e b_e |R|^p\,\mathrm{d}S. \tag{2.81}$$

It follows that

$$\int_e b_e |R|^p\,\mathrm{d}S = a(e, b_e |R|^{p-1}) - \int_{\tilde{e}} b_e r |R|^{p-1}\,\mathrm{d}x. \tag{2.82}$$

Bounding the terms separately, we begin with

$$\begin{aligned}\int_{\tilde{e}} b_e r |R|^{p-1}\,\mathrm{d}S &\leqslant \|r\|_{\mathrm{L}^p(\tilde{e})} \left\|b_e |R|^{p-1}\right\|_{\mathrm{L}^q(\tilde{e})} \\ &= \|r\|_{\mathrm{L}^p(\tilde{e})} C_{\mathrm{inv}} h^{\frac{1}{q}} \|R\|^{p-1}_{\mathrm{L}^p(\tilde{e})}\,.\end{aligned} \tag{2.83}$$

Now, we have

$$\begin{aligned}a(e, b_e|R|^{p-1}) &\leqslant \|e\|_{\mathrm{W}^{1,p}(\tilde{e})} \left\|b_e|R|^{p-1}\right\|_{\mathrm{W}^{1,q}(\tilde{e})} \\ &\leqslant \|e\|_{\mathrm{W}^{1,p}(\tilde{e})} C_{\mathrm{inv}} h^{\frac{1}{q}-1} \left\||R|^{p-1}\right\|_{\mathrm{L}^q(e)} \\ &= \|e\|_{\mathrm{W}^{1,p}(\tilde{e})} C_{\mathrm{inv}} h^{\frac{1}{q}-1} \|R\|^{p-1}_{\mathrm{L}^p(e)}\,.\end{aligned} \tag{2.84}$$

It finally follows that

$$\|R\|_{\mathrm{L}^p} \leqslant C\Big(h^{\frac{1}{q}} \|r\|_{\mathrm{L}^p(\tilde{e})} + h^{\frac{1}{q}-1} \|e\|_{\mathrm{W}^{1,p}(\tilde{e})} \Big)\,. \tag{2.85}$$

Combining with (2.79) completes the proof. □

2.7 Adaptive mesh refinement

In this section we detail the practical aspects of adaptive algorithms used to optimise a computational mesh for the solution of elliptic problems by the finite element method. We discuss the use of local error indicators to select which parts of the mesh that we wish to refine, how refinement of elements is conducted in practice, and how to ensure that mesh regularity does not degenerate in the process.

We begin with an outline of a general adaptive algorithm. An initial mesh $\mathscr{T}^0$ satisfying the conditions outlined in §2.5.1 is generated over the computational domain. During the solution process, $\mathscr{T}^{l+1}$ is obtained from $\mathscr{T}^l$ by adapting the mesh so that the local mesh size is smaller around cells marked for refinement and larger around cells marked for coarsening. This is generally achieved by adding or removing degrees of freedom and forming new mesh elements. The problem is then solved again on the new mesh, and the process is repeated until some stopping criterion is met. The high-level structure of the algorithm by which $\mathscr{T}^{l+1}$ is obtained from $\mathscr{T}^{l}$ is summarised as: SOLVE→ESTIMATE→MARK→REFINE (see e.g. [92] for a detailed description of the design of adaptive algorithms):

1. SOLVE the discretisation on the current mesh;
2. Calculate the local error ESTIMATE η_k;
3. Use η_k to MARK a subset of cells that we wish to refine or coarsen based on the size of the local indicator;
4. REFINE the mesh.

For residual-type error estimates such as the one given in Theorem 2.6.2, the ESTIMATE step is straightforward once the discrete solution has been found, while for other estimates that are not so explicit, more sophisticated techniques may be required (see §4.6.3 where the error estimate must be approximately computed). In the MARK step, the algorithm must choose which cells are to be refined. This is the most crucial step as it interprets the information given by the error estimate. Marking strategies have parameters that can be tuned by the user to suit the problem at hand. Here we summarise two of the most common marking strategies.

Maximum marking strategy

Let $0 < \beta \leqslant 1$. The maximum marking strategy marks all elements K such that

$$\eta_K \geqslant \beta \max_{K' \in \mathscr{T}} \eta_{K'} \tag{2.86}$$

The extremal case $\beta = 1$ refines only those elements whose local error indicator is maximal, while for small β, almost all elements will be refined. The aim of this rather simple algorithm is to approximately equidistribute error across all elements in the mesh, and was suggested in [26], where it is proven that, in a one-dimensional elliptic case, the approximation error tends to zero upon refinement using the marking strategy above. An analogous marking strategy for coarsening can be defined. For $\gamma \geqslant 1$ mark element K for coarsening if

$$\eta_K \leqslant \gamma \min_{K' \in \mathscr{T}} \eta_{K'}. \tag{2.87}$$

Dörfler marking strategy

The Dörfler marking strategy was used in [44] to guarantee error reduction in adaptive approximation of the solution to the Poisson problem, and was crucial in proving convergence of the adaptive algorithm. Choose $\beta, \gamma \in (0, 1)$, representing the proportion of the error to which refined and coarsened cells contribute respectively. If the estimate for the error is given by $\eta = \sum_{K \in \mathscr{T}} \eta_K$, we mark for refinement all elements $K \in \mathcal{M}$, where $\mathcal{M}$ is a minimal collection of elements such that

$$\sum_{K \in \mathcal{M}} \eta_K \geqslant \beta\eta. \tag{2.88}$$

Analogously, we mark for coarsening a maximal collection $\mathcal{M}'$

$$\sum_{K \in \mathcal{M}'} \eta_K \leqslant \gamma\eta. \tag{2.89}$$

Implementation of mesh refinement

In the numerical work that follows, we use quadrilateral meshes since it allows for efficient refinement as detailed below. If an element is marked for refinement it is quadrisected by placing new degrees of freedom at the centre of mass of the element, and the midpoints of the element edges, and adding edges by joining the centre to the edge midpoints.

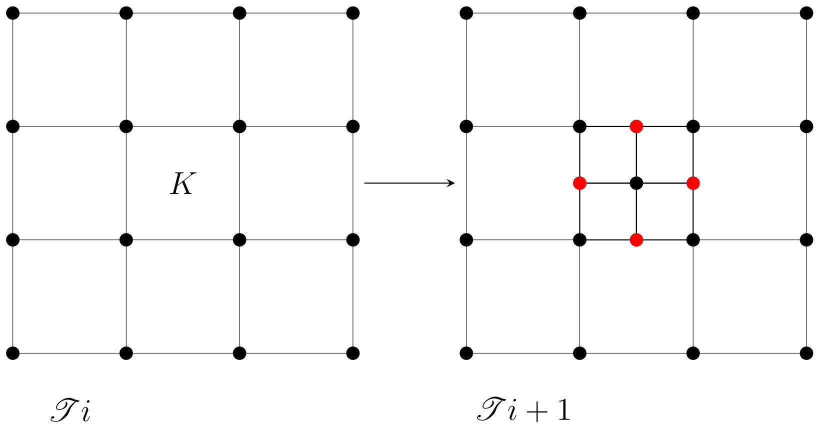


Figure 2.5: Adapting a mesh by refining the central element K. This results in hanging nodes shown in red. The only way to remove hanging nodes entirely by refining additional elements is to refine uniformly.

Thus, none of the existing degrees of freedom need to be moved meaning that the change of mesh is rather efficient. Moreover, we ensure that the shape of the elements will not degenerate as the mesh is refined.

This does, however, result in hanging nodes. It is impractical to refine additional neighbourng cells to avoid hanging nodes, since in the simplest example of such a mesh (uniform squares) all elements would have to be refined at once to avoid hanging nodes.

We therefore allow hanging nodes in our computations, and constrain function values at hanging nodes to ensure that the finite element method is still conforming, that is, discrete functions still lie in $\mathrm{H}^1(\Omega)$. Algebraic constraints on the degrees of freedom associated with hanging nodes are required to ensure that finite element functions remain continuous.

The result is a mesh where neighbouring cells could theoretically differ by any number of grid levels. It is therefore advantageous to allow a small amount of mesh smoothing such as setting a maximum difference of grid levels between adjacent cells to preserve the regularity of the mesh. In our computations, we insist that the difference in refinement level across a cell boundary is allowed to be at most one. In practice, this is achieved by refining some extra cells that were not originally marked to 'smooth' the mesh. For a more detailed explanation of the implementation of mesh refinement, we refer the reader to the extensive `deal.ii` documentation available online [5].

With regards to coarsening, due to the hierarchical structure of the meshes that result from this process, cells that have been refined 'parent', that is, a quadrilateral in $\mathscr{T}^i$ for some $i \leqslant l$ in which it is fully contained. If all four 'children' elements are marked for coarsening, the vertex at the middle of the four elements is removed and the parent cell is restored resulting in a locally coarser mesh. Note that it is not possible to coarsen beyond an initial macro triangulation or *coarse mesh* on which the computation is initiated.

Remark 2.7.1 (Hanging nodes)**.** *We note that a finite element function with hanging nodes is no longer the solution of a problem of the form* (2.41) *due to the constraints of the degrees of freedom that correspond to hanging nodes. It seems that this introduces an inconsistency to the problem. In this work, we will neglect errors that arise from this, but quantification of the effects of hanging nodes would be an interesting subject of further work.*

Interpolation onto new meshes

We observe that if a function U lies in a continuous piecewise polynomial finite element space defined on $\mathscr{T}^l$, and that $\mathscr{T}^{l+1}$ is obtained by refinement only, then U also lies in the finite element space defined on $\mathscr{T}^{l+1}$, as we have introduced degrees of freedom. This is not the case when coarsening occurs, and U needs to be interpolated onto the new mesh. This introduces new errors in the simulation of problems where solutions or data from one mesh is required on another. Typical examples are iterative solvers for nonlinear problems and time-dependent problems. In the former case, when the mesh changes the current iterate of the procedure needs to be transferred to the new mesh and used to calculate the next iterate. In the latter case, the finite element solution at one time-step is needed at following time-steps, again leading to calculations involving functions defined on different meshes. We note that the errors introduced by coarsening can be estimated a posteriori along with other sources of error (see e.g. [67]).

2.8 Conclusions

We have introduced the necessary notation, tools and results to conduct the analysis of later chapters, including foundational results from PDE theory and finite element approximation. We derived in detail an example of an a posteriori error estimate using the Aubin-Nitsche duality argument. For the boundary value problem studied in this section, we have the best-case scenario of rigorous upper and lower bounds on the error. While the upper bound is global, the cell-wise contributions can serve as useful *indicators* or error to drive mesh adaptivity, as will be seen in later sections. Lower bounds are an important measure of sharpness but are not always available. We concluded the chapter with a description of mesh adaptivity, including algorithmic aspects and practical concerns.

Chapter 3

Duality based error control for the Signorini problem

3.1 Abstract

In this chapter we study the a posteriori bounds for a conforming piecewise linear finite element approximations of the Signorini problem. We prove new rigorous a posteriori estimates of residual type in the L^4 norm, fully computable up to constants which are independent of discretisation parameters. This new analysis treats the positive and negative parts of the discretisation error seperately, requiring a novel sign- and bound-preserving interpolant, which is shown to have optimal approximation properties. The estimates rely on the sharp dual stability results on the problem in $\mathrm{W}^{2,(4-\varepsilon)/3}$ in a two-dimensional domain. We summarise extensive numerical experiments aimed at testing the robustness of the estimator to validate the theory.

3.2 Introduction & literature review

Let $\Omega \subset \mathbb{R}^2$ be open and convex with polygonal boundary $\partial\Omega$. Let $\boldsymbol{n}$ denote the outward pointing normal to $\partial\Omega$. We consider a variational inequality derived from the PDE (2.17), re-stated here for convenience.

$$-\Delta u + u = f \text{ in } \Omega, \tag{3.1}$$

coupled with boundary conditions

$$u \geqslant 0, \quad \nabla u \cdot \boldsymbol{n} \geqslant 0, \quad u \nabla u \cdot \boldsymbol{n} = 0 \text{ on } \partial\Omega. \tag{3.2}$$

Note that the structure of the conditions implies that if $u > 0$ on $I \subset \partial\Omega$ then $\nabla u \cdot \boldsymbol{n} = 0$ over I and vice versa.

This problem and its variants has a wide range of application including elasto-plasticity (the initial motivation for the study of the Signorini problem), fluid flow in porous media [23] and finance [116]. After its formulation as a variational inequality, the problem was comprehensively analysed in [68] where existence and uniqueness of solutions for a wide class of such problems was established.

A priori error estimates for finite element approximations to variational inequalities have been obtained of optimal order in energy norms [24, 51, 73, 91], and in $L^\infty(\Omega)$ [77]. Results in other norms are far less common; [72] gives an a priori bound in $\mathrm{L}^2(\Omega)$ but only in one spatial dimension. Other results include bounds on trace norms measuring error along the Signorini boundary [100]. More recently bounds in several low-order global norms were proved in [34].

The theory of a posteriori error estimation for elliptic variational inequalities is significantly less well-developed than that for elliptic boundary value problems. Nonetheless, there have been several works on the Signorini problem as well as the related problem where the obstacle is interior to the domain rather than on the boundary (usually referred to as the obstacle problem). In [58], hierarchical estimates are used under the assumption that quadratic finite elements give better approximation than linears. Error estimates for variational inequalities with nontrivial boundary data are given in [66], estimates for an alternative form using Lagrange multipliers are given in [18, 114], with results for the discontinuous Galerkin method given in [112].

Rigorous a posteriori error bounds of residual type were derived for the variational inequality of the first kind for the Signorini problem in [32] for piecewise linear finite elements. Here, the authors give an upper bound in $\mathrm{H}^1(\Omega)$. Several other works have considered a posteriori control in energy norms, see [57]. Pointwise error control was established in [82]. Recent regularity results for the problem in which constraints only apply at the boundary [34] show that better rates can be achieved in this case, paving the way for a posteriori error estimates in lower order norms.

We remark that error estimates have been derived for this problem using the dual-weighted residual methodology (see for example [103]) to estimate the error in a target quantity. In particular, estimates follow for the $\mathrm{L}^2(\Omega)$-error, however the necessary stability of the dual problem enters as an assumption.

A priori and a posteriori error estimates in low order norms are established using duality arguments, and the resulting error bounds in for example $\mathrm{L}^2(\Omega)$ are standard in the literature for boundary value problems . The Aubin-Nitsche duality argument allows a priori and a posteriori bounds of higher order than in energy norms or $\mathrm{L}^\infty(\Omega)$, see [2] and the references therein. See also section 2.6 for a derivation of a posteriori bounds in $\mathrm{L}^p(\Omega)$ for an elliptic problem. Application of this methodology has proved to be a challenge in the case of variational inequalities since the dual problem can lack the necessary regularity. Indeed, even given smooth data, solutions of the dual Signorini problem are smooth away from the boundary, but suffer from singular points at the boundaries of the contact set, that is, the set of points at which the boundary change type.

Using the ideas of an a priori argument recently given in [34], it is possible to use $\mathrm{W}^{2,p}(\Omega)$-regularity of the primal solution for $p > 2$ to compensate for this fact. The approach taken in [72] in which different dual problems are considered for the positive and negative parts of the error is adapted for the a posteriori setting and combined with a two-sided approximation that takes the place of the usual Lagrange interpolant. The novelty in this work is that we are able to prove rigorous a posteriori error control in $\mathrm{L}^4(\Omega)$. The results in [34] suggest that the exponent 4 is a critical value in the sense that piecewise linear finite element approximations achieve orders of convergence in $\mathrm{W}^{1,p}(\Omega)$ of $1-\varepsilon$ for any $\varepsilon > 0$, for $p = 4$, but with increasingly suboptimal order for $p > 4$.

The rest of the chapter is set out as follows: In §3.3 we introduce notation, formalise the model problem and describe the finite element approximation. In §3.4, we summarise recent work on a priori analysis of finite element approximation in non-energy norms. This material helps to motivate the approach taken in later sections. We introduce a new interpolant that incorporates bilateral bounds on its values in §3.5, and detail results on this two-sided approximation in §3.6, to be used in §3.7 where we state and prove the key estimates and main results. These are then tested numerically in §3.8.

3.3 Problem setup

This section contains a summary of the problem at hand, as well as a brief review of the current state of the art for regularity of the variational inequality and its finite element approximation. We note that the article [34] considers the problem on the unit square. For this chapter, to make use of their results, we will do the same, but remark that arguments can be extended to convex polygonal domains.

We define the test and trial space

$$\mathcal{K} = \{v \in \mathrm{H}^1(\Omega) : v \geqslant 0 \text{ on } \partial\Omega\} \subset \mathrm{H}^1(\Omega) \tag{3.3}$$

that is appropriate to define the weak form of (3.1)–(3.2). It reads: Find

$u \in \mathcal{K}$ such that

$$a(u, v-u) := \langle \nabla u, \nabla v - \nabla u \rangle + \langle u, v-u \rangle \geqslant \langle f, v-u \rangle =: l(v-u) \quad \forall v \in \mathcal{K}. \tag{3.4}$$

This leads us to state some well known properties of the problem (3.1)–(3.2) as well as some more recent regularity results. The following result holds on convex domains, and is therefore true for Ω as defined above. It is an application of the general result to our specific case.

Proposition 3.3.1 (Regularity of the primal problem [52], Theorem 3.2.3.1.)**.** *Given $f \in \mathrm{L}^2(\Omega)$ there exists a unique solution $u \in \mathrm{H}^2(\Omega)$ of (3.4). Moreover, there exists a constant $C > 0$ such that*

$$\|u\|_{\mathrm{H}^2(\Omega)} \leqslant C \|f\|_{\mathrm{L}^2(\Omega)}. \tag{3.5}$$

Remark 3.3.2. *There are variants of the Signorini problem for which this regularity result cannot be applied. For example, if the boundary also contains portions where Neumann and/or Dirichlet conditions are applied, then the boundary conditions can not be expressed in the correct form to apply proposition 3.3.1. In this case we are not guaranteed $u \in \mathrm{H}^2(\Omega)$. Such problems do arise in practical applications. Indeed they are studied in [17, 7] as well as in §4 of this thesis.*

Proposition 3.3.3. *[Improved Regularity of the Primal Problem [34].] Suppose that u solves (3.4). Then if $f \in \mathrm{L}^p(\Omega)$ for $2 < p < 4$, then $u \in \mathrm{W}^{2,p}(\Omega)$.*

Remark 3.3.4. *[$p = 4$.] Proposition 3.3.3 says nothing about the regularity of the solution in the case where $p = 4$. This is because $\mathrm{W}^{2,4-\varepsilon}(\Omega)$ is a regularity limit for the solution, no matter how smooth the datum f, so that if $f \in \mathrm{L}^4(\Omega)$, we still have at most $u \in \mathrm{W}^{2,4-\varepsilon}(\Omega)$, as remarked in [34] and [100], with examples to demonstrate.*

Finite element discretisation

We recall that $\mathscr{T}$ is a conforming triangulation of Ω, namely, $\mathscr{T}$ is a finite family of sets such that satisfies the conditions enumerated in §2.5. We will make the assumption that the triangulation is shape-regular.

Recalling notation from §2.5.2, we let $P(K)$ denote the space of piecewise linear polynomials over a triangular or quadrilateral mesh element K, and recall the *finite element space*

$$\mathcal{V}_h := \{\phi \in \mathrm{H}^1(\Omega) : \phi|_K \in P(K)\}, \tag{3.6}$$

the usual space of continuous piecewise affine polynomial functions. We also define a discrete analogue of $\mathcal{K}$ as

$$\mathcal{K}_h := \{v \in \mathcal{V}_h : v \geqslant 0 \text{ on } \partial\Omega\} = \mathcal{K} \cap \mathcal{V}_h \subset \mathrm{H}^1(\Omega). \tag{3.7}$$

We are now in a position to state the finite element approximation of (3.1)–(3.2) which is to find $U \in \mathcal{K}_h$ such that

$$a(U, \Phi - U) \geqslant l(\Phi - U) \quad \forall \Phi \in \mathcal{K}_h, \tag{3.8}$$

Note that we restrict our attention to either piecewise linear finite element spaces on triangular meshes or piecewise bilinear spaces on quadrilateral meshes. This is because we will need to use a one-sided polynomial approximation of the dual solution. In particular, we must ensure that positivity is preserved. There are alternative ways to do this, that is, by design of bound-preserving finite element methods, see [11] for a recent review of such methods, but we shall not consider this here.

Proposition 3.3.5 (The discrete problem is well posed [65, Thm 2.1])**.** *For all $h > 0$ there exists a unique solution U of (3.8).*

Remark 3.3.6. *We remark that proposition 3.3.5 is exactly the same the-*

orem that gives existence and uniqueness of a solution and stability for the continuous problem. There is an analogy with Lax-Milgram, which gives existence and uniqueness for continuous and discrete solutions of the associated boundary value problem.

Definition 3.3.7 (Contact Set)**.** *Let u be the unique solution of problem* (3.4)*. Since we have $u \in \mathrm{H}^2(\Omega)$ from Proposition 3.3.1, u has a continuous representative on $\overline{\Omega}$ which we may use to define the contact set*

$$\mathcal{A} := \{x \in \partial\Omega \mid u(\boldsymbol{x}) = 0\},$$

with $\mathring{\mathcal{A}}$ the relative interior of this set, that is, the interior with respect to the boundary. The discrete contact set is defined analogously:

$$\mathcal{A}_U = \{x \in \partial\Omega \mid U(\boldsymbol{x}) = 0\}.$$

Remark 3.3.8. *The stronger regularity results for the solution of* (3.4) *given in 3.3.3 follow from recently proven results on the structure of $\mathcal{A}$ given in [4], where it is shown that $\mathcal{A}$ consists of finitely many connected components and isolated points, and in particular this set of isolated points does not contain accumulation points. This means that $\mathring{\mathcal{A}}$ is a union of open intervals in $\partial\Omega$.*

3.4 A priori error control

The a posteriori analysis in this work was motivated by a priori error control in $\mathrm{L}^4(\Omega)$ proven in [34]. Here, we outline the main ideas and results of that work. The authors make use of the Ritz projection $R_h : \mathrm{H}^1(\Omega) \to \mathcal{V}_h$ such that equation

$$a(R_h u, \Phi) = a(u, \Phi) \tag{3.9}$$

holds for all $\Phi \in \mathcal{V}_h$ which corresponds to an approximation of an unconstrained problem. Classical estimates are available for the Ritz projection which are modified by the authors for rectangular domains. Using the observation that $R_h(u)$ solves the discrete form (3.8) of an obstacle problem where the obstacle is $R_h(u) - u$ instead of the zero function in (3.2), they show that $R_h(u) - U$ is controlled in H^1 by discrete functions w_h satisfying $R_h(u) - u \leqslant w_h \leqslant R_h(u)$ on the boundary. The construction of such a function provides a priori error control in various norms. In particular, for any $\varepsilon \in (0, 1/2)$,the bounds

$$\begin{aligned} \|u - U\|_{\mathrm{W}^{1,4}(\Omega)} &= \mathcal{O}(h^{1-\varepsilon}) \\ \|u - U\|_{\mathrm{W}^{1,\infty}(\Omega)} &= \mathcal{O}(h^{1/2-\varepsilon}) \\ \|u - U\|_{\mathrm{L}^{\infty}(\Omega)} &= \mathcal{O}(h^{3/2-\varepsilon}). \end{aligned} \tag{3.10}$$

hold as $h \to 0$.

To obtain optimal rates of convergence in lower order norms, a duality argument is used in combination with the above estimates. In the usual duality argument for elliptic PDEs, a test function for the dual problem is chosen to obtain a representation of the primal error (see §2.6). The required function in their case is not admissible, and is modified nodally by adding contributions to ensure that the inequality constraints are satisfied. The extra terms that result are controlled by the estimate in $\mathrm{W}^{1,\infty}(\Omega)$ and absorbed into the other terms. A careful quantification of the regularity of the dual problem allows a bound of optimal order to be derived, which is summarised in the following theorem.

Theorem 3.4.1 (Error control [34, Thm 6.1])**.** *Let u solve (3.4) for $f \in \mathrm{L}^{\infty}(\Omega)$, and let $U \in \mathcal{V}_h$ be the solution of (3.8). Then for all $\varepsilon \in (0, 1/2)$ we have as $h \to 0$,*

$$\left\|(u - U)^{+}\right\|_{\mathrm{L}^{4}(\Omega)} = \mathcal{O}(h^{2-\varepsilon}), \tag{3.11}$$

where the notation $g^{+} := \max(g, 0)$ and $g^{-} := \min(g, 0)$ for a real-valued function g. If we additionally assume that the discrete solution satisfies a *technical but not particularly restrictive assumption on the topology of the discrete contact sets (see condition ($\mathbf{A_h}$) in [34]), then we also have*

$$\left\|(u-U)^{-}\right\|_{\mathrm{L}^4(\Omega)} = \mathcal{O}(h^{2-\varepsilon}), \tag{3.12}$$

and therefore full a priori error control in $\mathrm{L}^4(\Omega)$.

Remark 3.4.2 (Condition ($\mathbf{A_h}$))**.** *Here for completeness we state condition ($\mathbf{A_h}$) as given in [34]. The condition is satisfied if there exist points $d_i \in \partial\Omega$ and numbers δ_i such that the open balls $B_{\delta_i}(d_i) \cap \partial\Omega$ have nonzero distance from each other and the corners of the domain, and such that for sufficiently small $h > 0$,*

1. *The sets $B_{\delta_i}(d_i) \cap \partial\Omega$ cover $\partial\Omega$ and each of these sets contains precisely one element of the relative boundary of $\mathcal{A}_U$.*

2. *Every connected component of $\mathcal{A}_U$ has non-empty interior.*

This condition is required to prove the stability result for the dual problem based on the negative part of the error in [34], and we quote their result in theorem 3.7.10.

3.5 Unilateral approximation

Let $\{x_i\}_{i=1}^N$ denote the nodes of the triangulation $\mathscr{T}$ and let $\phi_i \in \mathcal{V}_h$ be the i-th canonical basis function, with $\phi_i(x_j) = \delta_{ij}$ for $i, j = 1, \ldots, N$. Let

$$\widehat{x}_j := \cup\{K \in \mathscr{T} : \operatorname{supp}(\phi_j) \cap K \neq \emptyset\}. \tag{3.13}$$

We also recall $\mathcal{I}(z)$ is the nodal Lagrange interpolant.

Assume we have a function $0 \leqslant z \in \mathrm{W}^{2,q}(\Omega)$ then the aim of this section is to examine the question of existence of an interpolation operator

$\widehat{\Pi} : \mathrm{W}^{2,q}(\Omega) \to \mathcal{V}_h$ satisfying both

$$\begin{aligned} 0 \leqslant \widehat{\Pi}(z) \leqslant z \text{ and} \\ \left\| z - \widehat{\Pi}(z) \right\|_{\mathrm{L}^q(\Omega)} \leqslant Ch^2 \left\| z \right\|_{\mathrm{W}^{2,q}(\Omega)} . \end{aligned} \tag{3.14}$$

In other words, we wish to show that there exists a two-sided approximation that also has the optimal approximation properties enjoyed by the Lagrange interpolant (see proposition 2.5.6). This form of approximation theory arises in many areas of finite element analysis. In particular, the optimal approximation of non-smooth functions is a non trivial task relying on local averaging, as done by Clément originally and extended to zero boundary values in Scott-Zhang.

3.5.1 From below

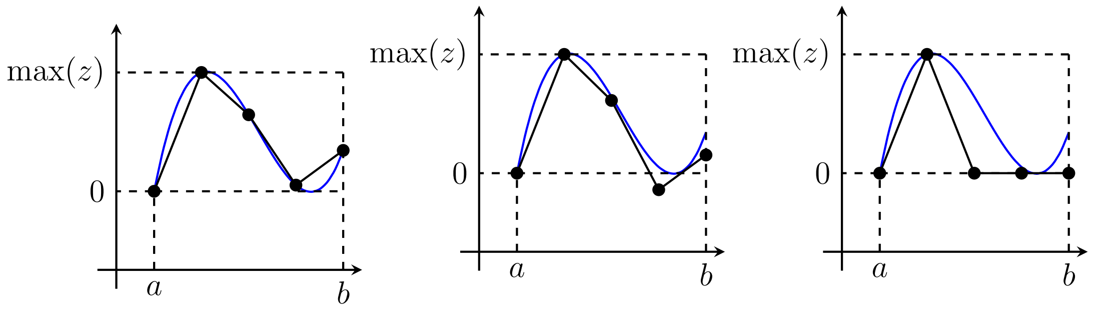
In this section we examine the design of functions that are positivity preserving. This means, for a function $z \geqslant 0$, the approximation also satisfies $\Pi(z) \geqslant 0$. Point values are well defined in $\mathrm{W}^{2,q}(\Omega)$ spaces if $2q > n$ (see remark 2.2.6). Also if $q = 1$ and $n = 2$ we have point values. If either of these cases are true, one can simply consider the Lagrange interpolant, this has the properties

$$\Pi(z) \geqslant 0 \text{ if } z \geqslant 0, \tag{3.15}$$

see Figure 3.1a, as well as optimal approximation

$$\left\| z - \Pi(z) \right\|_{\mathrm{L}^q(\Omega)} \leqslant Ch^2 \left\| z \right\|_{\mathrm{W}^{2,q}(\Omega)} . \tag{3.16}$$

If q and n don't satisfy these conditions and point values do not exist, the approximation is considerably more involved, although has been tackled in [33] where an interpolant is constructed through local mean-values of the function. It was shown in [33] that the constructed interpolant is L^q stable, second order accurate and linear. See also [108, 109] for related ideas using



(a) Approximation with the piecewise linear Lagrange interpolant, $\mathcal{I}(z)$. Notice it is positive.

(b) Approximation with a nonlinear interpolant $\Pi(z)$. Notice it is bounded above by the function, but is no longer bounded below by zero.

(c) Approximation with the modified bilateral approximation $\widehat{\Pi}(z)$. Notice that $0 \leqslant \widehat{\Pi}(z) \leqslant z$.

Figure 3.1: An illustration of three piecewise linear operators, black, applied to the same function, blue, that satisfies $z > 0$. The Lagrange interpolant, $\mathcal{I}(z) \geqslant 0$, the nonlinear interpolant $\Pi(z) \leqslant z$ and the bilateral approximation $0 \leqslant \widehat{\Pi}(z) \leqslant z$.

a nonlinear interpolation operator [43].

3.5.2 From above

In this section we examine the design of interpolants that are bounded above by specific functions. This means, for a function $z \in \mathrm{W}^{2,q}(\Omega)$ with $2q > n$, the interpolant satisfies $\Pi(z) \leqslant z$. In the case we can see the Lagrange interpolant does not satisfy the requirements, see Figure 3.1a. Indeed, any strictly convex function will see the Lagrange interpolant violate this bound. It can, however, be modified. Indeed, consider the function:

$$\Pi(z)(x) = \sum_{i=1}^{N} \left(\mathcal{I}z(x_i) - \max_{y \in \widehat{x}_i} (\mathcal{I}(z)(y) - z(y)) \right) \phi_i(x) =: \mathcal{I}(z)(x) - R(x), \tag{3.17}$$

where we recall from definition 2.5.5 that the points x_i are the vertices of the mesh, that is, the nodes from which point values are taken for interpolation.

Lemma 3.5.1. *For $0 \leqslant z \in \mathrm{W}^{2,q}(\Omega)$ the approximation $\Pi(z)$ defined in (3.17) satisfies*

$$\Pi(z) \leqslant z \text{ in } \Omega. \tag{3.18}$$

Proof. Note that by definition of R we have

$$\Pi(z) = \mathcal{I}(z) - R \leqslant \mathcal{I}(z) - (\mathcal{I}(z) - z) = z. \tag{3.19}$$

□

This is a nonlinear interpolant since for general $u, v \in \mathrm{W}^{2,q}(\Omega)$

$$\Pi(u+v) \neq \Pi(u) + \Pi(v). \tag{3.20}$$

This means we cannot directly apply Bramble-Hilbert to obtain optimal approximation under minimal regularity. Instead we must take a different approach.

Theorem 3.5.2 (Optimal approximation)**.** *Suppose $z \in \mathrm{W}^{2,q}(\Omega)$, where $2q > n$, then the approximation $\Pi(z)$ defined through (3.17) satisfies*

$$\left\| z - \Pi(z) \right\|_{\mathrm{L}^q(\Omega)} \leqslant C h^2 \left| z \right|_{\mathrm{W}^{2,q}(\Omega)}. \tag{3.21}$$

Proof. To begin, we note

$$\left\| z - \Pi(z) \right\|_{\mathrm{L}^q(\Omega)} \leqslant \left\| z - \mathcal{I}(z) \right\|_{\mathrm{L}^q(\Omega)} + \left\| R \right\|_{\mathrm{L}^q(\Omega)}. \tag{3.22}$$

we have from theorem 2.5.6 that

$$\left\| z - \mathcal{I}(z) \right\|_{\mathrm{L}^q(\Omega)} \leqslant C h^2 \left| z \right|_{\mathrm{W}^{2,q}(\Omega)}. \tag{3.23}$$

To control the 2nd term, notice that since $R \in \mathcal{V}_h$, for a $K \in \mathscr{T}$ with vertices, $x_{1,2,3}$ we can write a lumped L^q-norm to see that (cf. [47, proposition 12.5])

$$\begin{aligned} \|R\|^q_{\mathrm{L}^q(K)} &\leqslant Ch^n \sum_{i=1}^3 |R(x_i)|^q \\ &\leqslant Ch^n \sum_{i=1}^3 \|\mathcal{I}(z) - z\|^q_{\mathrm{L}^\infty(\widehat{x}_i)} . \end{aligned} \tag{3.24}$$

Now, note a further property of the Lagrange interpolant (see [20, corollary 4.4.7]) is

$$\|z - \mathcal{I}(z)\|_{\mathrm{L}^\infty(K)} \leqslant Ch^{2-\frac{n}{q}} |z|_{\mathrm{W}^{2,q}(K)} . \tag{3.25}$$

Hence substituting into (3.24) we have

$$\|R\|^q_{\mathrm{L}^q(K)} \leqslant Ch^{2q} |z|^q_{\mathrm{W}^{2,q}(\widehat{K})} . \tag{3.26}$$

Now

$$\|R\|^q_{\mathrm{L}^q(\Omega)} = \sum_{K\in\mathscr{T}} \|R\|^q_{\mathrm{L}^q(K)} \leqslant \sum_{K\in\mathscr{T}} Ch^{2q} |z|^q_{\mathrm{W}^{2,q}(\widehat{K})} \leqslant Ch^2 |z|^q_{\mathrm{W}^{2,q}(\Omega)} . \tag{3.27}$$

completing the proof, where we have used $2q > n$ to infer the final step: assuming that $h \leqslant 1$, $h^{2q} < h^n = h^2$. □

Remark 3.5.3. *Note the restriction on q in Theorem 3.5.2. For the application we have in mind we wish to work in $q = \frac{4}{3}$ which restricts $n = 1, 2$. For $n = 3$ the appproximation would no longer be optimal.*

3.6 Bilateral approximation

There is much less in the literature on the design of off two sided approximations. The only arguments date back to Mosco and Strang [74, 101, 102] where the authors consider piecewise linear approximations for $n = 1, 2, 3$

and $q = 2$. Unfortunately these arguements do not appear to extend to this case.

The difficulty in this problem can bee seen by examining Figure 3.1b around where $z \sim 0$. To keep a bilateral constraint on the approximation, any appropriate function is *squeezed* and the only mechanism is to force it to zero on a surrounding patch.

3.6.1 Optimal approximation over $P(K)$

We modify Π (defined in equation (3.17)) in the following way by requiring at the degrees of freedom:

$$\widehat{\Pi}(z)(x_i) = \begin{cases} \Pi(z)(x_i) \text{ if } \Pi(z) \geqslant 0 \text{ on } \widehat{x}_i \\ 0 \text{ otherwise,} \end{cases} \tag{3.28}$$

as illustrated in Figure 3.1c.

Lemma 3.6.1. *Let* $0 \leqslant z \in \mathrm{W}^{2,q}(\Omega)$, *and define*

$$\mathcal{Z}_h := \{z_h \in \mathcal{V}_h : 0 \leqslant z_h \leqslant z\}. \tag{3.29}$$

Then $\widehat{\Pi}(z) \in \mathcal{Z}_h$. *Further, for* $n < 2q$ *we have*

$$\left\| z - \widehat{\Pi}(z) \right\|_{\mathrm{L}^q(\Omega)} \leqslant C_b h^2 \left| z \right|_{\mathrm{W}^{2,q}(\Omega)} . \tag{3.30}$$

Proof. To begin, notice that due to the regularity of z we can find a constant C independent of h such that

$$\left\| z - \widehat{\Pi}(z) \right\|_{\mathrm{L}^q(\Omega)} \leqslant \left\| z - \Pi(z) \right\|_{\mathrm{L}^q(\Omega)} + C \left\| S \right\|_{\mathrm{L}^q(\Omega)} , \tag{3.31}$$

where

$$S = \max(-\Pi(z), 0) . \tag{3.32}$$

Control of the first term is given in Theorem 3.5.2. For the second, notice that S only has support when $\mathcal{I}(z) > z$ in a vicinity of when z vanishes. This can only happen if z is locally convex. Further, since $\mathcal{I}(z) > 0$ whenever $z > 0$ we have that

$$|\max(-\Pi(z), 0)| \leqslant |R| \tag{3.33}$$

and hence

$$\|S\|_{\mathrm{L}^q(\Omega)} \leqslant \|R\|_{\mathrm{L}^q(\Omega)} \leqslant Ch^2 \, |z|_{\mathrm{W}^{2,q}(\Omega)} \,, \tag{3.34}$$

concluding the proof. □

Remark 3.6.2. *Two sided bounds are conjectured in [74] for higher dimension, at least with $q = 2$, although their results only seem to hold with $n \leqslant 3$. Indeed, it certainly cannot be generalised to $n \geqslant 5$ since there exists a non-trivial function $z \geqslant 0$ vanishing on a dense set (according to [73], though the authors do not provide it) ensuring the only bilateral approximation in $\mathcal{Z}_h$ is $z_h \equiv 0$.*

The conjecture we make is that optimal bilateral approximations in $\mathrm{L}^q(\Omega)$ are only possible when $2 - \frac{n}{q} > 0$, hence we believe our operator is a constructive example of that studied in [74], i.e., is valid for $n = 3$ and $q = 2$.

3.7 A posteriori error control

The dual problem used in [34] worked with the modification because a priori bounds in $\mathrm{W}^{1,\infty}(\Omega)$ were available to replace norms in this space by powers of the mesh size, giving optimal rates in the a priori case. However, the bound is not computable, and is therefore not available in the a posteriori framework. We must choose the dual problem so that we are able to select a test function that will lead to a bound that is fully computable up to constants. Motivated by the approach used in [72] to derive a priori $\mathrm{L}^2(\Omega)$ estimates, we therefore define the following space

$$\mathcal{M} = \{v \in \mathrm{H}^1(\Omega) \mid v \leqslant 0 \text{ on } \mathring{\mathcal{A}}\}, \tag{3.35}$$

and consider the problem of finding $z \in \mathcal{M}$ such that

$$a(z, v - z) \geqslant \left\langle \max(0, u - U)^3, v - z \right\rangle \quad \forall v \in \mathcal{M}. \tag{3.36}$$

Remark 3.7.1. *Note the similarity with the dual elliptic boundary value problem* (2.49)*, and observe that $z + u - U$ lies in the space $\mathcal{M}$, since by the definition of $\mathring{\mathcal{A}}$, u vanishes there and $U \geqslant 0$ on $\partial\Omega$ by definition of $\mathcal{K}_h$,* (3.7)*. The disadvantage of using this space is that in general the Lagrange interpolant does not preserve the boundary constraints, forcing us to use an alternative approximation.*

To derive a posteriori error bounds, we require results on the stability and regularity of the dual problem (3.36). The key difference between the primal and dual problems is that in the primal case the inequality constraints are posed on the whole boundary, whereas the dual problem has a mixed condition, which generally involve more difficult analysis. In the primal case, one can derive a weak formulation based upon the minimisation of a functional defined by a convex function that determines the boundary conditions as in theorem 2.4.5. Based upon this weak formulation, regularity and stability is shown in [52]. In the dual case, this formulation cannot be applied. Indeed, examination of the dual problem gives weaker regularity, as we will now quantify.

Before presenting the necessary results, we make the following definitions regarding the corners of the domain and the boundary conditions.

Definition 3.7.2. *We let $\{x_1, ..., x_{N_p}\}$ be the union of the set of points that make up the boundary of $\mathring{\mathcal{A}}$ and the set of vertices of the polygonal domain boundary. Note that this is a finite set consisting of the vertices of the domain boundary and the points at which there is a change in boundary condition. We let ω_k denote the angle at the boundary at x_k, that is, $\omega_k = \pi$ at a change* *of boundary conditions or $0 < \omega_k < \pi$ at the corners of the domain, since we assume that the domain is convex. We will denote by r_k, θ_k the polar coordinates centred at point x_k, ζ_k a smooth cut-off function identically equal to 1 in a neighbourhood of x_k, and $\phi_k = \phi_k(\theta_k)$ a trigonometric function. These functions are the key tool in quantifying the behaviour of the solution around the points x_k.*

The properties of the dual problem that we will need are summarised in the following lemma. Our argument is simpler than the one given in [34] as their dual problem is mesh-dependent, whereas ours only depends on the contact set of the continuous problem. However, we note that since in both cases the portion of the boundary on which inequality constraints are imposed are unions of open intervals, the argument given in [34] is still applicable to our problem.

Proposition 3.7.3 (Properties of the dual problem [34])**.** *Let u solve (3.4) for some $f \in \mathrm{L}^2(\Omega)$, and $U \in \mathcal{V}_h$ solve (3.8). Then the dual problem (3.36) is uniquely solvable in $\mathrm{H}^1(\Omega)$. In addition, we have $z \geqslant 0$ in Ω and $z = 0$ on $\mathcal{A}$, and for any $\varepsilon \in (0, 1/2)$ there exists $C > 0$ independent of h such that the stability bound*

$$\|z\|_{\mathrm{H}^1(\Omega)} \leqslant C \left\| \max(0, u - U)^3 \right\|_{\mathrm{L}^{(4-\epsilon)/3}(\Omega)} \tag{3.37}$$

holds. Furthermore, we have the following regularity result for any $\varepsilon \in (0, \frac{1}{2})$

$$\|z\|_{\mathrm{W}^{2,(4-\varepsilon)/3}(\Omega)} \leqslant C \left\| \max(0, u - U)^3 \right\|_{\mathrm{L}^{(4-\epsilon)/3}(\Omega)}. \tag{3.38}$$

Proof. We first claim that $a(z^+, z^-) = 0$. To see this, note that by definition we have $z = z^+ + z^-$.

$$a(z^+, z^-) = \int_\Omega \nabla z^+ \cdot \nabla z^- + z^+ z^- \,\mathrm{d}x \tag{3.39}$$

It is clear that $z^+ z^- \equiv 0$ since z cannot be both positive and negative.

Further, we can write $\Omega = \{x \in \mathbb{R}^2 : z < 0\} \cup \{x \in \mathbb{R}^2 : z \geqslant 0\}$. Note that in each set, precisely at least one of z^+ and z^- is identically the zero function, and therefore if the set has nonzero measure, its weak gradient must be zero there. Contributions to the integral (3.39) must therefore be zero and the claim is proved.

Since $z \in \mathcal{M}$, $z \leqslant 0$ on $\mathcal{A}$, its positive part must be zero there, and we may take $v = z^+$ as test function in (3.36), which yields

$$-\left\|z^-\right\|^2_{\mathrm{H}^1(\Omega)} = a(z, -z^-) \geqslant \left\langle \max(0, u - U)^3, -z^- \right\rangle . \tag{3.40}$$

In other words, since $\max(0, u - U)^3$ is non-negative and z^- is non-positive,

$$0 \leqslant \left\|z^-\right\|^2_{\mathrm{H}^1(\Omega)} \leqslant \left\langle \max(0, u - U)^3, z^- \right\rangle \leqslant 0, \tag{3.41}$$

so that $z^- \equiv 0$, which proves that $z \geqslant 0$. This in turn gives $z = 0$ on $\mathring{\mathcal{A}}$.

The bound (3.37) follows after noting that we can take $v = 0$ in (3.36) to see

$$a(z, -z) \geqslant \left\langle \max(0, u - U)^3, -z \right\rangle \tag{3.42}$$

and $v = 2z$ in (3.36), yielding

$$a(z, z) \geqslant \left\langle \max(0, u - U)^3, z \right\rangle . \tag{3.43}$$

Therefore,

$$\begin{aligned} \|z\|^2_{\mathrm{H}^1(\Omega)} &= \left\langle \max(0, u - U)^3, z \right\rangle \\ &\leqslant \left\|\max(0, u - U)^3\right\|_{\mathrm{L}^{(4-\varepsilon)/3}(\Omega)} \|z\|_{\mathrm{L}^{(4-\varepsilon)/(1-\varepsilon)}(\Omega)} \\ &\leqslant C_{\mathrm{Sob}} \left\|\max(0, u - U)^3\right\|_{\mathrm{L}^{(4-\varepsilon)/3}(\Omega)} \|z\|_{\mathrm{H}^1(\Omega)} , \end{aligned}$$

where we have used the embedding of $\mathrm{H}^1(\Omega)$ in $\mathrm{L}^{(4-\varepsilon)/(1-\varepsilon)}(\Omega)$. The result now follows after dividing by $\|z\|_{\mathrm{H}^1(\Omega)}$.

Now, z is the weak solution of the boundary value problem

$$-\Delta z + z = \max(0, u - U)^3, \tag{3.44}$$

with zero Dirichet boundary condition on $\mathcal{A}$ and zero Neumann condition on the remainder of the boundary. The problem therefore fits into the framework of [52, thm 4.4.3.7], which allows us to quantify the regularity of z to the following extent. With the notation of definition 3.7.2, there exist unique real numbers $c_{k,m}$ such that

$$z - \sum_{\substack{1 \leqslant k \leqslant N_p \\ -2/q < \lambda_{k,m} < 0 \\ \lambda_{k,m} \neq -1}} c_{k,m} r^{-\lambda_{k,m}} \zeta_k \phi_k \in \mathrm{W}^{2,p}(\Omega) \tag{3.45}$$

where $\lambda_{k,m}$ are eigenvalues of a Laplacian operator which depends upon ω_k (see section 4.4 of [52] for complete enumeration of the $\lambda_{k,m}$ in all cases). In our case where Ω is a convex polygonal domain, the regularity of z is determined by the term in (3.45) with the lowest power of r, which is $r^{1/2}\eta_k\phi_k$. The singularity of type $r^{1/2}$ occurs at points where the boundary condition changes (compare with [4, §3] on this point).

Now, we observe that $r^{1/2}\eta_k\phi_k \in \mathrm{W}^{s,t}(\Omega)$ if and only if $s - \frac{2}{t} < \frac{1}{2}$. This gives a limit on the regularity of the second derivatives of z: $z \in \mathrm{W}^{2,t}(\Omega)$ only if $t < \frac{4}{3}$.

The proof of the stability bound in $\mathrm{W}^{2,(4-\varepsilon)/3}(\Omega)$ is a technical argument that we will not reproduce here, but can be found in full in lemma 5.2 of [34].

□

Remark 3.7.4 (Smoothness of the dual problem)**.** *It is important to note that the limit on regularity for the dual problem is imposed by the assumption of convexity of the domain and the nature of the boundary conditions, not the regularity of the problem data f. Indeed, the left hand side of equation* (3.45) *posesses higher regularity depending on the problem data, but the regularity of* *the singular components of u is limited by the geometry. For certain boundary conditions including the Signorini condition* (3.2)*, $\mathrm{H}^2(\Omega)$-regularity follows from proposition 3.3.1 which implies that the singular part of the solution vanishes. For the mixed boundary condition, this result does not apply, and the regularity is therefore constrained by the least regular of the components in* (3.45)*.*

Remark 3.7.5 (Choice of norm for a posteriori error estimation.)**.** *Proposition 3.7.3 also motivates a posteriori error estimation in $\mathrm{L}^4(\Omega)$. Comparing to the situation in §2.6, where we derive error estimates in $\mathrm{L}^p(\Omega)$ by using dual stability in $\mathrm{W}^{2,q}(\Omega)$, we see that we can only take this approach here for $q < \frac{4}{3}$, whose Hölder conjugate is 4. Thus, $\mathrm{L}^4(\Omega)$ is the strongest space in which estimates of the traditional residual form can be established.*

Remark 3.7.6. *Let u be the solution of (3.1)-(3.2) and suppose that $w \in \mathrm{H}^1(\Omega)$ has $tr(w) = 0$ on $\mathring{\mathcal{A}}$. Then*

$$a(u, w) = \langle f, w \rangle \tag{3.46}$$

In light of the structure of the contact set, described in remark 3.3.8, it can be seen (see [34, thm 2.3]) that u solves the boundary value problem

$$\begin{aligned} -\Delta u + u = f &\quad \in \Omega, \\ u = 0 &\quad on\, \Gamma_i, i = 1, ..., N, \\ \nabla u \cdot \boldsymbol{n} = 0 &\quad on\, \Gamma_i, i = N+1, ..., N+M. \end{aligned} \tag{3.47}$$

Here, $\Gamma_i \subset \Omega$ are disjoint open line segments such that $\bigcup_{i=1}^{N+M} \bar{\Gamma}_i = \partial\Omega$, and such that there is a set R of measure zero such that

$$\mathcal{A} = \bigcup_{i=1}^{N} \bar{\Gamma}_i \cup R.$$

The claim now follows immediately by testing (3.47) *with w and integrating by parts.*

Lemma 3.7.7. *Let u be the solution of (3.1)-(3.2), U be the finite element approximation satisfying (3.8), z be the dual solution of (3.36) and $\widehat{\Pi}(z)$ be the bilateral approximation of z. Then*

$$a\Big(\widehat{\Pi}(z), u - U\Big) \leqslant 0 \tag{3.48}$$

Proof. By equation (3.46) and since $0 \leqslant \widehat{\Pi}(z) \leqslant z = 0$ on $\mathring{\mathcal{A}}$, we have immediately that

$$a\Big(u, \widehat{\Pi}(z)\Big) = \Big\langle f, \widehat{\Pi}(z)\Big\rangle . \tag{3.49}$$

Since $\widehat{\Pi}(z) \geqslant 0$, we may take $\Phi = U + \widehat{\Pi}(z)$ in (3.8), giving

$$a\Big(U, -\widehat{\Pi}(z)\Big) \leqslant -\Big\langle f, \widehat{\Pi}(z)\Big\rangle . \tag{3.50}$$

The result then follows after combining (3.49) and (3.50). □

We are now ready to state and prove the main result of this chapter, a rigorous a posteriori bound in $\mathrm{L}^4(\Omega)$ for the problem (3.1) - (3.2). We prove this in two parts, bounding seperately the quantities $\|(u-U)^+\|_{\mathrm{L}^4(\Omega)}$ and $\|(u-U)^-\|_{\mathrm{L}^4(\Omega)}$.

Theorem 3.7.8 (A posteriori upper bound for the positive part of the error)**.** *Let u solve the variational inequality and U be the finite element approximation. Let*

$$\begin{aligned} p &:= \frac{4-\epsilon}{1-\varepsilon} \\ q &:= \frac{4-\varepsilon}{3}, \end{aligned} \tag{3.51}$$

for $\varepsilon > 0$. *Then*

$$\left\|(u-U)^+\right\|^4_{\mathrm{L}^4(\Omega)} \leqslant \eta(U,f,h) := C \sum_{K\in\mathscr{T}} \left(\eta_K^p + \frac{1}{2}\eta_J^p\right), \tag{3.52}$$

where

$$\begin{aligned} \eta_K &= h^2 \left\|-\Delta U + U - f\right\|_{\mathrm{L}^p(K)} \\ \eta_J &= h^{2-1/q} \left\|[\![\nabla U]\!]\right\|_{\mathrm{L}^p(\partial K)}. \end{aligned} \tag{3.53}$$

Proof. We may select $v = z + u - U$ in (3.36); this function is shown to belong to $\mathcal{M}$ in remark 3.7.1. We therefore have

$$\int_\Omega (u-U)_+^4 \,\mathrm{d}x \leqslant a(z, u-U). \tag{3.54}$$

We may use lemma 3.7.7 to introduce $\widehat{\Pi}(z)$ as follows:

$$\int_\Omega (u-U)_+^4 \,\mathrm{d}x \leqslant a\Big(z - \widehat{\Pi}(z), u-U\Big). \tag{3.55}$$

Since z and $\widehat{\Pi}(z)$ both have zero trace on $\mathcal{A}$, we can use remark 3.7.6 to arive at

$$\int_\Omega (u-U)_+^4 \,\mathrm{d}x \leqslant l\Big(z - \widehat{\Pi}(z)\Big) - a\Big(z - \widehat{\Pi}(z), U\Big).$$

The rest of the argument follows standard a posteriori techniques.

$$\begin{aligned} l\Big(z-\widehat{\Pi}(z)\Big) - a\Big(U, z-\widehat{\Pi}(z)\Big) &= \int_\Omega f\Big(z-\widehat{\Pi}(z)\Big) + \nabla U\cdot\Big(\nabla z - \nabla\widehat{\Pi}(z)\Big) - U\Big(z-\widehat{\Pi}(z)\Big)\,\mathrm{d}x \\ &= \int_\Omega (f + \Delta U - U)\Big(z-\widehat{\Pi}(z)\Big)\,\mathrm{d}x - \int_{\mathscr{E}} [\![\nabla U]\!]\Big(z-\widehat{\Pi}(z)\Big)\,\mathrm{d}S \\ &=: R_1 + R_2. \end{aligned} \tag{3.56}$$

Splitting the integral elementwise and making use of the Cauchy-Schwarz inequality, we see that

$$\begin{aligned} R_1 &= \int_\Omega (f + \Delta U - U)\Big(z - \widehat{\Pi}(z)\Big)\,\mathrm{d}x \\ &\leqslant \sum_{K\in\mathscr{T}} \left\| h^2(f + \Delta U - U) \right\|_{\mathrm{L}^p(K)} \left\| h^{-2}\Big(z - \widehat{\Pi}(z)\Big) \right\|_{\mathrm{L}^q(K)}. \end{aligned} \tag{3.57}$$

Invoking the optimal approximation result of lemma 3.6.1, we obtain

$$R_1 \leqslant C_b \sum_{K\in\mathscr{T}} \left\| h^2(f + \Delta U - U) \right\|_{\mathrm{L}^p(K)} |z|_{\mathrm{W}^{2,q}(K)}. \tag{3.58}$$

Similarly, we bound R_2,

$$\begin{aligned} R_2 &= -\sum_{e\in\mathscr{E}} \int_e [\![\nabla U]\!]\Big(z - \widehat{\Pi}(z)\Big) \\ &\leqslant \frac{1}{2}\sum_{K\in\mathscr{T}} \left(\sum_{e\in\partial K} \left\| h^{2-1/q} [\![\nabla U]\!] \right\|_{\mathrm{L}^p(e)} \left\| h^{-2+1/q}\Big(z - \widehat{\Pi}(z)\Big) \right\|_{\mathrm{L}^q(e)}. \right) \end{aligned} \tag{3.59}$$

Using again the optimal approximation result in lemma 3.6.1 combined with trace estimates (see proposition 2.5.4) we arrive at

$$R_2 \leqslant C_b C_{\mathrm{tr}} \frac{1}{2} \sum_{K\in\mathscr{T}} \left\| h^{2-1/q} [\![\nabla U]\!] \right\|_{\mathrm{L}^p(\partial K)} \|z\|_{\mathrm{W}^{2,q}(\widehat{K})}. \tag{3.60}$$

Collecting (3.58)–(3.60), we have

$$l\Big(z - \widehat{\Pi}(z)\Big) - a\Big(U, z - \widehat{\Pi}(z)\Big) \leqslant C \sum_{K\in\mathscr{T}} \left(\eta_K + \frac{1}{2}\eta_J \right) \|z\|_{\mathrm{W}^{2,q}(\widehat{K})}. \tag{3.61}$$

Hence, the result follows from a discrete Cauchy-Schwarz inequality and the regularity bound on z given in equation (3.38). □

We now prove that the negative part of the error also satisfies a bound of the form (3.52).

Once again taking an analogous approach to that in [72], we define a second dual problem as follows. Let $\bar{\mathcal{M}} = \{v \mid v \geqslant 0 \text{ on } \mathcal{A}_U\}$ where we recall that $\mathcal{A}_U$ is the discrete contact set. Find $\bar{z} \in \bar{\mathcal{M}}$ such that

$$a(\bar{z}, v - \bar{z}) \geqslant \left\langle \min(0, u - U)^3, v - \bar{z} \right\rangle \quad \forall v \in \bar{\mathcal{M}}. \tag{3.62}$$

Key properties follow in a manner analogous to proposition 3.7.3, and are summarised in the following proposition.

Proposition 3.7.9 (Properties of dual problem for the negative part of the error.)**.** *Let u solve (3.4) for some $f \in \mathrm{L}^2(\Omega)$, and $U \in \mathcal{V}_h$ solve (3.8). Then the dual problem (3.62) is uniquely solvable in $\mathrm{H}^1(\Omega)$. In addition we have $\bar{z} \leqslant 0$ in Ω and $\bar{z} = 0$ on $\mathcal{A}_U$. For any $\varepsilon \in (0, 1/2)$ there exists $C > 0$ independent of h such that the stability bound*

$$\|\bar{z}\|_{\mathrm{H}^1(\Omega)} \leqslant C \left\| (\min(0, u - U)^3 \right\|_{\mathrm{L}^{(4-\epsilon)/3}(\Omega)} \tag{3.63}$$

holds. In addition, for sufficiently small h, we have the following regularity result for any $\varepsilon \in (0, 1/2)$

$$\|\bar{z}\|_{\mathrm{W}^{2,(4-\varepsilon)/3}(\Omega)} \leqslant C \left\| \min(0, u - U)^3 \right\|_{\mathrm{L}^{(4-\epsilon)/3}(\Omega)}. \tag{3.64}$$

We may now once again use the two-sided approximation. Let $\widehat{\Pi}(\bar{z})$ be a finite element function such that $\bar{z} \leqslant \widehat{\Pi}(\bar{z}) \leqslant 0$.

Theorem 3.7.10 (A posteriori upper bound for the negative part of the error)**.** *Let u solve the variational inequality and U be the FE approximation. Let*

$$\begin{aligned} p &:= \frac{4 - \epsilon}{1 - \varepsilon} \\ q &:= \frac{4 - \varepsilon}{3}, \end{aligned} \tag{3.65}$$

for $\varepsilon > 0$. Then

$$\left\|(u-U)^{-}\right\|_{\mathrm{L}^4(\Omega)}^4 \leqslant \eta(U,f,h) := C \sum_{K\in\mathscr{T}} \left(\eta_K^p + \eta_J^p\right), \tag{3.66}$$

where

$$\begin{aligned} \eta_K &= h^2 \left\|-\Delta U + U - f\right\|_{\mathrm{L}^p(K)} \\ \eta_J &= h^{2-1/q} \left\|[\![\nabla U]\!]\right\|_{\mathrm{L}^p(\partial K)}. \end{aligned} \tag{3.67}$$

Proof. We begin by assuming that condition $\mathbf{A_h}$) holds, see remark 3.4.2). We may immediately take $v = \bar{z} + u - U$ in the dual problem since $U = 0$ and $u \geqslant 0$ on $\mathcal{A}_U$. We therefore have

$$\int_\Omega \min(0, u-U)^4 \leqslant a(u-U, \bar{z}). \tag{3.68}$$

Since we know that $U = 0$ on $\mathcal{A}_U$ and non-negative on $\partial\Omega$, and since we also have $\bar{z} = 0$ on $\mathcal{A}_U$ and non-positive on Ω, we can choose sufficiently small $s > 0$ so that $U \pm s\widehat{\Pi}(\bar{z}) \in \mathcal{K}_h$. Taking $\Phi = U \pm s\widehat{\Pi}(\bar{z})$ as test functions in the discrete problem (3.8) gives

$$\begin{aligned} a\Big(U, \widehat{\Pi}(\bar{z})\Big) &\geqslant \Big\langle f, \widehat{\Pi}(\bar{z})\Big\rangle \\ a\Big(U, -\widehat{\Pi}(\bar{z})\Big) &\geqslant -\Big\langle f, \widehat{\Pi}(\bar{z})\Big\rangle \end{aligned} \tag{3.69}$$

and therefore we must have

$$a\Big(U, \widehat{\Pi}(\bar{z})\Big) = \Big\langle f, \widehat{\Pi}(\bar{z})\Big\rangle. \tag{3.70}$$

Choosing $v = u - \widehat{\Pi}(\bar{z}) \in \mathcal{K}$ in (3.4) we also have

$$a\Big(u, -\widehat{\Pi}(\bar{z})\Big) \geqslant -\Big\langle f, \widehat{\Pi}(\bar{z})\Big\rangle. \tag{3.71}$$

Together with (3.68), we have shown

$$\int_\Omega \min(0, u - U)^4 \leqslant a\Big(u - U, \bar{z} - \widehat{\Pi}(z)\Big)\,. \tag{3.72}$$

To introduce the problem data, f, of the primal problem, we choose $v = u + \widehat{\Pi}(\bar{z}) - \bar{z}$ as test function in (3.4). Since $\widehat{\Pi}(\bar{z})$ lies between the (non-positive) $\bar{z}$ and zero, this function lies in $\mathcal{K}$, and we obtain

$$a\Big(u, \widehat{\Pi}(\bar{z}) - \bar{z}\Big) \geqslant \Big\langle f, \widehat{\Pi}(\bar{z}) - \bar{z}\Big\rangle\,, \tag{3.73}$$

or, equivalently.

$$a\Big(u, \bar{z} - \widehat{\Pi}(\bar{z})\Big) \leqslant \Big\langle f, \bar{z} - \widehat{\Pi}(\bar{z})\Big\rangle\,. \tag{3.74}$$

Inserting this in (3.72), we now have

$$\int_\Omega \min(0, u - U)^4 \leqslant \Big\langle f, \bar{z} - \widehat{\Pi}(\bar{z})\Big\rangle - a\Big(U, \bar{z} - \widehat{\Pi}(\bar{z})\Big) \tag{3.75}$$

The proof now proceeds exactly as in Theorem 3.7.8. □

Remark 3.7.11 (A posteriori lower bound)**.** *We remark here with interest that in light of remark 3.7.6, the derivation of the lower bound in §2.6.1 for the analogous boundary value problem goes through here with no modifications. Indeed, the argument does not require detailed regularity analysis, does not use a dual problem and since u satisfies a weak form as shown in 3.7.6, is not hindered by the inequality constraints at the boundary and can proceed in exactly the same fashion. We therefore have two-sided bounds on the error.*

3.8 Numerical Experiments

In this section, we present numerical results to demonstrate the effectiveness of the error estimate and adaptive routine against an exact solution within the framework of sections 3.3 and 3.7, that is, convex $\Omega \subseteq \mathbb{R}^2$ with polygonal boundary. We then present some more challenging situations in which theory of the preceding sections may fail, but in which the error estimate may still prove useful. In particular, we test the estimate on a non-convex domain in $\mathbb{R}^2$ with a re-entrant corner as well as a three-dimensional example.

All simulations presented here are conducted using `deal.II`, an open source C++ software library providing tools for adaptive finite element computations [5]. We note here that `deal.II` uses quadrilateral meshes. Since all coarse meshes are uniform meshes consisting of squares, and refinement is by quadrisection, regularity of the mesh does not degrade under heavy refinement. At each stage of the iterative process used to solve the variational inequality, a preconditioned conjugate gradient method is used to solve the algebraic problem. As the mesh is refined, hanging nodes are created, and are constrained so that the resulting numerical solution is continuous.

3.8.1 Example 1

For our first example, we choose the exact solution used by the authors of [34]. Let r, θ denote polar coordinates centred at $(0.5, 0)$, that is

$$r(x,y) = ((x-0.5)^2 + y^2)^{1/2} \tag{3.76}$$

$$\theta(x,y) = \arccos\left(\frac{x-0.5}{r}\right) \tag{3.77}$$

We define the function

$$\tilde{u}(r,\theta) := -r^{3/2}\sin(\tfrac{3}{2}\theta), \tag{3.78}$$

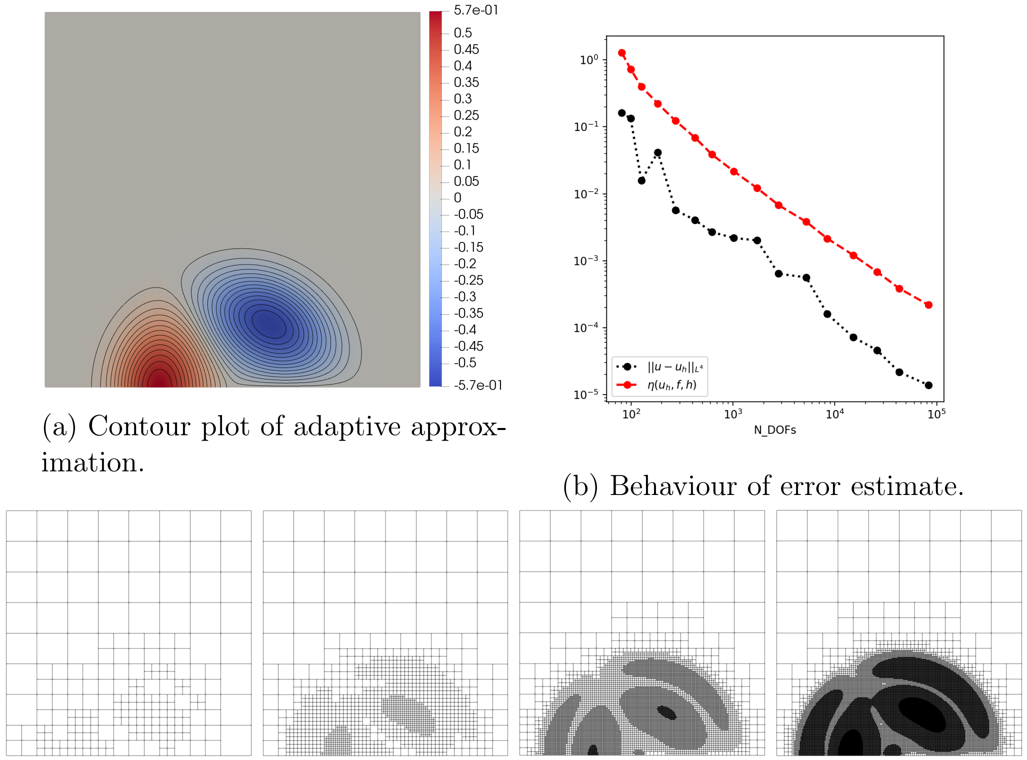
and define ψ to be a ninth order spline defined by endpoint values and gradients as follows (see also section 6 of [34]). We impose the following values to determine ψ.

$$\psi(0) = 1,\ \psi'(0) = ... = \psi''''(0) = \psi(0.45) = \psi'(0.45) = ... = \psi''''(0.45) = 0, \tag{3.79}$$

and $\psi(x) = 0$ for any $x < 0$ or $x \geqslant 0.45$. Then if we choose f accordingly, $10\psi\tilde{u}$ solves (3.4).

The problem is initialised on a coarse uniform mesh with $h = 1/8$. For the adaptive algorithm we use Dörfler marking (see §2.7) with refine fraction $\beta = 0.9$ and no coarsening. A numerical approximation to problem (3.4) with datum f chosen so that $10\psi\tilde{u}$ is the exact solution is shown in figure 3.2a, represented on the final mesh $\mathcal{T}^{15}$ produced by the adaptive algorithm, consisting of around 80,000 degrees of freedom. The progression of the $\mathrm{L}^4(\Omega)$-error, $\|u - u_h\|_{\mathrm{L}^4(\Omega)}$, and error estimate $\eta^{1/4}$ as defined in equation (3.52) under adaptive mesh refinement is shown in figure 3.2b. For a given number of degrees of freedom, the adaptive algorithm reduces the error and appears to give a slightly better convergence rate than uniformly refining the mesh in terms of degrees of freedom. We observe that the over-estimation factor (or effectivity index as it is often called) of the error estimate, is approximately 20. As expected from the theory, this factor is approximately constant once the asymptotic regime is reached. The two-sided bounds on the error ensure that the the true error decreases at the same rate as the error estimate. Comparing with figure 3.4a, we see that asymptotically, error is lower under adaptive refinement than uniform in terms of degrees of freedom, although both exceed the optimal rate.

A selection of adaptive meshes are shown in figures 3.2c-3.2f. These meshes show refinement around the main features of the solution. In particular, the mesh is heavily refined around large gradients, along with significant refinement where the boundary conditions change type and constraints are active. We remark here that accurate resolution early on in the adaptive process is particularly important due to the assumptions made on the discrete contact set in theorem 3.7.10.



(a) Contour plot of adaptive approximation.

(b) Behaviour of error estimate.

(c) $\mathcal{T}^4$ (d) $\mathcal{T}^8$ (e) $\mathcal{T}^{12}$ (f) $\mathcal{T}^{15}$

Figure 3.2: Example 1 §3.8.1, contour plot and various iterations of the adaptive mesh $\mathcal{T}^i$ of an approximation to a function $u \in \mathrm{H}^2(\Omega)$. Figure 3.2b shows a double-logarithmic plot comparing the error in $\mathrm{L}^4(\Omega)$ (black line) with the error estimate (red line).

3.8.2 Example 2: Re-entrant Corner

To test the estimate in the presence of a geometric singularity, we introduce a re-entrant corner to the domain. In this example, data f is chosen to ensure that both boundary constraints are active. The problem data is selected to try and force the solution to be close to

$$w := \sin(2\pi(r-b)^2) - 0.5, \tag{3.80}$$

where $r = (x^2+y^2)^{\frac{1}{2}}$ and b can be varied to force different behaviours of the solution. For this example we make the choice $b = 0.91$ and set $f = \Delta w + w$.

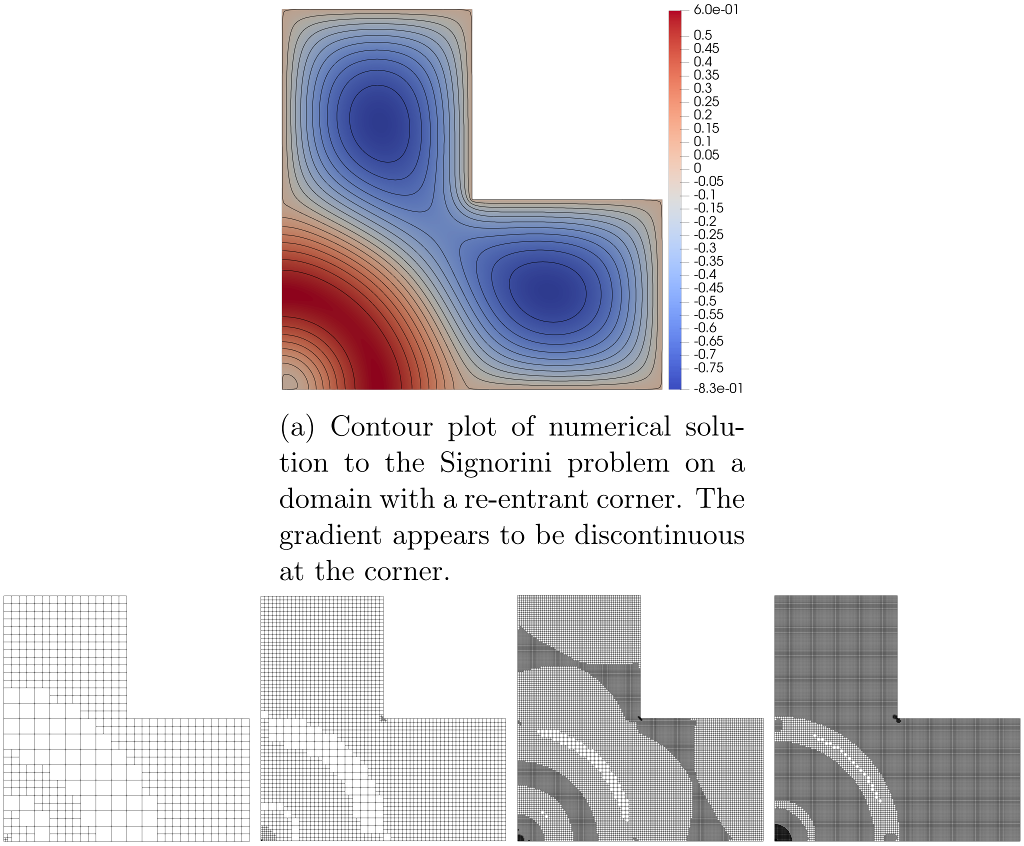
A numerical solution to this problem is shown in figure 3.3a. Under uniform mesh refinement, the error estimate converges to zero at a suboptimal rate. Optimality is restored using an adaptive routine utilising Dörfler marking with refinement fraction of 0.8 and no coarsening. A sample of the meshes produced is given in figures 3.3b-3.3e. The behaviour of the error estimate under uniform and adaptive refinement is shown in figure 3.5. The slopes of the error-reduction curves reveal suboptimal convergence in the uniform case which is somewhat recovered by adaptive mesh refinement, but to different degrees. This time we see most refinement around gradients and features in the solution. Interestingly, there is little refinement around the re-entrant corner where the solution is expected to lose regularity. Instead, the error estimate prioritises resolution of the solution near where the boundary constraints are active. This simulation was performed with the Dörfler marking criterion with refine fraction 0.8 and no coarsening. The algorithm was initialised on a coarse uniform mesh with $h = 1/16$.

3.8.3 Example 3, $d = 3$.

Working in three dimensions and in spherical polar coordinates, centred at $(0.5, 0, 0.5)$, we now consider a test analogous to example 1. Let (r, θ, ϕ) denote spherical polar coordinates centred at $(0.5, 0, 0)$. We note that the function

$$u = -10\psi(r) r^{3/2} \sin\left(\frac{3}{2}\varphi\right) \sin\left(\frac{3}{4}\theta\right) \tag{3.81}$$

satisfies appropriate constraints on the boundary and so solves (3.1), (3.2) for appropriately defined f. Contours of (3.81) are shown in figure 3.6. Again



(a) Contour plot of numerical solution to the Signorini problem on a domain with a re-entrant corner. The gradient appears to be discontinuous at the corner.

(b) $\mathcal{T}^3$

(c) $\mathcal{T}^8$

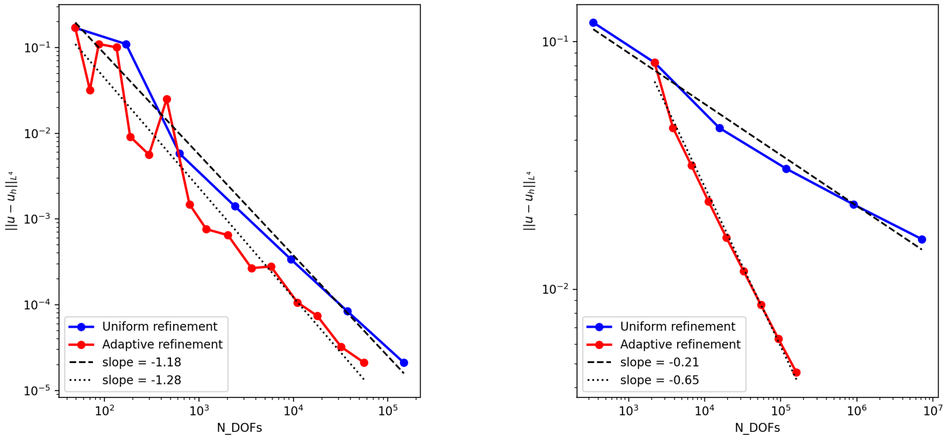
(d) $\mathcal{T}^{12}$

(e) $\mathcal{T}^{15}$

Figure 3.3: Example 2 §3.8.2, contour plot and various iterations of the adaptive mesh $\mathcal{T}^i$ of an approximation to a function $u \in \mathrm{W}^{2,(4-\varepsilon)/3}(\Omega) \setminus \mathrm{H}^2(\Omega)$. In this case the mesh refines around the reentrant corner as well as along qualitative features of the solution and where the boundary conditions change type.

we use Dörfler marking with refine fraction of 0.8 and no coarsening.

Adaptive meshes from several stages of the adaptive algorithm are displayed in 3.7 (note that these are slices of a three-dimensional mesh). We observe that mesh resolution is greater close to the boundary where the Signorini constraints are active (in this case the plane defined by $y = 0$) and provides good resolution of the contact set. Refinement is also present



(a) Example 1. Expected slope from theory is -1.

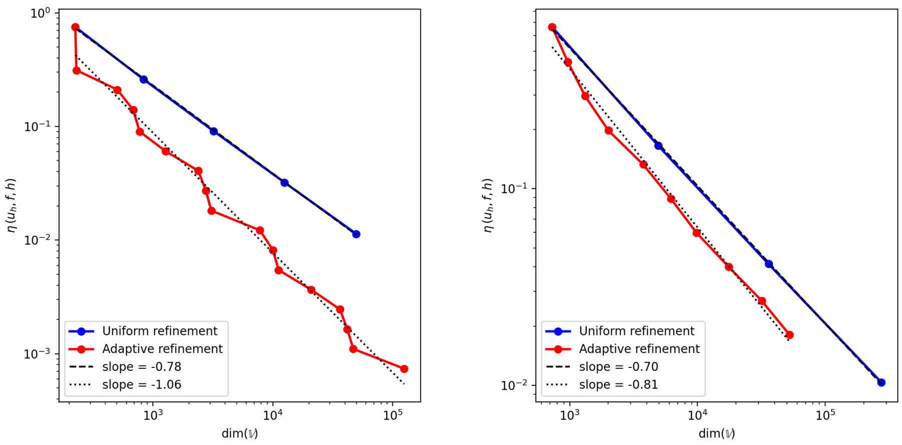
(b) Example 3. Expected slope from theory is $-\frac{2}{3}$.

Figure 3.4: Comparison of $L^4(\Omega)$ error computed against exact solutions for uniform and adaptive meshes for those examples for which an exact solution is available. The orders of convergence of the error estimate in each case (that is, the slope of the error values) is approximated from the log-scaled data by performing standard least squares linear regression.

around key features of the solution, as obsered for example 1, figures 3.2a and 3.2c-3.2f

3.9 Conclusions & discussion

In this chapter, we derived rigorous error estimates for the scalar Signorini problem in a non-energy norm using a duality argument. The error estimates were benchmarked against known exact solutions and shown to have optimal order, and to be sharp to a satisfactory degree (overestimation factor of order 10). Adaptive mesh refinement was shown to decrease error for a given number of degrees of freedom compared to uniform meshes, particularly for a challenging three-dimensional problem. The error estimate was tested on



(a) Example 2. Expected slope from theory is -1.

(b) Example 3. Expected slope from theory is $-\frac{2}{3}$.

Figure 3.5: Double logarithmic plot of error estimate against number of degrees of freedom under uniform and adaptive mesh refinement for examples 2 and 3, that is, the examples that do not meet the necessary assumptions for the theory above to hold. The orders of convergence of the error estimate in each case is approximated from the log-scaled data by performing standard least squares linear regression.

more challenging test cases, both in situations where the assumptions of the main theorem were satisfied, and in a case where they were not (three spatial dimensions, re-entrant corners). In the three-dimensional case where the theory does not hold (see section 3.8.3 and figure 3.5b, we remark that the error estimate is too optimistic in the sense that the error estimate decreases at the expected rate under uniform refinement, whereas the error does not (see figure 3.4b). This means that although adaptivity did produce superior error reduction in this case, we cannot use the error estimate as a means to meet a given tolerance.

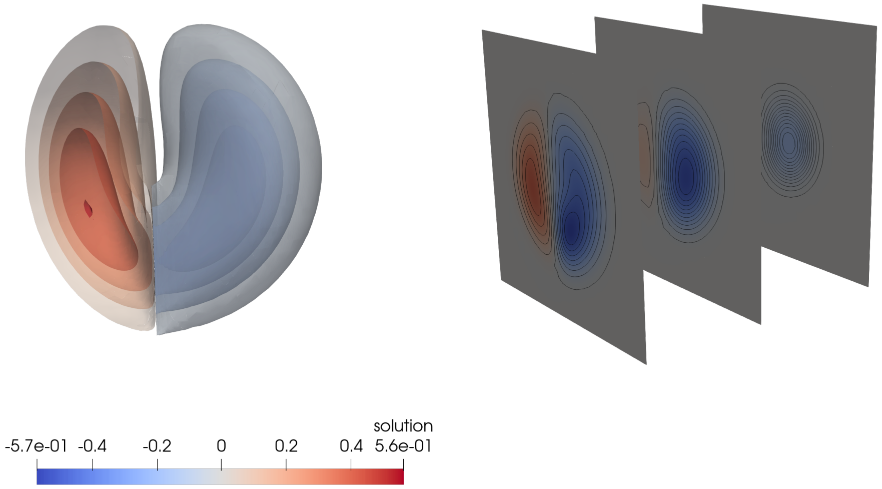


Figure 3.6: Left: contours of the numerical solution to example 3 §3.8.3, Right: slices of the numerical solution parallel to planes $y = c$ for $c = 0.1, 0.2$ and 0.3. Note that for visualisation purposes the right hand plot is not to scale; the planes have been translated along the y-axis.

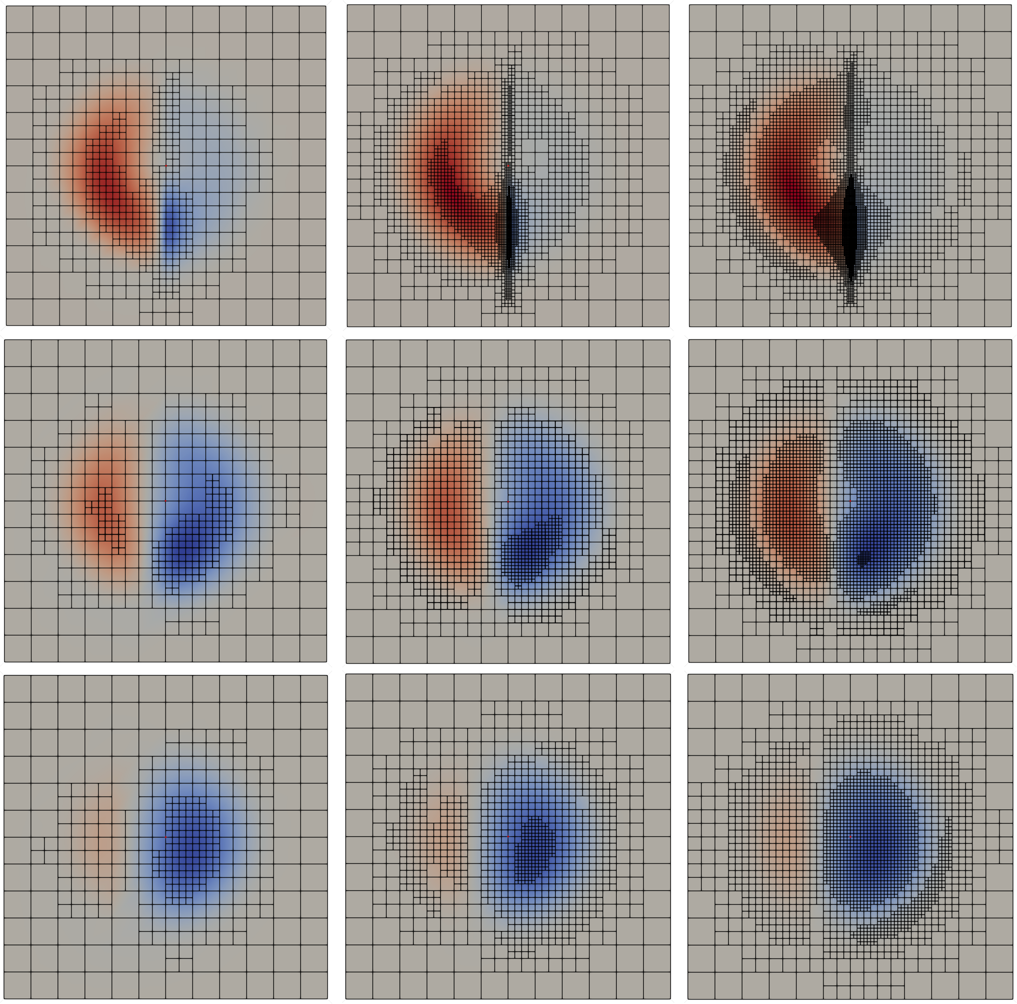


Figure 3.7: Adaptively generated meshes for example 3. The slices shown are the intersections of the mesh with the planes $y = 0.01$ (top row), $y = 0.1$, (middle row) and $y = 0.2$ (bottom row) and show the mesh after refinement cycles 2 (left hand column), 5 (middle column) and 8 (right hand column). The mesh is heavily refined along the z axis due to both active boundary constraints and gradient singularities in the solution.

Chapter 4

Adaptive modelling of variably saturated seepage problems

4.1 Abstract

This chapter is adapted from the paper [7] published during the candidate's PhD studies. In this chapter we present a goal-oriented adaptive finite element method for a class of subsurface flow problems in porous media, which exhibit seepage faces. We focus on a representative case of the steady state flows governed by a nonlinear Darcy-Buckingham law with physical constraints on subsurface-atmosphere boundaries. This leads to the formulation of the problem as a variational inequality similar in form to those introduced in §2.4 and analysed in §3 with the additional complications of mixed boundary conditions and nonlinear coefficients. The solutions to this problem are investigated using an adaptive finite element method based on a dual-weighted a posteriori error estimate, derived with the aim of reducing error in a specific target quantity. The quantity of interest is chosen as volumetric water flux across the seepage face, and therefore depends on an a priori unknown free boundary. The key contributions of this chapter are the application of the dual-weighted residual methodology to variably

saturated seepage problems and its application to challenging numerical examples. These example include specific case studies, from which this research originates, illustrating the major difficulties that arise in practical situations. We summarise extensive numerical results that clearly demonstrate the designed method produces rapid error reduction measured against the number of degrees of freedom.

4.2 Introduction

The modelling of subsurface flows in porous media presents a multitude of mathematical and numerical challenges. Heterogeneity in soils and rocks as well as sharp changes of several orders of magnitude in hydraulic properties around saturation are the multi-scale phenomena that are particularly difficult to capture in numerical models. In addition, physically realistic domains include a wide variety of boundary conditions, some of which depend upon a free (phreatic) surface and therefore also upon the problem solution itself. These boundary conditions are described by inequality constraints. At points where the active constraint switches from one to the other, gradient singularities in the solution can arise which must be resolved well to avoid polluting the accuracy of the solution. The situation is analogous to a thin obstacle problem, for which gradient discontinuities arise around the thin obstacle [71]. For these reasons, such problems are good candidates for h-adaptive numerical methods, where a computational mesh is automatically refined according to an indicator for the numerical error. It is the aim of such methods to provide the necessary spatial resolution with greater efficiency than is possible with structured meshes.

A common model for steady flow in porous media in the geosciences is a free surface problem where the medium is assumed to be either saturated with flow governed by Darcy's law or dry with no flow at all. The free surface is the boundary between the two regions with a no-flow condition applied across it. Some authors solve this as a pure free boundary problem where the computational domain is unknown a priori such as in [39]. However, this means that as the domain is updated, expensive re-meshing must take place, allowing fewer of the data structures to be re-used from one iteration to the next. To avoid the difficulties of this approach, in [21], the problem formulation is modified to a fixed domain in which flow can take place (such as a dam) and the pressure variable defined on the whole domain, removing the need for changes in problem geometry and costly re-meshing during numerical simulations. The theory of this type of formulation is described in detail in [83]. A good approximation theory is available for finite element methods applied to such problems. It should be noted though that this model is a simplification, owing to the fact that it does not allow for unsaturated effects.

To avoid the computational complexities of a changing domain, in this work we consider the porous medium to be variably saturated, and therefore we solve for pore pressure over the entire domain (cf [117]). The results presented in [90] suggest that this approach is in fact necessary to accurately represent the subsurface. It is also expected that this framework will allow relatively easy extension to unsteady cases where unsaturated effects are extremely important for the dynamics.

Although there has been much study of this problem, there are relatively few examples of adaptive finite element techniques being used. This is because the partial differential equation governing subsurface flow presents difficulties for the traditional theory of a posteriori estimation. This stems from the behaviour of the coefficient of hydraulic conductivity, which depends on the solution itself and approaches zero in the dry soil limit, leading to degeneracy of the PDE problem. This violates the standard assumption of stability in elliptic PDE problems.

Traditional residual a posteriori estimation for finite element methods such as those described in §2 and §3 gives upper bounds of the form

$$\|u - U\|_E \leqslant C\varepsilon(U, h, f) \tag{4.1}$$

where u is the exact solution to some partial differential equation, C is a positive constant, U is the numerical solution, h is the mesh function and f is problem data. C is usually only computable for the simplest domains and meshes, and can be large. The norm, $\| \cdot \|_E$, is a global measure chosen so that the asymptotic convergence rate of the method is optimal. In practical computations, however, the user is often not interested in asymptotic rates that may never be reached, but would prefer a sharp estimate of the error to give confidence in the approximation.

The dual-weighted residual framework for error estimation was inspired by ideas from optimal control as a means to estimate the error in approximating a general quantity of interest. Pursuing this analogy, the objective functional to be minimised is the error in numerically approximating a solution to the PDE problem, the constraints are the PDE problem and boundary conditions, and the control variables are local resolution in the spatial discretisation.

There has been a huge amount of work on error estimation and adaptivity using the dual-weighted approach and it has shown to be extremely effective in computing quantities which depend upon local features in steady-state problems in [53], heterogeneous media [37] and variable boundary conditions in variational inequalities [17, 103]. In almost all cases the performance of the goal based algorithm cannot be bettered in efficiency. The goal-based framework also extends to time dependent problems, where it has been applied to the heat equation by [96] and the acoustic wave equation by [9] among others.

A common feature of numerical methods for seepage problems in the literature is that they are designed around getting a good representation of the phreatic surface, namely the level set of zero pressure head that divides saturated from unsaturated soil. There are however many other possible quantities of interest such as flow rate over a seepage face that could represent the productivity of a well. In this work, correct representation of the phreatic surface is prioritised only if it is important for the calculation of the quantity of interest, and we let local mesh refinement do the work for us, rather than expensive re-meshing of the free surface. Indeed, in the current framework, mesh refinement is rather simple to implement and relatively cheap.

The dual-weighted residual method has been applied to linear problems with similar characteristics. In [17], a simplified version of the Signorini problem is solved. The authors of [37] consider a groundwater flow problem in which the focus is to estimate the error in the nonlinear travel time functional. In both cases, the underlying PDE operator is linear.

The key step in deriving an a posteriori error bound for this variational inequality is the introduction of an intermediate function that solves the unrestricted PDE corresponding to the inequality. This allows the removal of the exact solution from the resulting bound. Finally, the unrestricted solution allows the problem data to enter into the problem, allowing a fully computable a posteriori error bound. In this chapter, we apply these cutting edge techniques of a posteriori error estimation and adaptive computing to complex and relevant problems informed by geophysical applications. We demonstrate that the error bound is sharp and allows for highly efficient error reduction in the target quantity in a variety of situations which include geometric singularities, multi scale effects in layered media and complex boundary conditions at the seepage face.

The remainder of the chapter is set out as follows. In section 4.3, we describe the seepage problem and derive a weak formulation. The problem is discretised with a finite element method in section 4.4. Section 4.5 is devoted to the derivation of a dual-weighted a posteriori estimate for the finite element error. Sections 4.6 and 4.6.3 describe the particulars of the adaptive algorithm and our implementation of it. Section 4.7 contains numerical experiments, to illustrate the performance of the error estimate and adaptive routine in two test cases. Finally, section 4.8 contains the application of our adaptive routine to two case studies with experimental data chosen to illustrate some of the most difficult cases that arise in practice.

4.3 Description of Problem

In this section, we give the mathematical formulation of the seepage problem and derive its weak form. We note that the notation is slightly different to previous chapters to emphasise the flux and therefore the physical motivation for this problem. Let u denote the pressure head of fluid flowing in a porous medium in a bounded, convex domain $\Omega \subseteq \mathbb{R}^N$, $N = 2$ or 3 with boundary $\partial\Omega$. The flow of the fluid is described by the flux density vector $\mathbf{q}(u)$. Note that $\mathbf{q}(u)$ is not the fluid velocity $\mathbf{v}$, but is related to it by

$$\mathbf{v} = \frac{\mathbf{q}(u)}{\phi}, \tag{4.2}$$

where ϕ is the porosity of the medium, that is, the proportion of the medium that may be occupied by fluid. Flux density is related to the pressure field by

$$\mathbf{q}(u) := -k(u)\nabla\left(u + h_z\right), \tag{4.3}$$

where h_z is the vertical height above a fixed datum representing the action of gravity upon the fluid and k is a nonlinear function that characterises the hydraulic conductivity of the medium. We refrain from defining k precisely here as our analytic results only require abstract assumptions on the specific form of k. For our computational experiments we will make use of a van Genuchten model [106], which is defined in (4.56) and illustrated in Figure 4.2. The modification of Darcy's law following the observation that hydraulic conductivity depends upon the capillary potential u is due to [27], and is a generalisation of the standard Darcy law that applies to soil that is completely saturated. In this case, the coefficient k introduces strong nonlinearity into the problem.

Now consider the steady state and suppose that f is a source/sink term. Then we can combine (4.3) with the mass balance equation

$$\nabla \cdot \mathbf{q}(u) = f \tag{4.4}$$

to obtain the equation of motion for steady-state variably saturated flow

$$-\nabla \cdot (k(u)\nabla(u + h_z)) = f. \tag{4.5}$$

To complete the above system and solve it, boundary conditions must be specified. We briefly review the most relevant here and point an interested reader to [13] for a more complete list.

Boundaries that are in contact with a body of water can be modelled by enforcing a Dirichlet boundary condition $u = g$, where g is some function chosen based upon the assumption that the body has a hydrostatic pressure distribution. The boundary condition therefore enforces continuity of pressure head across the boundary. A hydrostatic condition can also be used to set the water table, and can represent the prevailing conditions far from the soil-air boundary.

The flow of water across a boundary is given by the component of the Darcy flux, (4.3), that is normal to the boundary. We will set $\mathbf{q}(u) \cdot \boldsymbol{n} = 0$ where $\boldsymbol{n}$ is the unit outward normal vector to $\partial\Omega$ to represent an impermeable boundary.

At subsurface-air boundaries, a set of inequality constraints must be satisfied. The ambient atmospheric pressure is set as the zero point, and the pressure of water in the soil at such a boundary can therefore not exceed zero. When this pressure is reached, water is forced out of the soil, creating a flux out of the domain. The portion of a subsurface-air boundary at which there is outward flux is known as a seepage face, and it is characterised by the following conditions:

$$u \leqslant 0, \quad \mathbf{q}(u) \cdot \boldsymbol{n} \geqslant 0, \quad \mathbf{q}(u) \cdot \boldsymbol{n} u = 0. \tag{4.6}$$

We note the similarity in structure, but different signs involved when compared with the Signorini boundary conditions in (3.2).

We are now ready to state the full problem. We divide the boundary of Ω, $\partial\Omega$, into Γ_A, Γ_N and Γ_D such that $\overline{\partial\Omega} = \overline{\Gamma_A} \cup \overline{\Gamma_N} \cup \overline{\Gamma_D}$. Here Γ_A stands for the portion of the boundary at which a seepage face may form, and Γ_N and Γ_D respectively denote portions of the boundary where it is known a priori that Neumann (respectively Dirichlet) boundary conditions are to be applied. The problem is to find u such that

$$\nabla \cdot \mathbf{q}(u) := -\nabla \cdot k(u)\nabla(u + h_z) = f \quad \text{in} \quad \Omega \tag{4.7}$$

$$\mathbf{q}(u) \cdot \boldsymbol{n} = 0 \quad \text{on} \quad \Gamma_N \tag{4.8}$$

$$u = g \quad \text{on} \quad \Gamma_D \tag{4.9}$$

$$u \leqslant 0, \quad \mathbf{q}(u) \cdot \boldsymbol{n} \geqslant 0, \quad \mathbf{q}(u) \cdot \boldsymbol{n} u = 0 \quad \text{on} \quad \Gamma_A, \tag{4.10}$$

where f denotes a source/sink and $g = g(h_z)$ is an affine function representing hydrostatic pressure. We also define the contact set to be the portion of the boundary along which the constraint $u \leqslant 0$ is active which is precisely the seepage face

$$\mathcal{A} := \{x \in \Gamma_A \mid u(x) = 0\}. \tag{4.11}$$

We refer to figure 4.1 for a visual explanation.

4.3.1 Weak Formulation

In this section, we write the seepage problem (4.7) - (4.10) in weak form. To that end, we define the following function sets:

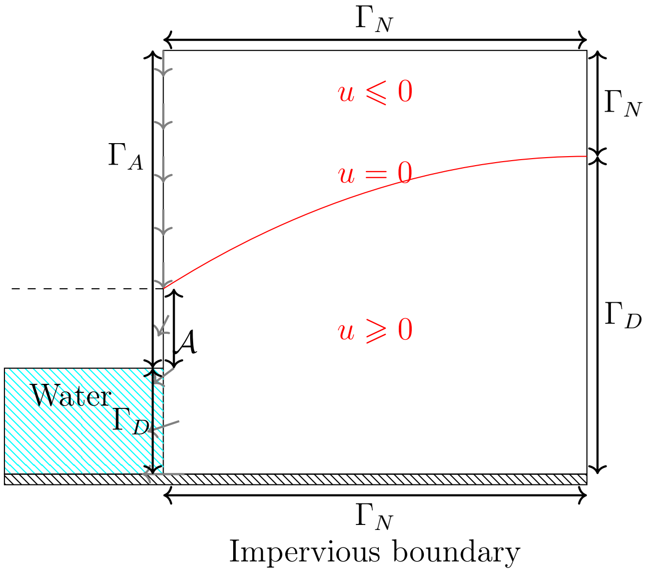


Figure 4.1: A typical seepage problem. The upper part of the left lateral boundary is in contact with the atmosphere, while the lower part is underwater. The height at which the level set $u = 0$ meets the boundary (marked with a dashed line) is a key unknown in seepage problems.

$$\mathcal{V}^g = \{v \in H^1(\Omega) \mid v = g \text{ on } \Gamma_D\} \tag{4.12}$$

$$\mathcal{K}^g = \{v \in \mathcal{V}^g \mid v \leqslant 0 \text{ on } \Gamma_A\} \tag{4.13}$$

$$\mathcal{V}^0 = \{v \in H^1(\Omega) \mid v = 0 \text{ on } \Gamma_D\} \tag{4.14}$$

$$\mathcal{K}^0 = \{v \in \mathcal{V}^0 \mid v \leqslant 0 \text{ on } \Gamma_A\}, \tag{4.15}$$

where boundary values are to be understood in the trace sense.

We seek a weak solution $u \in \mathcal{K}^g$ satisfying (4.7) - (4.10). To that end, multiplying (4.7) by a test function $v \in \mathcal{K}^0$ and integrating by parts, taking into account (4.8) gives

$$\langle \mathbf{q}(u), \boldsymbol{n}\, v\rangle_{\Gamma_A} - \langle \mathbf{q}(u), \nabla v\rangle = \langle f, v\rangle \quad \forall v \in \mathcal{K}^0. \tag{4.16}$$

By the boundary conditions and the definition of the space $\mathcal{K}^0$, the boundary integral is negative so that (4.16) can be written as:

$$\langle -\mathbf{q}(u), \nabla v\rangle \geqslant \langle f, v\rangle \quad \forall v \in \mathcal{K}^0. \tag{4.17}$$

We now extend the boundary data g to a function $\bar{g} \in \mathcal{K}^g$ by imposing that $\bar{g} \equiv 0$ on Γ_A. We will address the choice of function $\bar{g}$ in Remark 4.4.1 but for now it is sufficient to assume such a choice with this property exists. We may therefore set $v = u - \bar{g} \in \mathcal{K}^0$ in (4.16) to give

$$\langle \mathbf{q}(u), \boldsymbol{n}\,(u-\bar{g})\rangle_{\Gamma_A} - \langle \mathbf{q}(u), \nabla(u-\bar{g})\rangle = \langle f, u-\bar{g}\rangle\,. \tag{4.18}$$

Note that by (4.10) and the fact that $\bar{g}$ vanishes on Γ_A, the boundary contribution of (4.18) is zero. This result can be subtracted from (4.17) to obtain the variational inequality in the standard and more compact form for such problems. The problem is then to seek $u \in \mathcal{K}^g$ such that

$$\langle -\mathbf{q}(u), \nabla(v+\bar{g}-u)\rangle \geqslant \langle f, v+\bar{g}-u\rangle \quad \forall v \in \mathcal{K}^0. \tag{4.19}$$

In the seminal paper [68], existence and uniqueness of solutions is proved for problem (4.19) in the case where $k(u) \equiv 1$, see also [64]. This is extendable to monotone nonlinear operators, however note the coefficient k that parametrises the soil properties is often such that the operator does not satisfy this assumption as can be seen by the sharp gradients in Figure 4.2, although the coefficients can be regularised to mitigate this [12]. We will explore this idea further in chapter 6.

In the case $k(u) \equiv 1$, the obstacle problem on a convex domain where $\Gamma_A = \partial\Omega$ is studied in [21] (see also §3) and the regularity result $u \in H^2(\Omega)$ is established. To the authors' knowledge, no such result is available for van

Genuchten type nonlinearities, and in practical situations the nature of the domain and boundary conditions mean this level of regularity is unlikely. Indeed, our numerical results indicate this cannot be the case as the problem lacks regularity around the boundary of the contact set, shown in figure 4.1 as the boundary between $\mathcal{A}$ and $\Gamma_A \backslash \mathcal{A}$. This is precisely the form of the singularity that occurs in the dual problems of §3.7, which suggests regularity is not guaranteed beyond $\mathrm{W}^{2,(4-\varepsilon)/3}(\Omega)$.

4.4 Finite Element Method

In this section, we introduce a finite element method to discretise (4.19). Let us assume that the domain Ω is polyhedral and introduce a shape regular triangulation $\mathscr{T}$. We assume that Γ_A aligns with the mesh in the sense that for all $K \in \mathscr{T}$, $\partial K \cap \partial\Omega$ is either fully contained in Γ_A or else intersects Γ_A in at most one point ($N = 2$) or one edge ($N = 3$). We make a similar assumption on elements lying on Γ_D. For this choice of $\mathscr{T}$ define the space

$$\mathcal{V}_h^g = \{v \in \mathcal{V}^g \mid v \in P(K)\, K \in \mathscr{T}\} \tag{4.20}$$

and the subset

$$\mathcal{K}_h^g = \{v \in \mathcal{V}_h^g \mid v \leqslant 0 \text{ on } \Gamma_A\}. \tag{4.21}$$

Here we make use of the fact that g is affine so that $\mathcal{V}_h^g$ is a subset of $\mathcal{V}^g$.

Remark 4.4.1 (Choice of the function $\bar{g}$)**.** *Now we are in a position to describe the construction of an appropriate extension $\bar{g}$ of g. We define the space*

$$\mathcal{V}^{g,0} = \{v \in \mathcal{V}^g \mid v = 0 \;\textit{on}\; \Gamma_A\} \tag{4.22}$$

and corresponding finite element space

$$\mathcal{V}_h^{g,0} := \mathcal{V}_h^g \cap \mathcal{V}^{g,0} \tag{4.23}$$

and let $\bar{g}$ be the solution to the following finite element problem: Find $\bar{g} \in \mathcal{V}_h^{g,0}$

$$\langle \nabla \bar{g}, \nabla \Phi \rangle = 0 \quad \forall \Phi \in \mathcal{V}_h^{0,0} \tag{4.24}$$

$\bar{g}$ therefore has H^1 regularity over Ω, satisfies the boundary condition on Γ_D in the trace sense, and vanishes on Γ_A. We remark that this ensures also $\bar{g} \in \mathcal{K}^g$. In the following sections as an abuse of notation, we will identify g with $\bar{g}$ to simplify the exposition.

We are now ready to state the finite element approximation to this problem. We seek $U \in \mathcal{K}_h^g$ such that

$$\langle -\mathbf{q}(U), \nabla(\Phi + g - U) \rangle \geqslant \langle f, \Phi + g - U \rangle \quad \forall \Phi \in \mathcal{K}_h^0. \tag{4.25}$$

4.5 Automated error control

In this section we describe the derivation of an error indicator for the problem (4.7) - (4.10). In doing so we make use of a dual problem that is related to the linearised adjoint problem commonly used for nonlinear problems, but we keep only the zeroth order component of the linearisation. We then proceed in a similar manner to [17], where the authors consider a linear problem, to obtain a bound for the error in the quantity of interest.

4.5.1 Definition of Dual Problem

The definition of the dual problem is interwoven with the primal solution u as well as the finite element approximation U. To begin, we recall that the discrete contact set is defined as:

$$\mathcal{A}_U := \{x \in \Gamma_A \mid U(x) = 0\}. \tag{4.26}$$

We let

$$\mathcal{G} = \{v \in V \mid v \leqslant 0 \text{ on } \mathcal{A}_U \text{ and } \int_{\Gamma_A} -\mathbf{q}(u)(v+U)\cdot \boldsymbol{n}\,\mathrm{d}S \leqslant 0\}, \tag{4.27}$$

and suppose J is a form whose precise structure will be discussed later, and let $z \in \mathcal{G}$ be the solution to the following variational inequality:

$$\langle k(u)\nabla(\varphi - z), \nabla z\rangle \geqslant J(\varphi - z) \quad \forall \varphi \in \mathcal{G}. \tag{4.28}$$

Application of duality arguments to derive error bounds in non-energy norms require assumptions of well-posedness on the dual problem which may not hold. Sharp regularity bounds on the dual problem with $k(u) \equiv 1$ were only recently proven in [34] (see also §3) by a non-standard choice of dual problem. This motivates us to make the following assumption which we will use in the a posteriori analysis, the proof of which is currently the topic of ongoing research.

Assumption 4.5.1 (Convergence in L^2)**.** *With u solving (4.7) - (4.10) and U as defined in (4.25), there are constants $C > 0$ and $s > 1$ such that*

$$\|u - U\|_{L^2(\Omega)} \leqslant Ch^s. \tag{4.29}$$

Remark 4.5.2 (Assumption 4.5.1)**.** *We are motivated to make this assumption by the results of chapters 2, 3 where we saw two sided a posteriori bounds on the approximation error of a similar problem in* non-energy *norms. In addition, a priori convergence results in $\mathrm{L}^4(\Omega)$ for approximation of the Signorini problem are given in [***hristof1754finite***].*

Definition 4.5.3 (Unrestricted solution)**.** *We define a function $\tilde{U}$ to be the solution of the elliptic problem analogous to problem (4.7)-(4.9) but without the inequality constraint (4.10). That is, $\tilde{U} \in \mathcal{V}^g$ satisfies*

$$\left\langle -\mathbf{q}(\tilde{U}), \nabla w \right\rangle = \langle f, w\rangle \quad \forall w \in \mathcal{V}^0. \tag{4.30}$$

The omission of a boundary term in the weak form indicates that $\tilde{U}$ satisfies $\mathbf{q}(\tilde{U}) \cdot \mathbf{n} = 0$ on Γ_A.

4.5.2 Error Bound

Observe that by construction the function $z + u - U$ is a member of the set $\mathcal{G}$. Indeed, by (4.10) we have $u \leqslant 0$ on $\mathcal{A}_U$, by definition of $\mathcal{A}_U$ and $\mathcal{G}$ respectively we have $U = 0$ and $z \leqslant 0$ on $\mathcal{A}_U$. Further, we calculate

$$\int_{\Gamma_A} -\boldsymbol{q}(u)((z + u - U) + U) \cdot \boldsymbol{n} \, \mathrm{d}S = \int_{\Gamma_A} -\boldsymbol{q}(u) z \cdot \boldsymbol{n} \, \mathrm{d}S \leqslant 0, \tag{4.31}$$

since $\boldsymbol{q}(u)u \cdot \boldsymbol{n} = 0$ on Γ_A by the boundary conditions (4.10), and because $z \in \mathcal{G}$. We may therefore take $\varphi = z + u - U$ in (4.28) to obtain

$$J(u - U) \leqslant \langle k(u)\nabla(u - U), \nabla z \rangle \, . \tag{4.32}$$

Writing

$$\langle k(u)\nabla(u - U), \nabla z \rangle = \langle \mathbf{q}(U) - \mathbf{q}(u), \nabla z \rangle - \langle (k(u) - k(U))\nabla(U + h_z), \nabla z \rangle \, , \tag{4.33}$$

and expanding

$$k(u) - k(U) = \int_0^1 k'(U + s(u - U))(u - U) \, \mathrm{d}s, \tag{4.34}$$

we note that with the a priori assumption 4.5.1, we can assume that the second term on the right hand side of (4.33) is higher order in the error $u - U$ than the first term, and can therefore be neglected when the computation error becomes small. We will therefore focus on the first term in the following analysis.

In the following lemmata, we prove bounds on differences between the functions u, U and $\tilde{U}$.

Lemma 4.5.4 (Properties of the unrestricted solution)**.** *With u the primal solution defined through (4.17), U the finite element approximation to u given by (4.25), and $\tilde{U}$ the unrestricted solution defined in* (4.30)*, we have, for any $v \in \mathcal{K}^0$ and $\Phi \in \mathcal{K}_h^0$,*

$$\left\langle \mathbf{q}(u) - \mathbf{q}(\tilde{U}), \nabla(v + g - u) \right\rangle \leqslant 0 \quad \forall v \in \mathcal{K}^0 \tag{4.35}$$

and

$$\left\langle \mathbf{q}(U) - \mathbf{q}(\tilde{U}), \nabla(\Phi + g - U) \right\rangle \leqslant 0 \quad \forall \Phi \in \mathcal{K}_h^0. \tag{4.36}$$

Proof. We choose test functions $w = v+g-u$ and $w = \Phi+g-U$ respectively in (4.30) where $v \in \mathcal{K}^0$ and $\Phi \in K_h^0$ are arbitrary to see that

$$\langle -\mathbf{q}(U), \nabla(v + g - u) \rangle = \langle f, v + g - u \rangle \quad \forall v \in \mathcal{K}^0 \tag{4.37}$$

and

$$\left\langle -\mathbf{q}(\tilde{U}), \nabla(\Phi + g - U) \right\rangle = \langle f, \Phi + g - U \rangle \quad \forall \Phi \in \mathcal{K}_h^0. \tag{4.38}$$

Subtracting (4.19) from (4.37) and (4.25) from (4.38), we arrive at the desired result.

□

Definition 4.5.5 (Restricted solution set)**.** *We define the set*

$$\mathcal{W}_h^g = \{v \in \mathcal{V}_h^g \mid v \leqslant 0 \ \text{ on } \ \mathcal{A}_U\}. \tag{4.39}$$

Note that $\mathcal{W}_h^g$ is a slightly larger set than $\mathcal{K}_h^g$, but that $U \in \mathcal{W}_h^g$. This means that U in fact satisfies

$$\left\langle \mathbf{q}(U) - \mathbf{q}(\tilde{U}), \nabla(\Phi + g - U) \right\rangle \leqslant 0 \quad \forall \Phi \in \mathcal{W}_h^0. \tag{4.40}$$

Lemma 4.5.6 (Galerkin orthogonality)**.** *With u the primal solution defined through (4.17) and U the finite element approximation to u given by (4.25) we have*

$$\langle \mathbf{q}(U) - \mathbf{q}(u), \nabla z_h \rangle \leqslant \left\langle \mathbf{q}(\tilde{U}) - \mathbf{q}(u), \nabla(z_h + U - u) \right\rangle \quad \forall z_h \in \mathcal{W}_h^0, \tag{4.41}$$

in analogy to the usual Galerkin orthogonality result.

Proof. We can write

$$\begin{aligned} \langle \mathbf{q}(U) - \mathbf{q}(u), \nabla z_h \rangle = &\left\langle \mathbf{q}(\tilde{U}) - \mathbf{q}(u), \nabla(z_h + U - u) \right\rangle \\ &+ \left\langle \mathbf{q}(U) - \mathbf{q}(\tilde{U}), \nabla z_h \right\rangle \\ &+ \left\langle \mathbf{q}(\tilde{U}) - \mathbf{q}(u), \nabla(u - U) \right\rangle . \end{aligned} \tag{4.42}$$

Now suppose $z_h \in \mathcal{W}_h^0$. By setting $\Phi = U + z_h - g$ in (4.36), the second term on the right hand side of (4.42) is negative. Similarly, choosing $v = U - g$ in (4.35), the final term is also negative, and the result follows. □

Lemma 4.5.7 (Property of the dual solution)**.** *Let u be the primal solution defined through (4.17), z be the dual solution from (4.28) and U the finite element approximation to u given by (4.25). Then, we have*

$$\left\langle \mathbf{q}(\tilde{U}) - \mathbf{q}(u), \nabla(z + U - u) \right\rangle \leqslant 0. \tag{4.43}$$

Proof. By the definition of $\tilde{U}$ we have

$$\left\langle -\mathbf{q}(\tilde{U}), \nabla(z + U - u) \right\rangle = \langle f, z + U - u \rangle \tag{4.44}$$

and by (4.16),

$$\langle -\mathbf{q}(u), \nabla(z + U - u) \rangle = \langle f, z + U - u \rangle - \langle \mathbf{q}(u), \boldsymbol{n}(z + U - u) \rangle_{\Gamma_A} , \tag{4.45}$$

and therefore, noting that $(k(u)\nabla u)u = 0$ on Γ_A,

$$\begin{aligned}\left\langle \mathbf{q}(\tilde{U}) - \mathbf{q}(u), \nabla(z + U - u) \right\rangle &= \int_{\Gamma_A} -\mathbf{q}(u) \cdot \boldsymbol{n}(z + U - u)\,\mathrm{d}S \\ &= \int_{\Gamma_A} -\mathbf{q}(u) \cdot \boldsymbol{n}(z + U)\,\mathrm{d}S \leqslant 0,\end{aligned} \tag{4.46}$$

by the definition of the space $\mathcal{G}$.

□

We now state the main result of this section.

Theorem 4.5.8 (Error bound)**.** *Let u be the solution of (4.19) and U the finite element approximation to u. Let $\tilde{U}$ be the solution of the unrestricted problem (4.30), z the dual solution of (4.28) and $z_h \in W_h^0$ an arbitrary function. Then to leading order, we have*

$$J(u - U) \lesssim \left\langle \mathbf{q}(U) - \mathbf{q}(\tilde{U}), \nabla(z - z_h) \right\rangle. \tag{4.47}$$

Proof. Starting from (4.33) and neglecting the higher order term, justified by Assumption 4.5.1,

$$\begin{aligned} J(u - U) &\leqslant \langle \mathbf{q}(U) - \mathbf{q}(u), \nabla z \rangle \\ &= \langle \mathbf{q}(U) - \mathbf{q}(u), \nabla(z - z_h) \rangle + \langle \mathbf{q}(U) - \mathbf{q}(u), \nabla z_h \rangle.\end{aligned} \tag{4.48}$$

Combining with Lemma (4.5.6) gives

$$
\begin{aligned}
&\langle \mathbf{q}(U)-\mathbf{q}(u), \nabla(z-z_h)\rangle + \langle \mathbf{q}(U)-\mathbf{q}(u), \nabla z_h\rangle \\
&\leqslant \langle \mathbf{q}(U)-\mathbf{q}(u), \nabla(z-z_h)\rangle + \left\langle \mathbf{q}(\tilde{U})-\mathbf{q}(u), \nabla(z_h+U-u)\right\rangle \\
&= \langle \mathbf{q}(U)-\mathbf{q}(u), \nabla(z-z_h)\rangle + \left\langle \mathbf{q}(\tilde{U})-\mathbf{q}(u), \nabla(z+U-u)\right\rangle \\
&\quad + \left\langle \mathbf{q}(\tilde{U})-\mathbf{q}(u), \nabla(z_h-z)\right\rangle \\
&= \left\langle \mathbf{q}(U)-\mathbf{q}(\tilde{U}), \nabla(z-z_h)\right\rangle + \left\langle \mathbf{q}(\tilde{U})-\mathbf{q}(u), \nabla(z+U-u)\right\rangle,
\end{aligned}
\tag{4.49}
$$

upon rearranging. The second term is negative by Lemma 4.5.7, completing the proof. □

To illustrate the usefulness of this result, we state the following corollary to theorem 4.5.8.

Corollary 4.5.9 (A posteriori error indicator)**.** *With the notation of theorem 4.5.8, we have the local error estimate*

$$
J(u-U) \leqslant \sum_{K\in\mathscr{T}} \langle f-\nabla\cdot\mathbf{q}(U), z-z_h\rangle_K + \langle \mathbf{q}(U)\cdot\boldsymbol{n}, z-z_h\rangle_{\partial K}\,. \tag{4.50}
$$

Proof. Since $\tilde{U}$ solves (4.30), we can replace it in the right hand side of (4.47) and introduce the problem data:

$$
\left\langle \mathbf{q}(U)-\mathbf{q}(\tilde{U}), \nabla(z-z_h)\right\rangle = \langle f, z-z_h\rangle + \langle \mathbf{q}(U)\cdot\boldsymbol{n}, \nabla(z-z_h)\rangle\,. \tag{4.51}
$$

After integrating by parts over each element we obtain the stated result. □

Equation (4.50) gives a local quantity that we can approximately evaluate to give an estimate of the local numerical error. Given a suitable approximation of the dual error $z-z_h$, this quantity can be computed and used to inform adaptive mesh refinement. The approximate computation of the error estimate will be addressed in section 4.6.3.

Remark 4.5.10. *The analysis above allows the choice of J to be made by the user depending on the problem at hand. The resulting estimate used in an adaptive algorithm will prioritise the accurate computation of J. For example,*

1. *Fix $x_0 \in \Omega$ and set $J_1(\varphi) = \varphi(x_0)$ for all φ lying in a suitable test space. An adaptive routine based upon the resulting estimate would prioritise accurate computation of the point value of the solution at x_0.*

2. *Setting $J_2(\varphi) = \langle u - U, \varphi\rangle$ for all φ lying in a suitable test space would give an estimate of the error in the global error in $L^2(\Omega)$. Using suitable approximations, such an approach can be used in practice, see section 4 of [14].*

3. *In seepage problems, a common quantity of interest is the volumetric flow rate of water through the seepage face. Since by definition the soil is saturated along the seepage face, the hydraulic conductivity takes the constant value K_s (see section 4.6). The fluid velocity is given by (4.2) and therefore the volumetric flow rate is given by*

$$J(u) := -\int_{\Gamma_A} \frac{K_s}{\phi} \nabla(u + h_z) \cdot \boldsymbol{n} \, \mathrm{d}S = \int_{\Gamma_A} \frac{\mathbf{q}(u)}{\phi} \cdot \boldsymbol{n} \, \mathrm{d}S, \tag{4.52}$$

   where we recall that ϕ is the porosity of the soil.

4.6 Implementation Details

In this section we discuss various aspects of the practical solution of problem (4.7) - (4.10). We first discuss the choice of parametrisation of k in (4.7), then present the iterative numerical algorithm used to solve the nonlinear problem. Finally, we discuss aspects of the adaptive routine and the tools required to approximately evaluate the error estimate.

4.6.1 Hydrogeological Properties of the Medium

We make use of the popular model of [75] and [106] to parametrise the unsaturated hydraulic properties of the soil. Consider a volume V of a porous medium of total volume V_{total}. V is made up of the solid matrix and air- or fluid-filled pores. If V_{water} is the total volume of water contained in V, the volumetric water content θ is V_{water}/V_{total}, and therefore takes values between 0 and the porosity of the soil. Point values of water content can be defined in the usual way of taking the water content over a representative elementary volume around the point (we refer to section 1.3 of [13] for details). Water content is related to the pressure head in the soil, and can be modelled as a nonlinear function $\theta(u)$. The dimensionless water content Θ was defined by van Genuchten [106] as

$$\Theta(u) = \frac{\theta(u) - \theta_R}{\theta_S - \theta_R}, \tag{4.53}$$

where θ_R and θ_S are respectively the minimum and maximum volumetric water contents supported by a soil. Then the normalised water content Θ takes values between 0 and 1 with 1 corresponding to saturation. Hydraulic conductivity, that is the nonlinear coefficient k in (4.7) is modelled similarly, and takes strictly positive values reaching its maximum value at saturation. The shapes of the functions k and Θ are dictated by choice of dimensional parameters K_S and α, and non-dimensional parameter n. The units are $[K_S] = ms^{-1}$ and $[\alpha] = m^{-1}$. The soil parameters are often fitted following laboratory experiments for a given soil. The saturated hydraulic conductivity K_S is the maximum value that k can take. Finally, α and n are shape parameters whose physical meaning is less clear. The parameter m, introduced for ease of presentation, is defined by $m = (n-1)/n$. This model has been shown to give good predictions in most soils near saturation by [107].

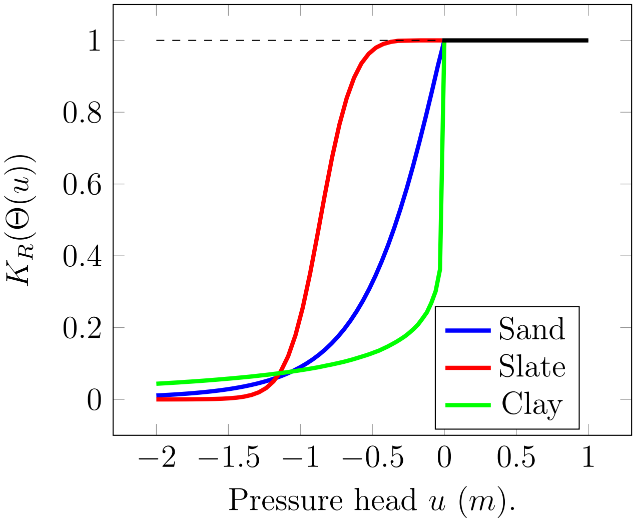


Figure 4.2: The permeability coefficient as a function of pressure head u for different soil types. Note that $k(u) \to 0$ as $u \to -\infty$ but $K_R > 0$ for all u. Further, observe the smoothness of K_R is quite different at $u = 0$ for different soil types. This lack of regularity makes the numerical simulation of, say clay, particularly challenging. We also note that these functions are scaled by the saturated hydraulic conductivity, K_S, which varies enormously between different soils. The mean value for different soil types is $5 \times 10^{-6} ms^{-1}$ (sand), $5 \times 10^{-9} ms^{-1}$ (slate) and $1 \times 10^{-8} ms^{-1}$ (clay).

$$\Theta(u) = \begin{cases} \frac{1}{[1+(-\alpha u)^n]^m} & u < 0 \\ 1 & u \geqslant 0 \end{cases} \tag{4.54}$$

$$K_R(\Theta(u)) = \begin{cases} \Theta(u)^{\frac{1}{2}} \left[1 - \left(1 - \Theta(u)^{\frac{1}{m}}\right)^m\right]^2 & u < 0 \\ 1 & u \geqslant 0 \end{cases} \tag{4.55}$$

from which k is then obtained by scaling by the saturated hydraulic conductivity:

$$k(u) = K_S \, K_R(\Theta(u)). \tag{4.56}$$

Examples of hydraulic behaviour of different soils are shown in figure 4.2. The smoothness of the function K_R as it approaches saturation is largely determined by the parameter n, with larger n resulting in a smoother transition from unsaturated to saturated soil.

4.6.2 Solution Methods

To solve the nonlinear problem, we use a Picard iterative technique, common in the literature for computations in variably saturated flow [84]. As described in [94], we choose to implement the seepage face boundary condition using a type of active set strategy in a way that allows it to be updated within the Picard iteration during the solution process of the PDE. This has clear benefits for the accurate resolution of the seepage face, and it is especially important in the adaptive framework that the exit point be allowed to move to take advantage of increasing resolution during the adaptive process. A practical way of achieving this within the nonlinear iteration was first presented in [76], but its focus on representing a single seepage face in an a priori assumed part of the boundary limits the range of applicability. The procedure was generalised in [38] to allow any number of seepage faces by checking for unphysical behaviour at boundary nodes. This is essentially the method used here, but assignment is element-wise. Pressure and flux is checked along boundary faces which are then assigned as being on the seepage face or not, determining the boundary condition to be enforced at the next iteration. It was observed that this approach resulted in less oscillation of the exit point through the iterative process. This process can be thought of as a physically motivated version of a projection method for solving variational inequalities, as described in section 2 of [83]. The algorithm is illustrated below (see Algorithm 1).

The nonlinear iteration is controlled by monitoring the difference in $\mathrm{L}^2(\Omega)$-norm between successive iterates normalised by the norm of the newest iterate. Since we are concerned with the error in the finite element approxima-

Algorithm 1 An Iterative Scheme for the Seepage Problem

```
Require: $u^0$, TOL, $N$
Ensure: $U$, the approximation to the solution of the variational inequality
Set $i = 1$;
while $i < N$ do
    Set $\mathcal{A}_U := \{x \in \Gamma_A \mid U^{i-1}(x) = 0\}$;
    for degrees of freedom, $x_q$, over $\Gamma_A$ do
        if $U^{i-1}(x_q) > 0$ and $x_q \notin \mathcal{A}_U$ then
            Constrain $U^i(x_q) = 0$;
        else if $(\mathbf{q}(U^i) \cdot \mathbf{n})(x_q) < 0$ and $x_q \in \mathcal{A}_U$ then
            Constrain $(\mathbf{q}(U^i) \cdot \mathbf{n})(x_q) = 0$;
        else
            Leave boundary conditions unchanged;
    Find $U^i$ such that: $\int_\Omega k(U^{i-1})\nabla(U^i + h_z) \cdot \nabla\Phi = \int_\Omega f\Phi$ for all $\Phi$ over a space with boundary conditions as above;
    if $e := \|U^i - U^{i-1}\|_{L^2(\Omega)} <$ TOL then
        Set $U := U^i$;
        Break;
    $i$++;
```

tion, a very small iteration tolerance is set to ensure that the nonlinear error is small compared to discretisation error. The iteration registers a failure if this tolerance is not met within 30 steps, but in practice no convergence failures occurred.

4.6.3 Adaptive Algorithm

As described in section 2.7, we adapt the mesh using SOLVE→ ESTIMATE → MARK → REFINE. Cells are marked for refinement using Dörfler marking. It remains to describe the approximate evaluation of the error estimate.

Evaluating the Estimate

Recall the error estimate of corollary 4.5.9:

$$\eta = \sum_{K \in \mathscr{T}} \eta_K, \tag{4.57}$$

where

$$\eta_K = \langle f - \nabla \cdot \mathbf{q}(U), z - z_h \rangle_K + \frac{1}{2} \langle [\![\mathbf{q}(U)]\!] \,, z - z_h \rangle_{\partial K} . \tag{4.58}$$

Note that η_K can only be approximately calculated since the exact dual solution z is not available. There are several strategies for doing this which produce similar results [8]. For computational efficiency, we choose a cheap averaging interpolation to obtain a higher order approximation of the dual solution as follows.

The dual problem is solved on the same finite element space as the primal problem to obtain an approximation z_h. A function $\bar{z}_h$ is then constructed from z_h in the following manner. Consider the mesh $\bar{\mathscr{T}}^l$ such that refining every element of $\bar{\mathscr{T}}^l$ produces $\mathscr{T}^l$. In the simplest case of uniform meshes, each element of this new mesh corresponds to four elements of the original mesh, with nine nodal values of z_h. These values are sufficient to define a biquadratic finite element function $\bar{z}_h$ on the mesh $\bar{\mathscr{T}}^l$. We note that z_h and $\bar{z}_h$ coincide at degrees of freedom of the original mesh, but differ away from them. We also emphasise that this computation is possible on locally refined grids. Similar techniques are sometimes used as post-processors to improve the quality of finite element approximations [40]. Since both z_h and $\bar{z}_h$ are piecewise polynomial, we can integrate their difference exactly, and we make the approximation

$$\eta_K \approx \langle f - \nabla \cdot \mathbf{q}(U), \bar{z}_h - z_h \rangle_K + \frac{1}{2} \langle [\![\mathbf{q}(U)]\!], \bar{z}_h - z_h \rangle_{\partial K} . \tag{4.59}$$

Remark 4.6.1. *We remark here that there are alternate ways to compute an approximation $\bar{z}_h$. One could compute on the same mesh but with piecewise quadratic finite elements, however this is significantly more expensive than the solving the primal problem whereas computational complexity is comparable for our approach.* If *the dual solution is smooth, the coarse mesh approximation is better as increasing polynomial degree gives higher order approximability in light of theorem 2.5.6. The different options availble are discussed in section 4.1 of [8].*

Remark 4.6.2 (Approximation of the space $\mathcal{G}$)**.** *We finally remark that in the practical implementation, we must solve the dual problem in the set $\mathcal{W}_h^g$ which may or may not be a subset of $\mathcal{G}$. This is due to the fact that the exact contact set is not available, and so we do not have access to $\mathcal{G}$. In fact, the authors of [17] further suggest approximating $\mathcal{G}$ by $\mathcal{G}^0 := \{v \in \mathcal{V}^0 \mid v = 0 \text{ on } \mathcal{A}_U\}$, and we also take this approach. This reduces the dual problem to a linear elliptic PDE, thereby simplifying the adaptive process.*

4.7 Numerical Benchmarking

In this section, we present numerical results to demonstrate the effectiveness of the error estimate and adaptive routine in a range of realistic situations of interest in the analysis of subsurface flow. In this sense, we aim to benchmark our work to justify its use in the next section where we tackle specific case studies.

All simulations presented here are conducted using `deal.II`, an open source C++ software library providing tools for adaptive finite element computations [5]. These were run on a Viglen Genie desktop computer with an Intel i7 processor and 16Gb RAM. All simulations were completed within an hour on this hardware. A high order quadrature formula (order 8) is used in the assembly of the finite element system for each linear solve to attempt to capture some of the variation in the coefficients. To avoid any possible issues with convergence of linear algebra routines, an exact solver, provided by UMFPACK, is used to invert the system matrix. This software is an implementation of multifrontal LU factorisation.

In all simulations we take as our quantity of interest the volumetric flow rate of water through the seepage face given in equation (4.52).

4.7.1 Example 1: Aquifer Feeding a Well

As a first two-dimensional example, let $\Omega = (0,1)^2$ represent a vertical section of a subsurface region. Spatial dimensions are given in metres. We refer to Figure 4.1 for a visual representation of this problem, and give the specifics here. The upper surface $\{(x, h_z) \mid h_z = 1\}$ represents the land surface while $\{(x, h_z) \mid h_z = 0\}$ is impermeable bedrock. In both cases no-flux boundary conditions are enforced. We remark that in certain cases the land surface could exhibit seepage faces, as we will see in Example 2, but we assume that this will not be the case here. On $\{(x, h_z) \mid x = 1\}$, a hydrostatic Dirichlet condition is enforced for the pressure with the water table height set at $0.8m$, that is, we set $u = 0.8 - h_z$ along this portion of the boundary. This corresponds to setting the groundwater table far from the well. Finally, $\{(x, h_z) \mid x = 0\}$ is the inner wall of the well. The well is filled with water up to a fixed level H_w, and a hydrostatic Dirichlet condition for the pressure is applied along the portion of the boundary that is in contact with the body of water. Above H_w, the seepage face boundary conditions apply. For the simulations presented here, we choose $H_w = 0.25m$. We remark that this simple setup and variations of it are common benchmarks for works on seepage problems [83, 38, 62, 118].

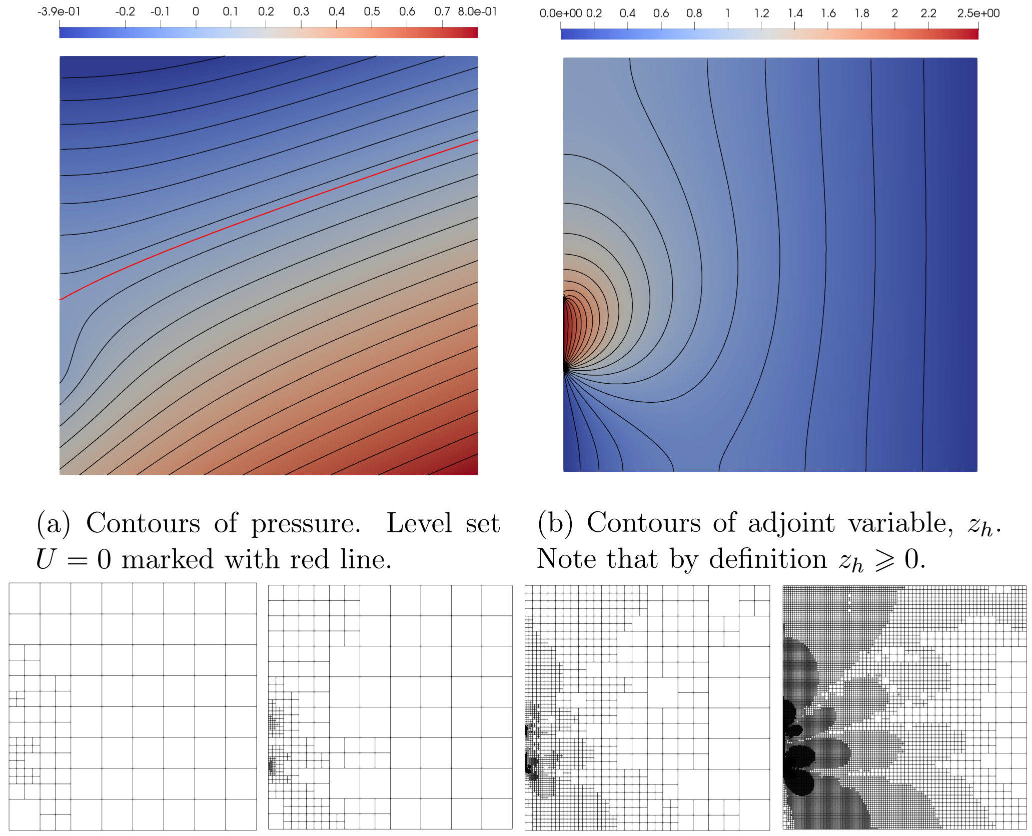
For the soil parametrisation, we make the choices $n = 2.06$, $\alpha = 1m^{-1}$, $K_S = 1ms^{-1}$. This results in a soil that has the characteristics of silt whose hydraulic conductivity has been scaled to have magnitude 1. Since we have taken the datum f to be zero in this example, scaling the diffusion coefficient by a constant has no effect on the pressure head.

Figure 4.3a shows an approximation to the solution of the problem in this case, with the associated adjoint solution in Figure 4.3b. Notice the adjoint solution takes its largest values along the seepage face along which the quantity of interest is evaluated, with values increasing along streamlines that terminate there. This is to be expected as it demonstrates that the flow upstream of the seepage face has the greatest influence upon the quantity of interest.

The simulation is initialised on a coarse mesh of 256 elements and uses the goal-based estimate as refinement criterion. A selection of meshes generated by the adaptive algorithm is given in figure 4.3c–4.3f. The solution is qualitatively comparable to those found for example in [38].

4.7.2 Example 2: Sloping Unconfined Aquifer with Impeding Layer

The second test case is taken from [94]. Its relevance was shown in [90] where the location of impeding layers was shown to have large effects on the saturation conditions of the soil. The domain setup is illustrated in Figure 4.4. This configuration leads to water flowing down the slope due to gravity, and allows multiple seepage faces to form. We introduce a forcing term, representing an underground spring, above the layer to force extra seepage



(a) Contours of pressure. Level set $U = 0$ marked with red line.

(b) Contours of adjoint variable, z_h. Note that by definition $z_h \geqslant 0$.

(c) $\mathscr{T}^1$

(d) $\mathscr{T}^4$

(e) $\mathscr{T}^7$

(f) $\mathscr{T}^{11}$

Figure 4.3: Example 1, flow through a single layered, silty soil. We show the pressure, adjoint solution and a sample of adaptively generated meshes showing refinement upstream of the seepage face. The primal variable U and the adjoint variable z_h are both represented on $\mathscr{T}^{11}$ which has approximately 66000 degrees of freedom.

faces. It is defined by:

$$f(x) = \begin{cases} 10 & \text{if } \operatorname{dist}(x, (9, 1.15)) < 0.2 \\ 0 & \text{otherwise.} \end{cases} \tag{4.60}$$

We make the same choice of soil parameters as in example 1, that is

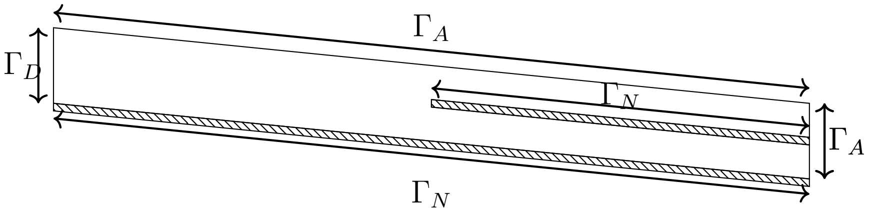


Figure 4.4: The domain models a slope lying on a layer of bedrock with a downstream external boundary. The domain is a parallelogram with corners $(0,1)$, $(0,2)$, $(10,1)$ and $(10,0)$ where all dimensions are in metres. The lower extent of the domain represents an impermeable boundary, as does a layer of rock parallel to the land surface towards the right hand side of the domain. This layer is 0.1m thick with corners $(5,0.95)$, $(5,1.05)$, $(10,0.45)$ and $(10,0.55)$. The water table is fixed with a Dirichlet boundary condition on the left hand boundary of the domain.

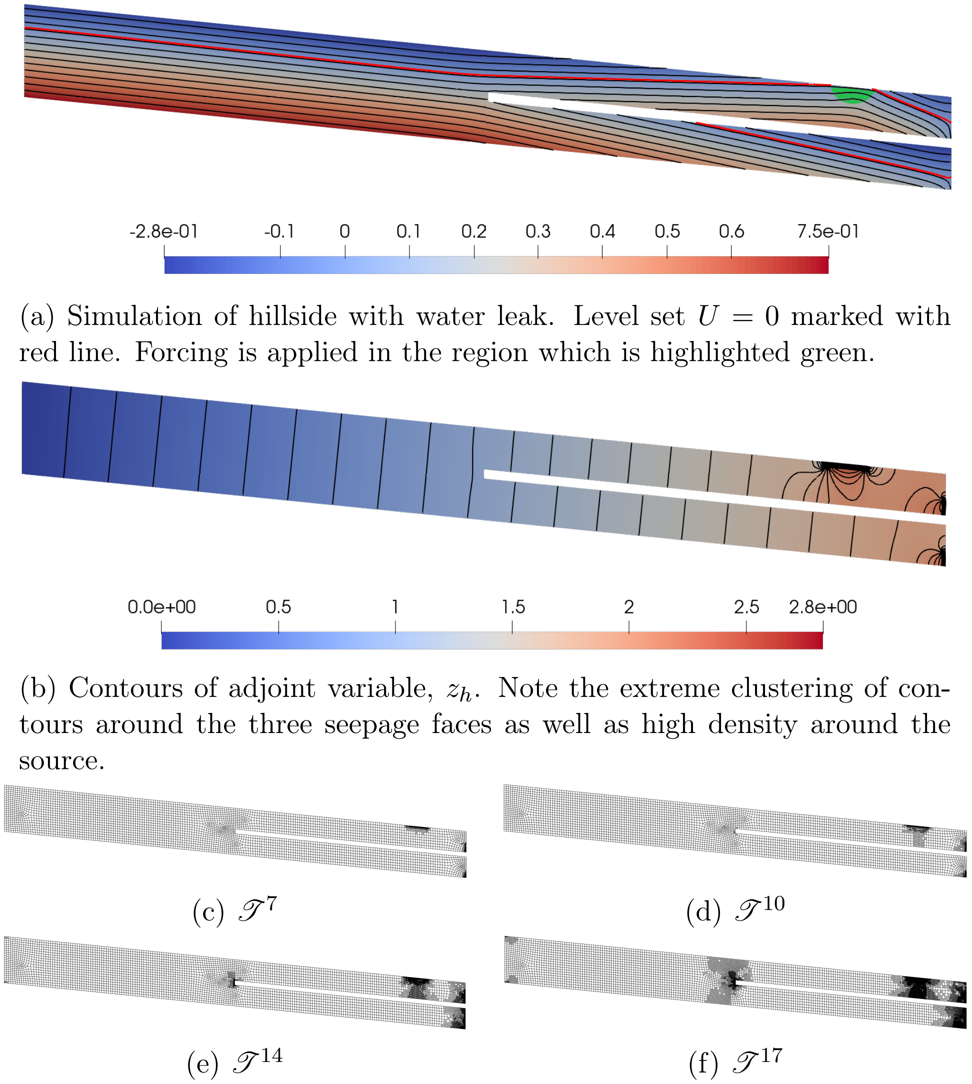
$n = 2.06$, $\alpha = 1m^{-1}$, $K_S = 1ms^{-1}$.

The results of this are given in Figure 4.5. As can be seen in Figure 4.5a, three disjoint seepage faces arise from this simulation, two on the right hand face, one above and one below the impermeable barrier, and another at the land surface. It should be noted that the seepage face at the land surface would generate surface run-off. This process is not taken into account by the model we use.

The simulation is initialised on a coarse mesh of 4036 elements. An adaptive simulation using the dual-weighted estimate produced the meshes in figure 4.5. The algorithm refines heavily around the source and all seepage faces, as well as resolving the corners around the impeding layer.

4.7.3 Estimator Effectivity Summary

In Examples 1 and 2 above we compute a reference value for $J(u)$ obtained from a simulation on a very fine grid. This was taken as the 'true' value to perform analysis of the behaviour of the estimate. In Figures 4.6a–4.6b, we



(a) Simulation of hillside with water leak. Level set $U = 0$ marked with red line. Forcing is applied in the region which is highlighted green.

(b) Contours of adjoint variable, z_h. Note the extreme clustering of contours around the three seepage faces as well as high density around the source.

(c) $\mathscr{T}^7$

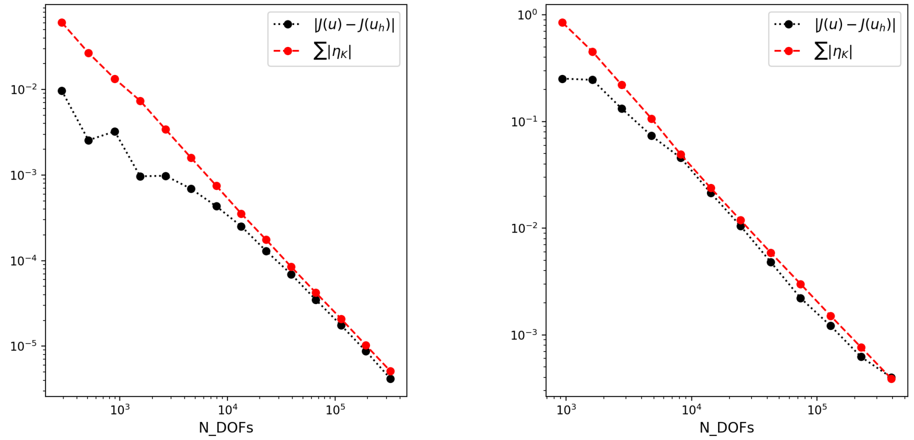
(d) $\mathscr{T}^{10}$

(e) $\mathscr{T}^{14}$

(f) $\mathscr{T}^{17}$

Figure 4.5: Example 2, flow through a sloped aquifer with impeding layer. We show the pressure, adjoint solution and a sample of adaptively refined meshes that capture multiple seepage faces as well as potential singularities in the pressure at the corner in the domain. The primal and adjoint variable are both represented on $\mathscr{T}^{17}$ which has approximately 7×10^5 degrees of freedom.

see that as the simulation progresses the effectivity of the estimate, defined as the ratio of the error to the estimate, becomes very close to 1.



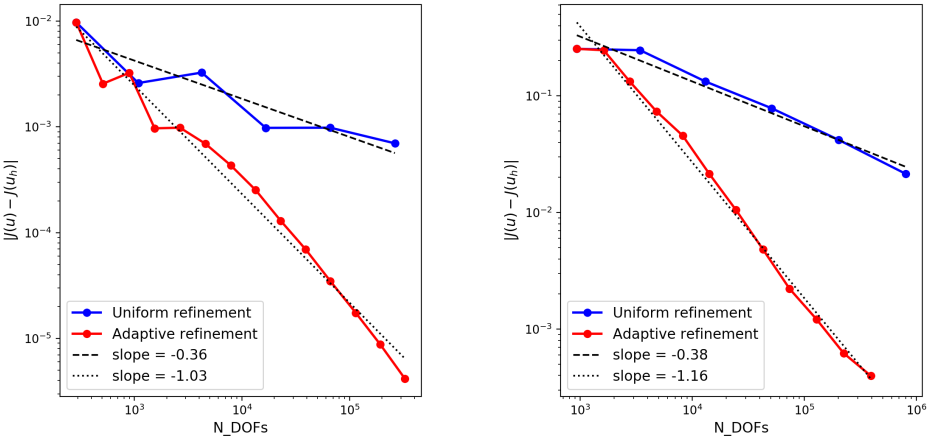
(a) Example 1.

(b) Example 2.

Figure 4.6: Sharpness of error estimates during adaptive mesh refinement. Notice that the dual-weighted estimate significantly over-estimates the error for the first few refinement cycles but as the simulation progresses the effectivity moves closer to one. This is a well known feature of this class of algorithm further described in [78]

4.7.4 Adaptive vs Uniform Comparison

To illustrate the gains obtained through adaptive refinement, we make a comparison between the uniformly refined simulation and the adaptive one. In each case uniform meshes perform extremely poorly with small and unpredictable reductions in error where the adaptive scheme produces fast and monotonic error reduction on all but the coarsest meshes. For comparison, two lines illustrating different rates are included in figure 4.7a illustrating that convergence of $J(U)$ is suboptimal for uniform meshes, and that in terms of degrees of freedom, this optimality can be restored using the goal-based estimate.



(a) Example 1.

(b) Example 2.

Figure 4.7: Comparison of orders of convergence in terms of number of degrees of freedom (N_{DOFS}) on uniform and adaptive grids. Notice the rate of error reduction is considerably slower for uniform simulations in all cases.

4.8 Case Studies with Layered Inhomogeneities

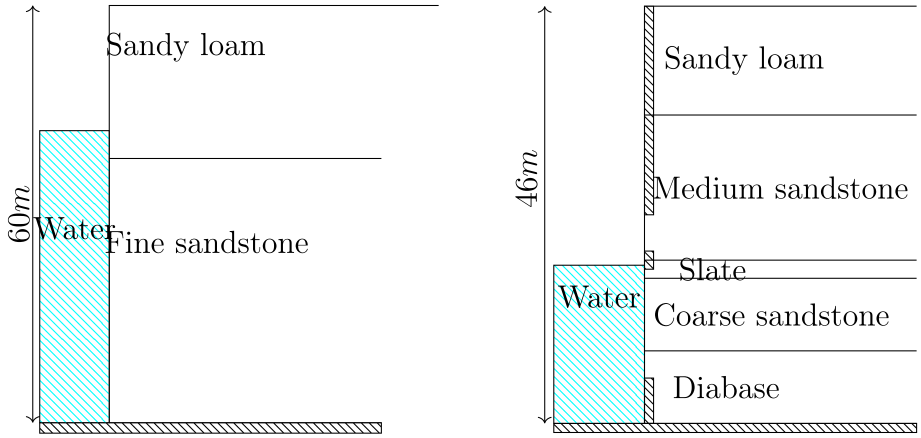
We present results making use of borehole data provided by CPRM (Brazilian Geological Survey) by the Siagas system [1]. The wells are used to supply water to two different cities in São Paulo State, Brazil, one in Ibirá and the other in Porto Ferrreira. Both cities are located over the Paraná Sedimentary Basin, but in places with different shallow geology. There are two different problem setups that we consider. In both cases the domain is a vertical section illustrated in Figure 4.8. We assumed the soil is in homogeneous layers, where there is no variation in the physical properties in the horizontal direction. The soil parameters used for the simulation are given in Table 4.1. The water table height far from the well is known and applied as a Dirichlet boundary condition for the pressure on the right hand lateral boundary. In both cases, the height of the water in the well gives the left lateral boundary condition, and water is continually pumped out of the well in such a way that the water height remains constant.

We work in cylindrical coordinates with the (r, ϕ, h_z) with the h_z-axis aligned with the centre of the well. The aim is to calculate the total flux into the well. We therefore use the functional J_2 to account for flux of water over the inner boundary below the water level, defined as follows.

$$J_2(u) := 2\pi r_0 \int_{r=r_0} \mathbf{q}(\mathbf{u}) \cdot \boldsymbol{n} \, \mathrm{d}h_z, \tag{4.61}$$

where r_0 denotes the radius of the well, that is, we integrate over the entire inner wall of the well, above and below the water.

[1] http://siagasweb.cprm.gov.br/layout/



(a) Case study 1, well within a two layered soil.

(b) Case study 2, well within a five layered soil.

Figure 4.8: Geometric setup of the industrial case study problems. Black shading represents an impermeable boundary. In case study two, the gaps between impermeable regions on the inner wall of the well are the filter locations. The far field boundary conditions are analogous to those given in Figure 4.1 and water is continually pumped out to maintain constant water height.

4.8.1 Case Study 1 - 2 layered well in Ibirá (CPRM reference 3500023601)

For these case studies, all lengths are given in metres. In the first case we set $\Omega = \{(r, \phi, h_z) \mid 0.0762 \leqslant r \leqslant 50,\ 0 \leqslant h_z \leqslant 60\}$. The medium consists of sandy loam for $38 \leqslant h_z \leqslant 60$ and fine sandstone for $0 \leqslant h_z \leqslant 38$. We refer to Table 4.1 for details of the parametrisations of these soils. Again, the base of the well is assumed to consist of impervious rock, and a no-flow boundary condition is enforced. There is assumed to be no water flow at the land surface. The water table has been measured in the vicinity of the well to be $49.8m$, so we set a hydrostatic boundary condition at $r = 50$ to represent the far field conditions around the well. The height of water in the well is $42.7m$. The initial mesh is aligned with the layers in the soil. The

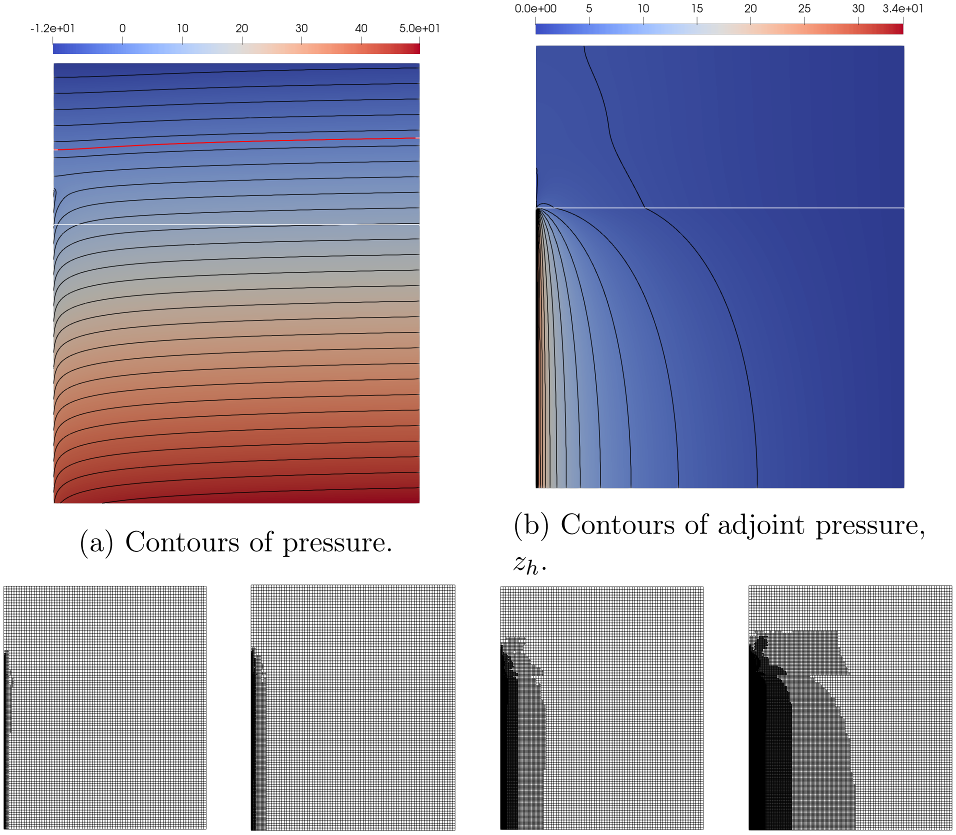
Table 4.1: Case study soil parameters. Parameters used in the van Genuchten-Mualem model for hydraulic conductivity in each of the several types of soil and rock. Note the differences of several orders of magnitude in the parameters K_S^i, causing strong discontinuities in the coefficient k.

Layer	K_S (ms^{-1})	n	α (m^{-1})
sandy loam	5E-6	1.65	0.66
med. sandstone	9E-6	1.36	0.012
slate	5.0E-9	6.75	0.98
fine sandstone	1.15E-6	1.361	0.012
diabase	2E-5	1.523	1.066

solution, together with a selection of adaptive meshes are given in Figure 4.9. The computed flux as a function of degrees of freedom is given in Figure 4.11a showing that the mathematical model is in good agreement with the experimental data.

4.8.2 Case study 2 - 5 layered well in Porto Ferreira (CPRM reference 3500009747)

The second case study is a challenging setup with five layers of highly varying hydraulic properties, as well as complex boundary conditions due to the fact that in this case the inner wall of the well is impermeable apart from two filters to allow water to flow into the well. One is below and one above the water, meaning that the former allows flow into the subsurface and the other allows flow out. Along the inner wall, filters cover the part of the wall with $5 \leqslant h_z \leqslant 17$ and $19 \leqslant h_z \leqslant 23$. The water level in the well is set at 17.44, with the other boundary conditions as in case study 1, with the water table at the far boundary set at 33.9. Once again we assume a radially symmetric solution. The domain is given by $\Omega = \{(r, \phi, h_z) \mid 0.1585 \leqslant r \leqslant 50, 0 \leqslant h_z \leqslant 46\}$. The medium consists of five layers. In order, with the top layer first, the layers consist of sandy loam, medium sandstone, slate, coarse sandstone and diabase. The boundaries between the layers are at $h_z = 34$, $h_z = 18$, $h_z = 16$



(a) Contours of pressure.

(b) Contours of adjoint pressure, z_h.

(c) $\mathscr{T}^{15}$

(d) $\mathscr{T}^{20}$

(e) $\mathscr{T}^{30}$

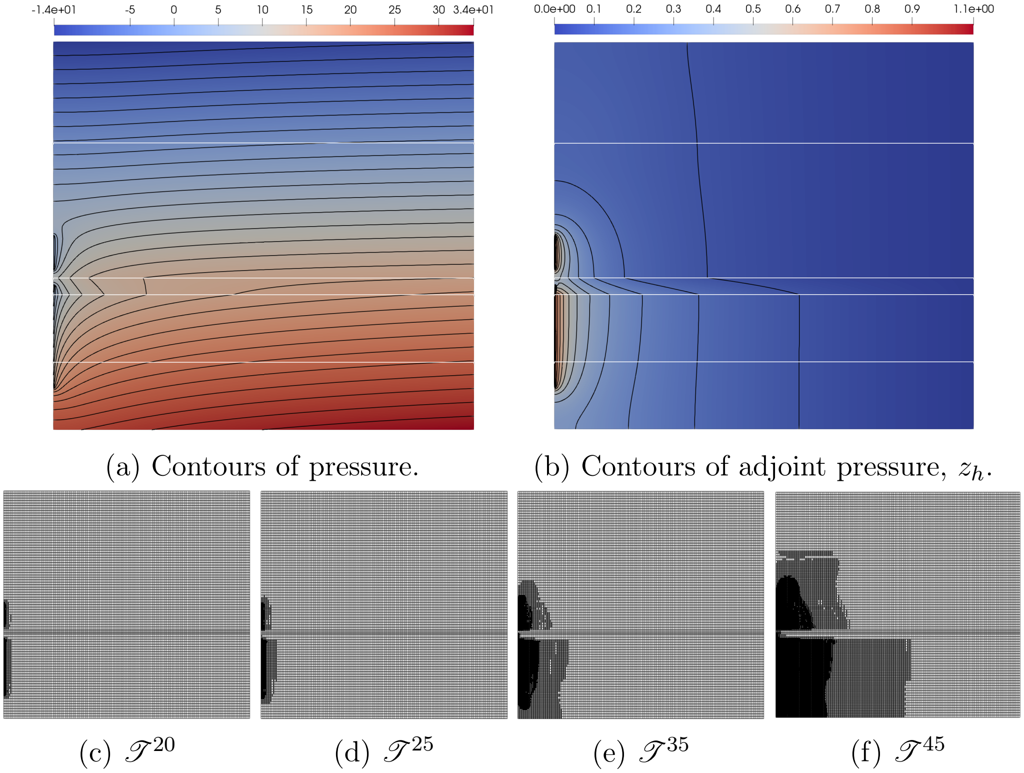
(f) $\mathscr{T}^{35}$

Figure 4.9: Case study 1, flow through a two layered soil. We show the pressure, the adjoint solution and a sample of adaptively generated meshes. The boundary between the soil layers is marked with a white line. Both solutions are represented on $\mathcal{T}^{35}$ which has approximately 1.5 million degrees of freedom.

and $h_z = 8$. We refer to Figure 4.8 for a visual description. The slate layer in particular causes this to be a difficult problem to simulate numerically due to its hydraulic conductivity being several of orders of magnitude smaller than those of the other soils and rocks. The initial mesh is aligned with the layers as well as the filter locations and the water level in the well. The solution, together with a selection of adaptive meshes are given in Figure 4.10. The computed flux as a function of degrees of freedom is given in Figure 4.11b showing a comparison between the mathematical model and the experimental data.

4.9 Conclusions & Discussion

In this chapter, we applied techniques from goal-oriented a posteriori error estimation to a challenging nonlinear problem involving a groundwater flow. For this class of problem, fine uniform meshes do not perform well. Indeed, in Figure 4.7 we see that convergence can be extremely slow on uniform meshes. By comparison, the dual-weighted error estimate was shown to perform well under a variety of conditions. It has been observed in previous studies (see for example [78]) that due to the approximations that must be made to evaluate the error representation numerically, the error estimate can perform poorly if the initial mesh in simulations is too coarse. In this particular case, we expect that the problem originates in the approximation of the dual problem. Since the dual solution must satisfy homogeneous Dirichlet boundary conditions on the seepage face defined by the primal solution, and since the forcing from the quantity of interest is largest here, there is a sharp boundary layer at the seepage face which is inevitably poorly resolved by a coarse mesh. Notwithstanding, the algorithm produces rapid error reduction with effectivity close to 1 once the mesh is sufficiently locally refined. This means that numerical error can be quantified with a high degree of confidence, and that the dual-weighted error estimate can be used as a termination criterion



(a) Contours of pressure.

(b) Contours of adjoint pressure, z_h.

(c) $\mathscr{T}^{20}$

(d) $\mathscr{T}^{25}$

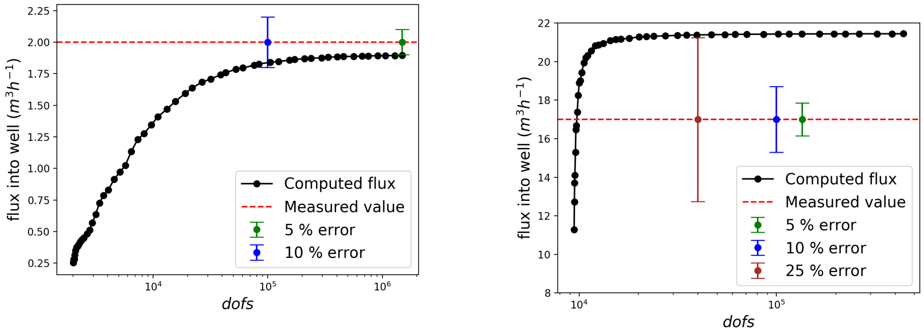
(e) $\mathscr{T}^{35}$

(f) $\mathscr{T}^{45}$

Figure 4.10: Case study 2, flow through a 5 layer soil. Level set $u = 0$ marked with red line. The boundaries between the soil layers are marked with white lines. See table 4.1 for a detailed description of the properties of each layer. Both solutions are represented on $\mathscr{T}^{45}$ which has approximately 500,000 degrees of freedom. Note that in this case the dual problem is much more interesting due to the structure of the inner wall of the well. The meshes appear to show that the soil layers have very different influences on solution accuracy. In particular, the slate layer shows little mesh refinement due to its low permeability relative to the other layers.

for an adaptive routine.

The case studies most clearly demonstrate the need for adaptive techniques in solving problems such as this. The multi-scale nature of inhomogeneous soil results in a problem which is extremely challenging to solve by conventional numerical methods. Indeed, the error remains large on uniform



(a) Case study 1. Note that the fully resolved model is within a 5% relative error of the experimental results with a 10% relative error at around 90000 degrees of freedom.

(b) Case study 2. The fully resolved model is around 25% relative error. This is already achieved with 20000 degrees of freedom.

Figure 4.11: Plots displaying the computed value of the water flux into the well under successive refinement cycles of the adaptive finite element method. This allows to infer the maximal amount of water pumped from the well whilst leaving the surrounding water table unchanged.

meshes even as the mesh approaches 10^5 degrees of freedom where in the adaptive case a steep and consistent reduction in error can be observed with successively refined meshes, see Figure 4.7. Applying these robust, computationally efficient methods to the case studies allows the accurate quantification of solutions to the variational inequality. Note, however, that these case studies are still extremely challenging. The assumption of layered soil, for example, may not always be physically meaningful. Indeed, we believe it is this assumption that affects the performance of case study 2. For highly variable soils we must use further information, for example those provided through resistivity methods. This is the subject of ongoing research, as explored in [6].

Chapter 5

Modelling of unsteady infiltration

5.1 Abstract

In this chapter we consider an unsteady model of water infiltration - Richards' equation. Richards' equation is known to be difficult to solve numerically, particularly in many cases of practical interest such as instense rainfall scenarios, fine grained soils and heterogeneous soil structure. We introduce Richards' equation, describe its weak formulation and discuss its solution by the finite element method, including linearisation and time discretisation. We present the results of numerical benchmark simulations to verify the performance of the scheme. This is followed by an application to more difficult problems including realistic soil parametrisations, illustrating the situations in which numerical schemes may fail to converge. The reasons for this failure are explored, and a regularisation of the nonlinear relations that describe permeability of the soil is suggested, and shown to mitigate the solver convergence problems.

5.2 Introduction

The use of potentials in a way analogous to the study of electric fields was one of the milestones in the introduction of mathematics into soil science. This methodology was introduced by Buckingham in a seminal work on soil science [27] which aimed to derive a theory for water flow in soil analogous to those for the movement of electricity and heat, respectively Ohm's and Fourier's laws. This methodology was taken up by Richards in [89], where the author made the case for studying unsaturated flow using negative values of the potential, backed up by the observation that in unsaturated soil the pressure of the fluid is less than that of the atmosphere. The capillary potential u was introduced as

$$u = \frac{p}{\rho_d}, \tag{5.1}$$

where p is the pressure of the fluid in the soil pores and ρ_d is the density of the fluid. This paper also presented experimental results showing what later became known as water retention curves, that is the relationship between the capillary potential and the amount of water held by the soil.

Following a few years later in [88], the author gives one of the earliest full descriptions of variably saturated flow of water in a porous medium in the form of a nonlinear partial differential equation, which subsequently became known as Richards' equation. His paper followed the observation that within certain ranges, unsaturated flow could be modelled with Darcy's law, but with variable permeability that depends on the water content of the soil.

There are three common formulations of Richards' equation. We shall be concerned with the mixed form, so-called because it contains both pressure and saturation.

$$\frac{\partial \theta(u)}{\partial t} - \nabla \cdot \{k(u) \nabla (u + h_z)\} = 0, \tag{5.2}$$

where h_z denotes the vertical height against a fixed datum, $\theta = \theta(u)$ is a function giving water content in terms of the capillary potential $u = u(\boldsymbol{x}, t)$, and $k = k(\theta(u))$ is the hydraulic conductivity, a nonlinear function of capillary potential. Here for example we may have in mind the functional forms of the van Genuchten-Mualem model described in §4.6.1.

We remark that there are other forms of Richards' equation, but we choose (5.2) as it is more applicable to heterogeneous soils, arises naturally from a conservation argument, and avoids the introduction of other functional forms such as diffusivity or specific moisture capacity which may not be well-defined in the saturated limit. Since our key regime of interest is infiltration, it is crucial that we are able to model variably saturated flows effectively.

Much of the difficulty associated with solving Richards' equation in practical contexts stems from the models of k and θ described in §4.6.1. It is not possible to capture all properties of the subsurface, as measurement is often performed indirectly with limited accuracy. Electroresistivity methods are commonly used to infer subsurface structure. They provide a non-invasive way to obtain resistivity profiles which can then be converted to other soil properties using empirical relationships [95]. These methods have several limitations. They typically produce a resistivity field at one moment in time, and therefore cannot contain information about how resistivity changes in time, meaning that it can be difficult to separate effects from soil properties and water content. They can however be very effective when combined with geological information to infer for example layered soil and rock structures, as well as obstacles such as pipes and other man-made structures that can inform models.

Soil also exhibits hysteresis, meaning that in general a drying soil will contain more water than a wetting soil at the same pressure. In this work, we will not consider hysteresis, that is, we will select a single function to describe the pressure-saturation relationship. For a review on models of hysteresis in soils, we refer the reader to [85].

We now turn our attention to modelling of hydraulic properties of individual soil types. Such models form the building blocks of more involved simulations such as those with soils consisting of homogeneous layers, or homogeneous soils with obstacles. Many models have been proposed for the hydraulic conductivity curve k and water retention curve θ. One group uses a power law to relate hydraulic conductivity to the dimensionless water content:

$$K_R = \Theta^\alpha, \tag{5.3}$$

see for example [25]. Another approach derives an expression for the hydraulic conductivity assuming knowledge of the water retention curve (see [75]). This was then combined with a flexible class of models for the water retention curve in [106] to give the popular van Genuchten-Mualem model. The model parameters are chosen by a curve fitting algorithm.

For certain types of soils (particularly fine-textured soils) the hydraulic conductivity function can be very steep, and can cause numerical instability, especially when combined with rapid infiltration or internal boundaries in the soil between clay and a more conductive soil [59]. In soil science, some authors modify the van Genuchten model in an attempt to obtain more physically correct parametrisations (see e.g. [111], where the modification seems to mitigate some numerical difficulties is specific test cases). This is an ongoing research area with new soil models still being proposed to replicate physical behaviour in very dry or very wet soils [113]. Later in this chapter, we develop a novel approximation to the original van Genuchten model that can potentially be applied more generally to problems with highly nonlinear coefficients.

One feature common to all models of the water retention curve and hydraulic conductivity is strong nonlinearity. Richards' equation is therefore expensive to solve numerically, and ensuring computational efficiency becomes very important.

Despite these challenges, this equation has proved a popular model for transient variably-saturated subsurface flow in numerical studies [12, 69, 115], practical engineering applications [94] and coupled with other processes as part of larger projects [28].

The rest of this chapter is set out as follows. We begin with an introduction to the numerical solution of Richards' equation by finite element methods in §5.3, where we discuss different linearisation approaches and time discretisation. These methods are applied to numerical examples in §5.4, and the situations in which standard methods can fail are addressed. In §5.5, we introduce a regularisation of the hydraulic conductivity model of van Genuchten and Mualem. Finally, this regularisation and its effect on the performance of nonlinear iterative schemes for Richards' equation is investigated numerically in §5.6.

5.3 Finite element methods for evolution problems

In this section we describe the finite element method and its application to solving Richards' equation. We first introduce the basic concepts necessary to formulate equation (5.2) weakly. We then give an overview of theoretical results that have been achieved for Richards' equation and related problems. We then discuss the time discretisation and linearisation of the strong form of the problem. Finally, the linearised problem is discretised in space using a finite element method.

To fix ideas, we seek solutions to the mixed boundary value problem given by (5.2) on a finite time interval $[0, T]$, augmented by the boundary and initial conditions

$$\begin{aligned} u(\boldsymbol{x}, t) &= u^0(\boldsymbol{x}) \\ u &= g_D \quad \text{on } \Gamma_D \\ k(u)\nabla(u + h_z) \cdot \boldsymbol{n} &= g_N \quad \text{on } \Gamma_N, \end{aligned} \tag{5.4}$$

with $\Gamma_D \cup \Gamma_N = \partial\Omega$.

To formulate this problem weakly, we will need to define additional function spaces. We first consider spatial boundary conditions, and define

$$\mathrm{H}^1_{g,\Gamma_D}(\Omega) := \{v \in \mathrm{H}^1(\Omega) \,:\, v|_{\Gamma_D} = g\}, \tag{5.5}$$

for a function g defined on Γ_D.

Now let

$$\mathcal{W} := \{v \in \mathrm{L}^2(0,T;\mathrm{H}^1_{g_D,\Gamma_D}(\Omega)) : \partial_t\theta(v) \in \mathrm{L}^2(0,T;\mathrm{H}^{-1}_{\Gamma_D}(\Omega))\} \tag{5.6}$$

where $\mathrm{H}^{-1}_{\Gamma_D}(\Omega)$ is understood as the dual of $\mathrm{H}^1_{0,\Gamma_D}(\Omega)$.

$$\mathcal{Y} := \mathrm{L}^2(0,T;\mathrm{H}^1_{0,\Gamma_D}(\Omega)) \tag{5.7}$$

We are now ready to state the weak formulation of Richards' equation. Seek $u \in \mathcal{W}$ such that

$$\int_0^T \langle \partial_t\theta(u), v\rangle + \langle k(u)\nabla(u+h_z), \nabla v\rangle \,\mathrm{d}t = \int_0^T \langle f, v\rangle \,\mathrm{d}t \quad \forall v \in \mathcal{Y}. \tag{5.8}$$

Remark 5.3.1 (Existence and uniqueness of solutions to (5.8))**.** *Due to the structure of the nonlinearities in the problem, the proof of existence of a solution to* (5.8) *is significantly more technical than some of its linear parabolic siblings, and will not be given here. The interested reader is referred to [3] for a detailed analysis.*

We will now define the finite element solution of (5.8). We recall the finite-dimensional space $\mathcal{V}_h$ from (2.33) and modify it to take account of the boundary conditions. We assume here the finite element partition aligns with the problem data in the sense that all edges $e \subseteq \partial\Omega$ are wholly contained in either Γ_D or Γ_N, and that g_D is piecewise polynomial so that elements of $\mathcal{V}_h$ may satisfy the boundary conditions exactly. Then we define

$$\mathcal{V}_h^{g_D} := \{v_h \in \mathcal{V}_h : v_h|_{\Gamma_D} = g_D\} \tag{5.9}$$

and

$$\mathcal{V}_h^0 := \{v_h \in \mathcal{V}_h : v_h|_{\Gamma_D} = 0\}. \tag{5.10}$$

Then the (semidiscrete) finite element solution is $U \in C^1(0, T; \mathcal{V}_h^{g_D})$ such that

$$\langle \partial_t \theta(U), \Phi\rangle + \langle k(U)\nabla(U + h_z), \nabla\Phi\rangle = \langle f, \Phi\rangle \quad \forall \Phi \in \mathcal{V}_h^0, \forall t \in [0, T]. \tag{5.11}$$

where we make the assumption that $\theta(U) \in C^1([0, T])$.

The result is a finite dimensional system of ODEs, and after choosing a basis for $\mathcal{V}_h^{g_D}$ becomes an algebraic problem, as described in §2.

5.3.1 Time-stepping and linearisation

In this section we work with a strong form of the problem to present the various linearisations we will consider for Richards' equation. We begin with a brief overview of the different iterative schemes used to solve Richards' equation. Newton's method is a popular choice for variably saturated flow, with versions of the method used in [12, 69] but is far from universal. Picard, or fixed-point iteration is the simplest choice, and has been used successfully in realistic dynamic case studies such as [94], but can be unstable. An alternative stable scheme is presented in [61]. It converges independently of the initial guess but requires an additional computational overhead.

There are several schemes that lie between Picard and Newton methods that are essentially stabilised fixed point iteration. The modified Picard method developed in [30] includes gradient information for the water content function but not hydraulic conductivity, since the former is generally smoother. In place of this gradient, the L-method (see [97]) uses a positive constant for stabilisation. In [97], the author proves convergence of a weak form of Richards' equation for arbitrary initial guess, making it a robust choice. In many cases the condition number of the linear system that must be solved at each iteration of the nonlinear solver is lower, however this improvement is often balanced by the fact that many outer iterations are needed for the nonlinear solver to reach convergence.

A detailed comparison of these schemes is conducted in [69] in which it is observed that, when it converges, Newton is the fastest, with the L-method generally being the slowest. However the authors registered convergence failures for Picard and Newton in challenging scenarios, suggesting that the L-method could be a useful tool in soils with particularly steep characteristic curves.

We now select a time-stepping scheme. A common choice for solving Richards' equation is the backward Euler scheme (see [69, 12, 94]) in view of its favorable stability properties, though others are available (cf. [115]). We utilise the backward Euler scheme for our simulations to help minimise the effect of the large variations in k as this scheme is *more dissipative* than the original problem.

Consider (5.2) on the time interval $[0, T]$. We define a subdivision of the time domain $[0, T]$ into a partition of N consecutive adjacent subintervals with endpoints denoted $0 = t^0 < t^1 < \cdots < t^N = T$. We denote the n-th timestep as $\tau^n = t^n - t^{n-1}$ and consistently use the shorthand $F^n(\cdot) = F(\cdot, t^n)$ for a time dependent function. Given u^0, for $n = 1, \ldots, N$ we define at each time level an approximate temporally semidiscrete u^n as the solution of

$$\theta(u^n) - \theta(u^{n-1}) - \tau^n \nabla \cdot k(u^n) \nabla (u^n + h_z) = \tau^n f^n. \tag{5.12}$$

This is then an iterative family of nonlinear problems, indexed by the timestep.

We define for each time step the nonlinear residual function evaluated at v

$$F^n(v) = \theta(v) - \theta(u^{n-1}) - \tau^n \nabla \cdot k(v) \nabla (v + h_z) - \tau^n f^n, \tag{5.13}$$

allowing us to formulate the nonlinear problem (5.12) as seeking u^n such that

$$F^n(u^n) = 0. \tag{5.14}$$

Let $DF^n(v, w)$ denote the directional derivative of F^n at v in direction w:

$$DF^n(v, w) = \lim_{\epsilon \to 0} \left(\frac{F^n(v + \epsilon w) - F^n(v)}{\epsilon} \right). \tag{5.15}$$

Further, let δ_j be the update to the current iterate. Then Newton's method applied to solve (5.14) starting from an initial guess $u_0^n = u^{n-1}$ and, assuming for now that θ and k have well-defined derivatives, is given by seeking δ_j for $j = 0, \ldots$ such that:

$$DF^n(u_j^n, \delta_j) = -F^n(u_j^n) \tag{5.16}$$

and

$$u_{j+1}^n = u_j^n + \delta_j. \tag{5.17}$$

Under the assumptions above on the differentiability of k and θ, one calculates

$$DF^n(v, w) = \theta'(v) w - \nabla \cdot (k(v) \nabla w + w k'(v) \nabla (v + h_z)) \tag{5.18}$$

We now use the fact that $u_{j+1}^n = u_j^n + \delta_j$ to write (5.16) as follows. Given $u_0^n, u_1^n, ...u_j^n$, find u_{j+1}^n solving

$$\begin{aligned} \tau^n f^n &= \theta(u_j^n) - \theta(u^{n-1}) + \theta'(u_j^n)(u_{j+1}^n - u_j^n) \\ &\quad -\tau^n \nabla \cdot \left[k(u_j^n) \nabla (u_j^n + h_z) + (u_{j+1}^n - u_j^n) k'(u_j^n) \nabla (u_j^n + h_z) \right]. \end{aligned} \tag{5.19}$$

Thus, we obtain a sequence of functions which will converge to the solution under assumptions on the initial guess and the regularity of the coefficients. At each stage, we can monitor the convergence of the method using the residual $F^n(u_j^n)$.

5.3.2 Alternatives to Newton's method

Newton's method is quadratically convergent once within its convergence radius, and so is an attractive choice when it can safely be applied. It is however only locally convergent, and requires the calculation of derivatives of both θ and k. As we have seen in §4.6.1 however, some of the most widely applied parametrisations for these coefficients do not have the necessary regularity.

Picard or fixed point iteration

The simplest possible iterative scheme is the Picard method, which at time level n is to find u_n solving

$$\theta(u_j^n) - \theta(u^{n-1}) + \tau^n \nabla \cdot \left[k(u_j^n)\nabla(u_{j+1}^n + h_z)\right] = \tau^n f^n. \tag{5.20}$$

All terms involving derivatives are omitted leaving a fixed-point iterative scheme in which the nonlinear coefficients are evaluated at the previous iterate. Although convergence will typically be first order and therefore slower in terms of number of iterations, each iteration will be cheaper due to the fact that there is no need to compute derivatives, and this scheme has been used successfully for realistic simulations of variably saturated flow [94].

Modified Picard method

It is often the case that the water retention curve, θ is smoother than the hydraulic conductivity, k, and so we can improve (5.20) by including a higher order approximation to the nonlinearity in θ to obtain the scheme given in

[30], essentially a Newton approximation for θ and a Picard approximation for k,

$$\theta(u_j^n) - \theta(u^{n-1}) + \theta'(u_j^n)(u_{j+1}^n - u_j^n) + \tau^n \nabla \cdot \left[k(u_j^n)\nabla(u_{j+1}^n + h_z)\right] = \tau^n f^n. \tag{5.21}$$

This comes at the modest expense of having to calculate a derivative of the water retention curve, but the cost is mitigated by the stabilisation effect gained from including it. The term $\theta'(u_j^n)(u_{j+1}^n - u_j^n)$ acts in a similar manner to the L term in the L-method presented below, but the modified Picard method lacks the guaranteed convergence enjoyed by the L-method.

L-method

Finally, another stabilised Picard method has been designed specifically for Richards' equation and analysed in [97],

$$\theta(u_j^n) - \theta(u^{n-1}) + L(u_{j+1}^n - u_j^n) + \tau^n \nabla \cdot \left[k(u_j^n)\nabla(u_{j+1}^n + h_z)\right] = \tau^n f^n. \tag{5.22}$$

Here $L > 0$ is a parameter that, if chosen correctly, guarantees convergence of this scheme.

A detailed comparison of these schemes is conducted in [69] in which it is observed that, when it converges, Newton is the fastest, with the L-method generally being the slowest. However the authors registered convergence failures for Picard and Newton in challenging scenarios, suggesting that the L-method could be a useful tool in soils with particularly steep characteristic curves.

5.3.3 Fully discrete form

We are now ready to define a fully discrete form as a spatial discretisation of (5.21) using finite elements. Let U^0 denote a discrete approximation of the initial condition u^0. We then define U_j^n iteratively for $n = 1, \ldots, N$, $j = 1, \ldots, J$ with $U_0^n = U^{n-1}$ where U^{n-1} is the final iterate from the previous time step, as the element of $\mathcal{V}_h^{g_D}$ such that

$$\begin{aligned}\left\langle\theta(U_{j-1}^n), \Phi\right\rangle + \left\langle\theta'(U_{j-1}^n)(U_j^n - U_{j-1}^n), \Phi\right\rangle + \tau^n \left\langle k(U_{j-1}^n)\nabla(U_j^n + h_z), \nabla\Phi\right\rangle = \\ \left\langle\theta(U^{n-1}), \Phi\right\rangle + \tau^n \left\langle f^n, \Phi\right\rangle \quad \forall\Phi \in \mathcal{V}_h^0.\end{aligned} \tag{5.23}$$

We have therefore defined a sequence of problems whose solutions U_J^n converge to the finite element approximation of the time-discrete solution u^n. Since the finite element method converges with order h, and backward Euler exhibits linear convergence in τ, assuming convergence of the nonlinear iteration, we should hope for $\|u - U^n\|_{\mathrm{H}^1} \leqslant C(\tau + h)$ and $\|u - U^n\|_{\mathrm{L}^2} \leqslant C(\tau + h^2)$.

We also state below the Newton's method for this problem to show the difficulties in fully linearising k. It is similar to (5.23), but with a derivative of the coefficient k included also. Find $U_j^n \in \mathcal{V}_h^{g_D}$ such that

$$\begin{aligned}\left\langle\theta(U_{j-1}^n), \Phi\right\rangle + \left\langle\theta'(U_{j-1}^n)(U_j^n - U_{j-1}^n), \Phi\right\rangle + \tau^n \left\langle k(U_{j-1}^n)\nabla(U_j^n + h_z), \nabla\Phi\right\rangle + \\ \tau^n \left\langle k'(U_{j-1}^n)\nabla(U_{j-1}^n + h_z)(U_j^n - U_{j-1}^n), \nabla\Phi\right\rangle \\ = \left\langle\theta(U^{n-1}), \Phi\right\rangle + \tau^n \left\langle f^n, \Phi\right\rangle \quad \forall\Phi \in \mathcal{V}_h^0.\end{aligned} \tag{5.24}$$

Notice that the linearisation in k appears as a convection term. For k defined for clay, for example as shown in figure 4.2, the problem could bcome heavily convection dominant, resulting in unstable solutions.

5.4 Numerical examples

In this section we present a collection of examples to introduce standard test cases for Richards' equation and observe the performance of the numerical scheme in some challenging cases.

5.4.1 Benchmarking covergence rates: Hornung-Messing problem

To verify correctness of the scheme and test convergence rates, we begin with a standard test case known as the Hornung-Messing problem ([19, 50, 97]). It is a two-dimensional problem in the horizontal plane, and therefore gravity has no effect on the flow. The problem does however retain many of the key difficulties of the problem, namely its double nonlinearity and degenerate parabolic-elliptic character and therefore serves as an appropriate benchmark for numerical schemes.

We consider the problem

$$\partial_t b(u) - \nabla \cdot (K_b(u) \nabla u) = f \quad \text{in} \quad \Omega \tag{5.25}$$

$$u = g \quad \text{on} \quad \Gamma_D \tag{5.26}$$

$$u(0, x, z) = u_0, \tag{5.27}$$

where $\Gamma_D \equiv \partial\Omega$ and

$$b(s) = \begin{cases} \frac{\pi^2}{2} - 2\arctan(s)^2 & s < 0 \\ \frac{\pi^2}{2} & s \geqslant 0 \end{cases} \tag{5.28}$$

$$K_b(s) = \begin{cases} \frac{2}{1+s^2} & s < 0 \\ 2 & s \geqslant 0. \end{cases} \tag{5.29}$$

Then if we let $s = x - y - t$, one may check that with appropriate boundary and initial data, problem (5.25)-(5.27) is satisfied exactly by

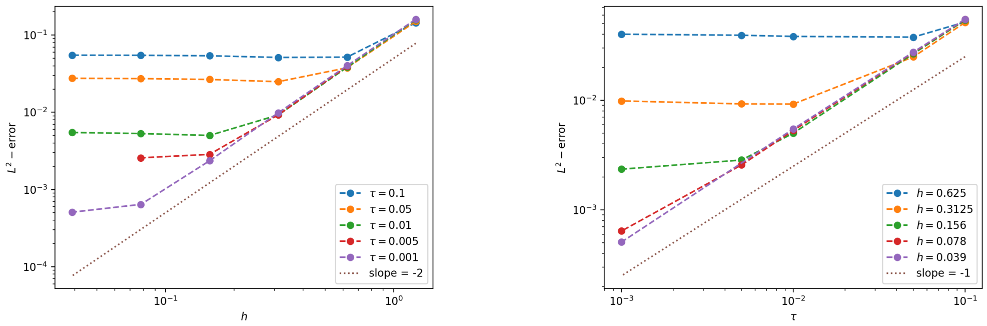
$$u_{\text{exact}}(x,y,t) = \begin{cases} -\dfrac{s}{2} & s < 0 \\ -\tan\left(\dfrac{e^s - 1}{e^s + 1}\right) & s \geqslant 0. \end{cases} \tag{5.30}$$

We should note here that u_{exact} has two continuous derivatives and is therefore smoother than we can expect in general for solutions of Richards' equation. Simulations were conducted on the domain $\Omega = (0,10)^2$ for $t \in [0,2]$. Uniform meshes consisting of square elements were used. The initial condition is taken to be the Lagrange interpolant of u_0. The problem is linearised using Newton's method (5.24) since in this case the coefficients are continuously differentiable, and in any case we are interested in the discretisation error. A tolerance of 10^{-9} is chosen as a stopping criterion for the Newton iteration to ensure that the linearisation error is small compared to the discretisation error.

The results of the suite of test simulations are shown in figure 5.1. They show that the numerical error in the pressure head measured in the $L^2(\Omega)$-norm is $\mathcal{O}(h^2 + \tau)$, which is the theoretical rate for piecewise linear finite elements combined with the backward Euler scheme for linear parabolic PDE problems such as the heat equation (see for example [105]).

5.4.2 Aquifer Recharge

In this section we present the results of simulations of Richards' equation in a standard benchmark test case ([12, 69, 115]). This is a relatively difficult and realistic problem in the sense that we do not expect a smooth solution, both due to the boundary conditions and the lack of regularity of coefficients. No exact solution is available. In addition, the scheme is presented with the difficult scenarios of steep infiltration and the joining of the infiltrating front with the water table. In both cases, rapid variation of the hydraulic



(a) The $L^2(\Omega)$ norm of the pressure head error against mesh size, each line having fixed time step.

(b) The $L^2(\Omega)$ norm of the pressure head error against time step, each line having fixed mesh size.

Figure 5.1: Test 1, §5.4.1. Second order convergence in space and first order convergence in time is observed. This is optimal for the discretisation used.

conductivity can occur, and cause instability of the method. This test case is therefore a useful tool to evaluate robustness.

Let $\Omega = \{(x, h_z) \mid 0 < x < 2,\, 0 < h_z < 3\}$ represent a vertical section of a subsurface region. For the purposes of this example, we assume that Ω is filled with homogeneous soil or rock. The upper portion of the boundary $\{(x, h_z) \mid h_z = 3\}$ represents the land surface while $\{(x, h_z) \mid h_z = 0\}$ is impermeable bedrock. To clearly describe the setup, we define the following sections of the boundary. Let $\Gamma_{D_1} = \{(x, h_z) \mid x = 2 \text{ and } 0 \leqslant h_z \leqslant 1\}$, $\Gamma_{D_2} = \{(x, h_z) \mid h_z = 3 \text{ and } 0 \leqslant x \leqslant 1\}$, and Γ_N be the remainder of the boundary so that $\partial\Omega = \Gamma_{D_1} \cup \Gamma_{D_2} \cup \Gamma_N$. On Γ_{D_1} a hydrostatic Dirichlet condition is enforced for the pressure with the water table height set at 1. This corresponds to setting the groundwater table in the surrounding soil.

We assume initially that the system has reached an equilibrium, that is, the initial condition is hydrostatic pressure consistent with the boundary condition on $\{(x, h_z) \mid x = 2\}$. The full problem specification is given below in equations (5.31) - (5.35).

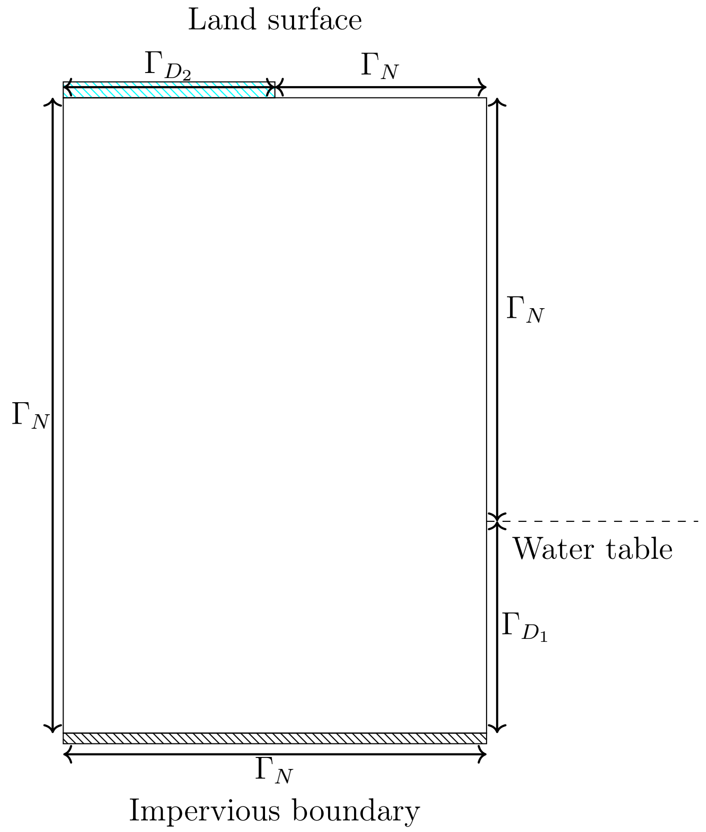


Figure 5.2: The aquifer recharge problem. The upper boundary is in contact with the atmosphere, with water pooling on the left hand part, while the lower boundary is assumed to be impermeable.

$$\partial_t\theta(u) - \nabla \cdot (k(u)\nabla(u + h_z)) = f \qquad \text{in} \quad \Omega \tag{5.31}$$

$$k(u)\nabla(u + h_z)(u) \cdot \mathbf{n} = 0 \qquad \text{on} \quad \Gamma_N \tag{5.32}$$

$$u = \Psi_0(1 - h_z) \qquad \text{on} \quad \Gamma_{D_1} \tag{5.33}$$

$$u = \min(-2\Psi_0 + 2(\Psi_0 + \Psi_1)t/t_D, 2\Psi_1) \qquad \text{on} \quad \Gamma_{D_2}, \tag{5.34}$$

$$u(0, x, h_z) = \Psi_0(1 - h_z) \tag{5.35}$$

The boundary condition on Γ_{D_2} is consistent with the initial condition at $t = 0$. Thereafter, the pressure head on Γ_{D_2} is increased smoothly up to a maximum of $2\Psi_1$ at a speed determined by the constant t_D to drive the flow. The constants Ψ_0, Ψ_1 and t_D allow the timescales and boundary conditions of the problem to be adjusted to allow for a wide range of soil types and infiltration intensities. Physically, this problem can be interpreted as water collecting in a drainage trench. We refer to figure 5.2 for a visual explanation.

The solutions have very different characteristics depending upon the choices of parametrisation. Below we present the results of simulations of homogeneous subsurface domains filled with sand, silt and clay.

5.4.3 Infiltration into sand

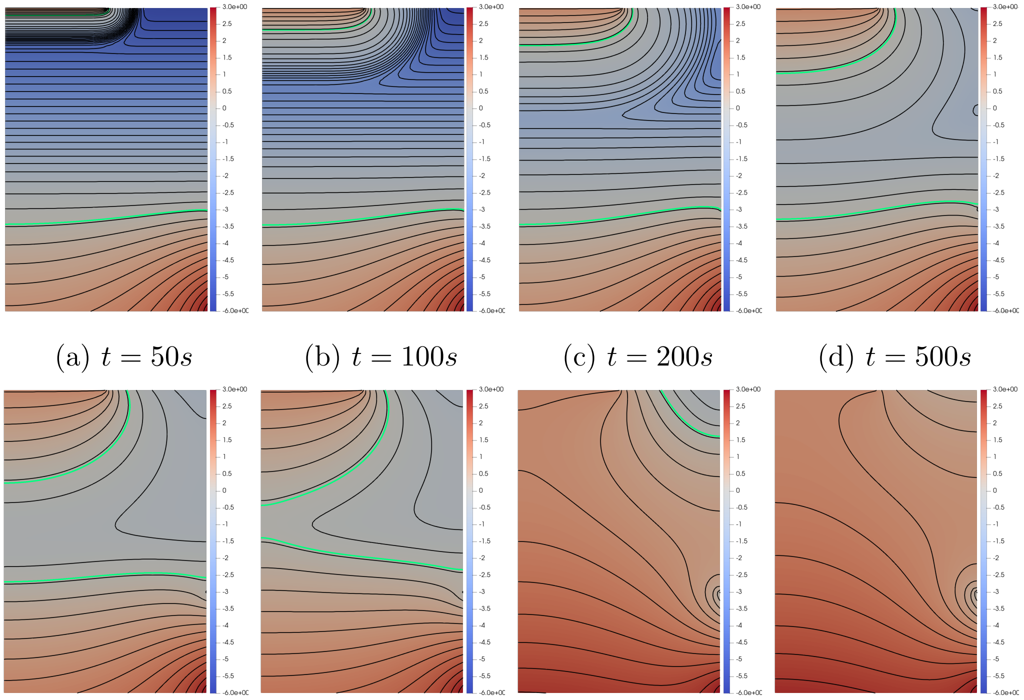
Following [12], we use a parametrisation of the water retention curve and hydraulic conductivity from [55], given in equations (5.36)-(5.37).

$$\theta(u) = \theta_R + (\theta_S - \theta_R)(1 + (-\alpha u)^n)^{-1}, \tag{5.36}$$

$$k(u) = K_S + (1 + (-\beta u)^m)^{-1}. \tag{5.37}$$

For this section only, time is measured in seconds. We select problem parameters $\Psi_0 = 3$, $\Psi_1 = 0.6$ and $t_D = 50s$ to force a strong infiltration over a short time. The soil parameters are chosen to be $\theta_R = 0.075$, $\theta_S = 0.3$, $\alpha = 2.71$, $n = 3.96$, $\beta = 2$, $m = 4.74$, $K_S = 10^{-4}ms^{-1}$. Plots of the pressure head are shown in figure 5.3.

Infiltration into sand attains a steady state after a little over half an hour. At the point where the domain becomes fully saturated, $\partial_t\theta = 0$ and the problem becomes elliptic, and the solution will not change in the absence of further forcing from boundary conditions or sources.



(a) $t = 50s$ (b) $t = 100s$ (c) $t = 200s$ (d) $t = 500s$

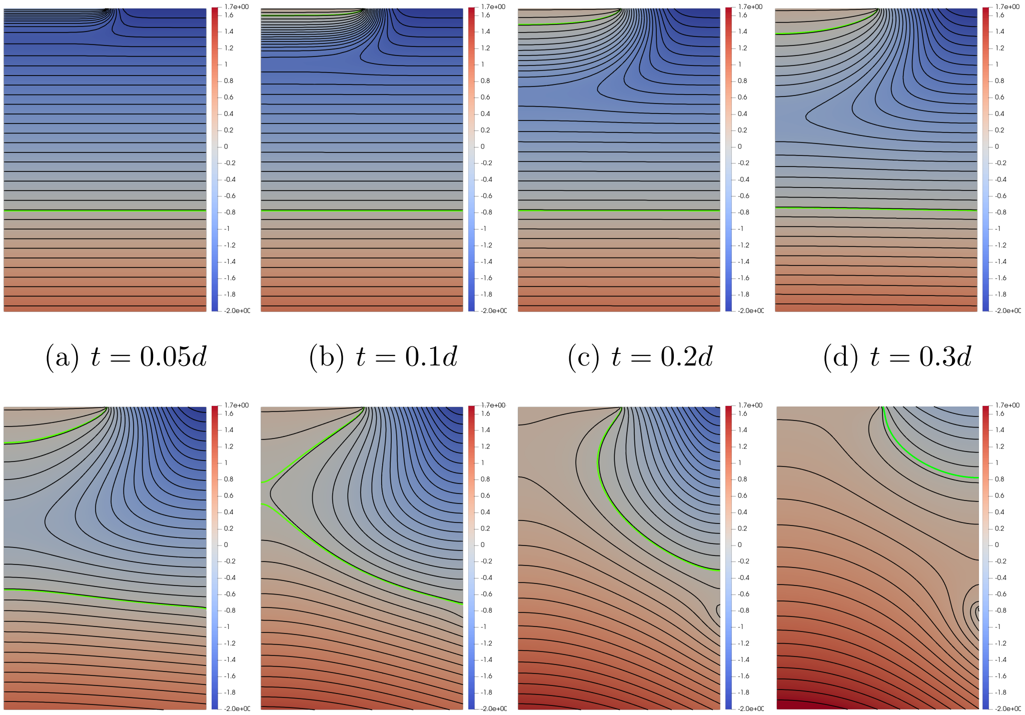
(e) $t = 1000s$ (f) $t = 1250s$ (g) $t = 2000s$ (h) $t = 2200s$

Figure 5.3: Infiltration into sand - §5.4.3. Level set $u = 0$ shown with green line. The gradient of u at the wetting front is initially large, but rapidly reduces due to relatively large permeability values. The saturated region joins the area below the water table at approximately $t = 1300s$, and the domain becomes fully saturated at approximately $t = 2200s$.

5.4.4 Infiltration into silt loam

We now simulate infiltration into silt loam. The soil is parametrised by $\theta_R = 0.131$, $\theta_S = 0.396$, $\alpha = 0.423$, $n = 2.06$, $K_S = 4.96^{-2} md^{-1}$. We choose parameters $\tau = 0.02d$, $t_D = 0.1d$, $\Psi_0 = 1$, $\Psi_1 = 0.1$. Selected plots at various times ae shown in figure 5.4.

In this case, the forcing at the surface is not sufficient to fully saturate the medium, and the solution tends asymptotically to a steady state. After approximately 1.5 days, the solution appears to have reached a steady state.



(a) $t = 0.05d$ (b) $t = 0.1d$ (c) $t = 0.2d$ (d) $t = 0.3d$

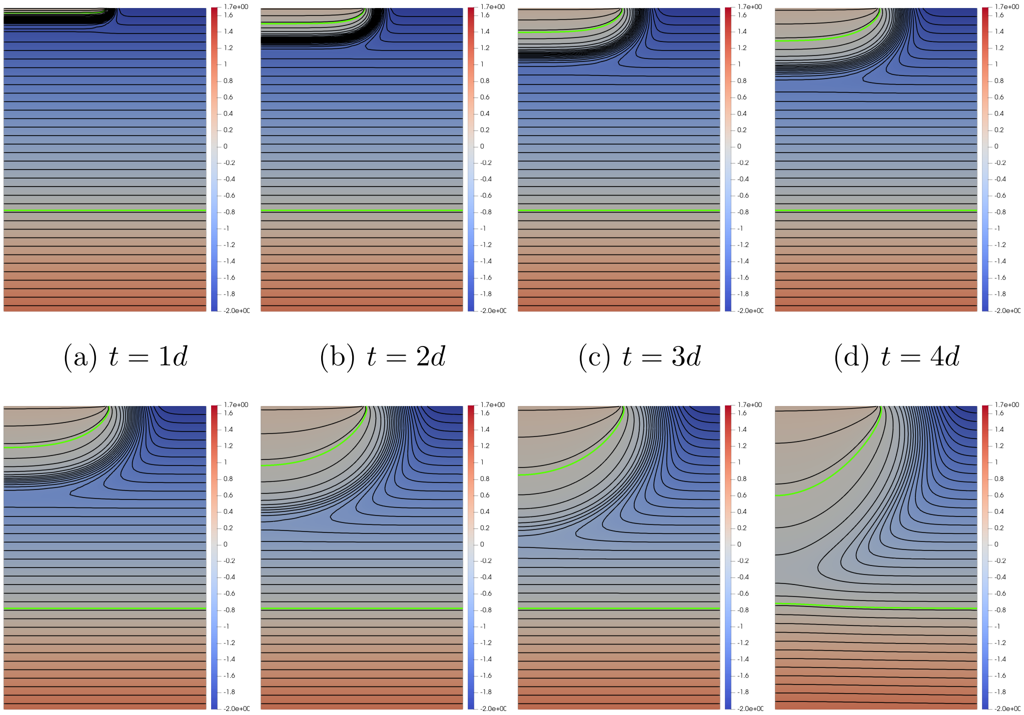
(e) $t = 0.4d$ (f) $t = 0.5d$ (g) $t = 1d$ (h) $t = 1.5d$

Figure 5.4: Infiltration into silt-§5.4.4. Level set $u = 0$ shown with green line. Silt soil exhibits a relatively smooth infiltrating front compared to the steeper fronts in figures 5.3 and 5.5. Figure 5.4h shows steady state after 1.5 days, but in this case the domain is not fully saturated.

5.4.5 Infiltration into clay

We now select $\Psi_0 = 1$, $\Psi_1 = 0.1$ and $t_D = 1$ day. We again use the van Genuchten model with parameters for Beit Natofa clay taken from [106] and converted to our units, namely $\theta_S = 0.446$, $\theta_R = 0$, $\alpha = 0.152$, $n = 1.17$ and $K_S = 8.2 \times 10^{-4} md^{-1}$. This is by far the most difficult case, since the van Genuchten parametrisation of hydraulic conductivity for clay has unbounded derivative as the saturation approaches its maximum value. In addition, the initial infiltrating front has very steep gradients which persist for longer times than in silt and sand due to the form of the hydraulic conductivity for clay.

For this problem, convergence failures of the nonlinear scheme were observed for Newton, modified Picard and L schemes. Depending on the time step and mesh size, failure occured reliably at one of two points in the simulation. The first is at around $t = 1d$ when the infiltrating front has fully developed. The second is at around $t = 10d$ when the infiltrating front meets the water table, and the topology of the saturated region changes. These scenarios cause instability of the numerical scheme manifesting around the wetting front, and the nonlinear iteration fails to converge. This leads us to consider a regularisation of the hydraulic conductivity function that will be discussed in the next section.



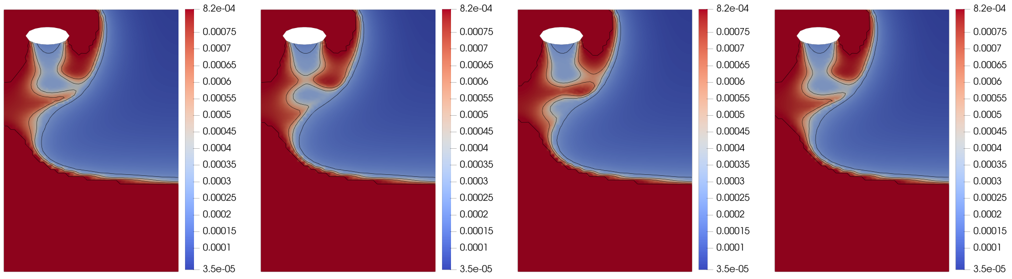
(a) $t = 1d$ (b) $t = 2d$ (c) $t = 3d$ (d) $t = 4d$

(e) $t = 5d$ (f) $t = 7d$ (g) $t = 8d$ (h) $t = 10d$

Figure 5.5: Infiltration into clay - §5.4.5. Level set $u = 0$ marked with green line.

5.4.6 Wetting front re-forming after passing obstacle

To elicit challenging solution behaviour, we place an impermeable obstacle in the path of the wetting front to obstruct the flow. As the two parts of the wetting front pass the obstacle and coalesce on the other side, the resulting singularity causes oscillations within the nonlinear iteration loop, and convergence failure results. The non-lipschitz nature of the hydraulic conductivity around saturation leaves the scheme unstable, as illustrated in figure 5.6. The profiles in figure 5.6 display unphysical oscillatory behaviour, as can be seen by the relatively large changes in hydraulic conductivity over very small time scales. This is caused by rapid oscillation in the hydraulic conductivity k as the two saturated regions (infiltrating front and groundwater zone) meet.



(a) $t = 11.44$ days (b) $t = 11.46$ days (c) $t = 11.48$ days (d) $t = 11.50$ days

Figure 5.6: Evolution of the hydraulic conductivity field for infiltration past an obstacle in clay soil.

5.5 Regularisation of the hydraulic conductivity

The instability of iterative schemes applied to challenging problems parametrised with the van Genuchten model was demonstrated in the numerical examples above. In this section we describe a method for regularising the van Genuchten hydraulic conductivity function. It is not uncommon for this relation to be modified to reduce instability in numerical calculations. For example, in [12], a spline approximation replaces k when the saturation exceeds 99%. We note that this actually requires modification of k for $u \in [-0.6911, 0]$, a relatively large range that makes the problem significantly easier to solve. The character of the solution is also changed when the hydraulic conductivity is altered (see figure 5.7).

We wish to construct an approximation that retains the shape of the curve, doesn't create numerical artefacts in the solution, and is able to be controlled by a regularisation parameter ε. When solving Richards' equation, the instability occurs for small negative pressure values where the hydraulic conductivity can oscillate rapidly. We set $\varepsilon > 0$ and define a function k_ε by modifying k in the interval $[-\varepsilon, 0]$.

$$k_\varepsilon(u) = \begin{cases} k(u), & \text{for } u \geqslant 0, u \leqslant -\varepsilon \\ p(u) & \text{for } -\varepsilon < u < 0 \end{cases} \tag{5.38}$$

where p is the unique quadratic polynomial such that $p(0) = k(0), p(-\varepsilon) = k(-\varepsilon)$ and $p'(-\varepsilon) = k'(-\varepsilon)$. In other words, we match derivatives at $u = -\varepsilon$, and ensure that the function k_ε is continuous. The result is a function that is not differentiable at $u = 0$, but is Lipschitz continuous for any $\varepsilon > 0$. Examples of k_ε are shown in figure 5.7e for three different values of ε.

Remark 5.5.1 (Choice of quadratic approximation.)**.** *When selecting a method of approximating the hydraulic conductivity k close to saturation, linear, quadratic and cubic approximations were considered. A linear approximation was discounted due to the non-matching derivative at $u = -\varepsilon$, which caused visual artefacts in the solution. A cubic spline would have been able to match derivatives at $-\varepsilon$ and 0, but as $\varepsilon \to 0$, this causes instability in the spline, leading to poor approximation. By contrast, the quadratic approach is well-behaved for small ε and respects the shape of the model curve.*

To investigate the effect of altering the hydraulic conductivity function, we conduct simulations of the aquifer recharge test case of §5.4.2 into clay soil for three choices of hydraulic conductivity. We choose a very small value of $\varepsilon = 0.04$ as the reference value, and compare with k_ε for $\varepsilon = 0.4$ and the cubic spline approximation used in [12]. Comparisons between pressure fields that result from these choices are shown in figure 5.7. We note that as more regularisation is used, the boundary between saturated and unsaturated soil lags behind, and gradients at the wetting front lessened. In the numerical examples of §4, it was demonstrated that the location of this boundary is an important feature that needs be well resolved.

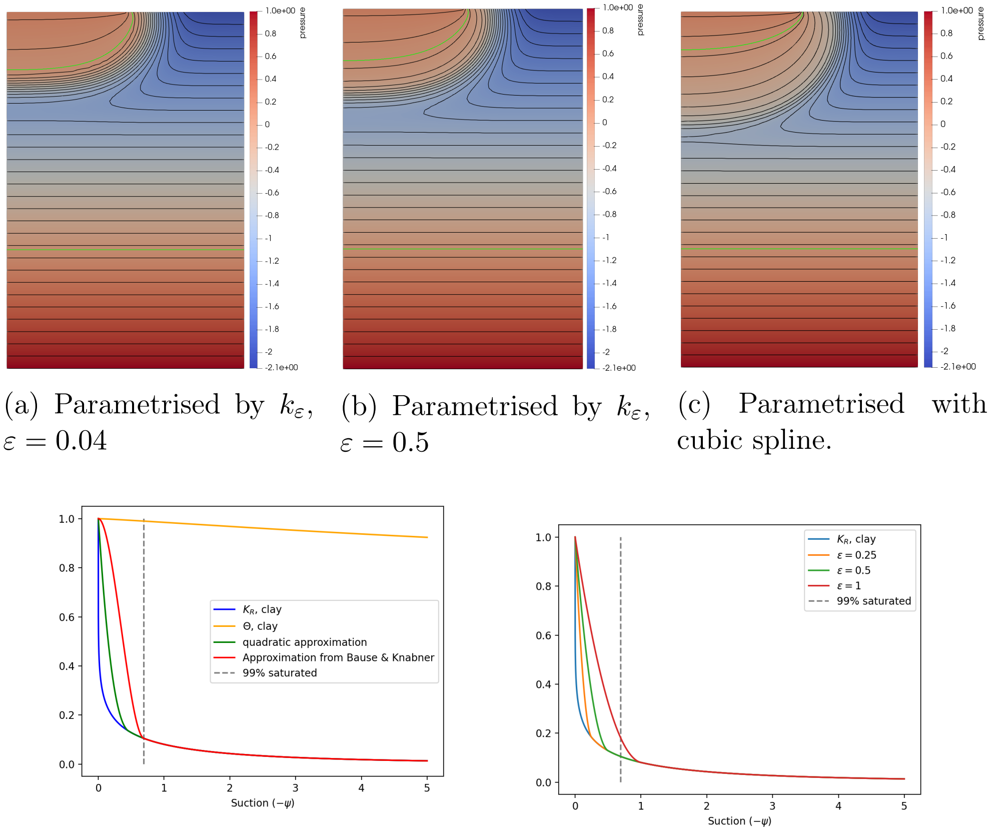
5.6 Numerical study of the effect of regularisation of k on a nonlinear solver for Richards' equation

In this section we investigate numerically the extent to which our modification of the nonlinear diffusion coefficient k above is able to improve the stabiliy of numerical solvers for Richards' equation. We perform numerical simulations of the aquifer recharge test case described in §5.4.2 for clay soil, and for different values of the regularisation parameter ε. Different time steps are used to demonstrate that smaller time steps do improve matters, but only up to a point, and that the principal issue is the rapid variation of k. All other problem parameters are kept constant.

The results are shown in figure 5.8. We see that the conditioning of the problem degenerates rapidly, and the timestep required becomes smaller and smaller. The behaviour of the nonlinear solver becomes unpredictable as $\varepsilon \to 0$, indicating a clear problem with robustness. Note also that when ε becomes small, reducing the time step to improve stability becomes impractical due to the extremely small values required.

5.7 Conclusions & discussion

In this chapter we introduced key concepts in numerical solution of parabolic partial differential equations, including time stepping and dealing with nonlinearity. We illustrated a key difficulty in the simulation of infiltration into soils with a steep hydraulic conductivity curve: the large variation in hydraulic conductivity between time steps can cause instability in nonlinear schemes, leading to convergence failure. This motivated the regularisation of the hydraulic conductivity. It was shown that this regularisation allows the nonlinear solver to converge, albeit to the solution of an approximate problem. To be useful in practice, it would be advantageous to quantify the error introduced by our regularisation. To this end, in the next chapter we conduct a posteriori analysis to determine how the regularisation should be used to obtain approximate solutions that converge to the correct solution upon sending h, τ and ε to zero.



(a) Parametrised by k_ε, $\varepsilon = 0.04$

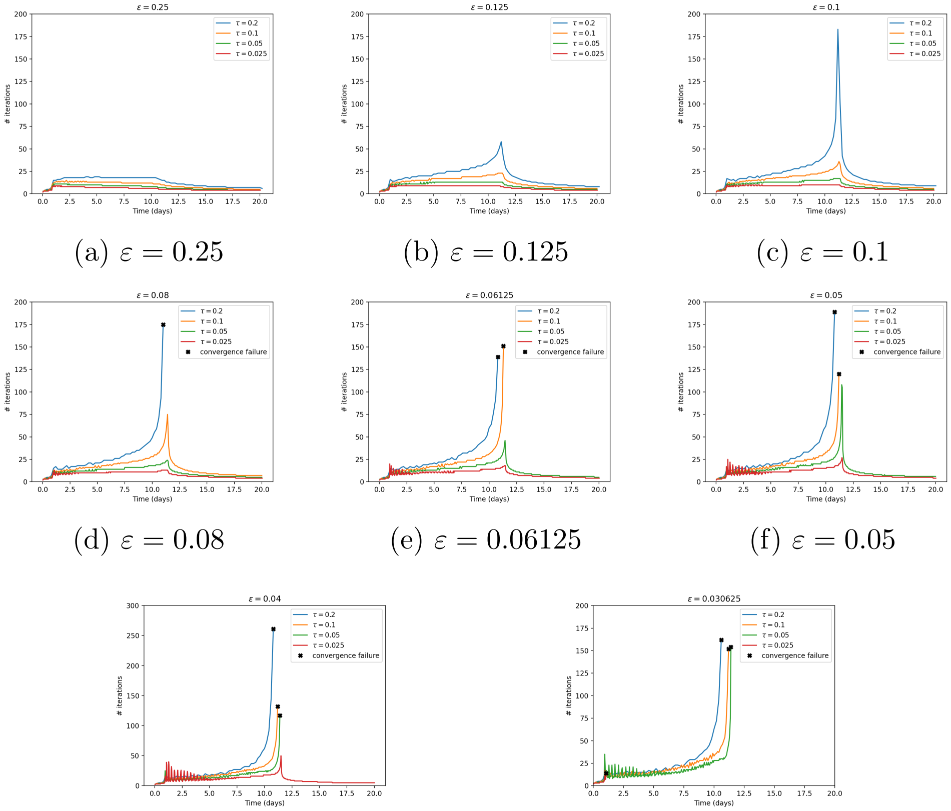
(b) Parametrised by k_ε, $\varepsilon = 0.5$

(c) Parametrised with cubic spline.

(d) Exact van Genuchten model, quadratic approximation and spline approximation used in [12].

(e) Comparison of quadratic approximations for several values of ε.

Figure 5.7: Pressure profiles at $T = 5$ days using three different models for the hydraulic conductivity. The regularised coefficient with small regularisation parameter $\varepsilon = 0.04$ in 5.7a serves as a reference solution for comparison. Approximating k results in a slightly spread out front but makes the problem easier to solve (i.e. nonlinear solver doesn't fail and no timestep adaptivity is required). The profile in 5.7c resulting from the regularised coefficient explains the difference between the simulations in clay soil presented here and those presented in [12]. Quadratic approximation of hydraulic conductivity model for clay soil retain the qualitative shape of the model, but have bounded derivative at zero for an $\varepsilon > 0$.



(a) $\varepsilon = 0.25$

(b) $\varepsilon = 0.125$

(c) $\varepsilon = 0.1$

(d) $\varepsilon = 0.08$

(e) $\varepsilon = 0.06125$

(f) $\varepsilon = 0.05$

(g) $\varepsilon = 0.04$

(h) $\varepsilon = 0.030625$

Figure 5.8: Comparison of the numbers of iterations required to solve the nonlinear problem at each time step. All axes use the same scale, each experiment uses the same spatial mesh and the same set of different time steps, the regularisation parameter ε is decreasing from left to right. A convergence failure was registered if the nonlinear solver had not reached tolerance within 400 iterations.

Chapter 6

Adaptive regularisation applied to Richards' equation

6.1 Abstract

In this chapter we investigate the regularisation of problem data in elliptic and parabolic PDE problems. We derive a posteriori error estimates that take account of the error due to the approximation of the coefficients for linear elliptic and parabolic cases. Motivated by convergence issues in nonlinear iterative solvers for Richards' equation detailed in the previous chapter, we apply indicators based on those derived in the linear cases to the nonlinear and time-dependent case. Here we are particularly concerned with the diffusion coefficient, which is a nonlinear and non-Lipschitz function of the solution.

6.2 Introduction

It is well known that numerical methods for solving Richards' equation are prone to suffering convergence failure in the nonlinear iteration, a phenomenon that was observed in chapter 5. The nonlinear coefficient of permeability changes rapidly with changes in soil pressure head, and therefore can rapidly oscillate from iteration to iteration, leading to the numerical solution blowing up or getting stuck in a loop. In the previous chapter, we introduced a regularisation of the permeability coefficient to stabilise the iteration and improve robustness of the numerical scheme. This came at the expense of accuracy, as around infiltrating fronts a small change in permeability can have a large effect. The goal is therefore to be able to control this regularisation locally, and obtain a sequence of stable approximations that converge to the solution.

In this chapter, we work towards a space-time adaptive scheme for Richards' equation in the following steps. We begin with a rigourous a posteriori error estimate for an elliptic problem with approximate coefficients. This takes the form of the usual residual estimate augmented by a term that quantifies the effect of approximating the coefficient. The latter term allows us to choose the regularisation parameter to be compatible with the mesh size to ensure that it does not negatively affect the convergence rate of the scheme as $h \to 0$. The analysis is then extended to a parabolic analogue for which the elliptic reconstruction is used to obtain a bound that takes the form of time-integrated error in an elliptic problem. This allows direct application of the previous result to obtain a rigorous bound in the parabolic case. Although we were unable to obtain an analogous bound in the degenerate, nonlinear Richards' equation case, we investigate the potential use of local error indicators inspired by the error estimate for the linear problem to combine mesh refinement with variable regularisation to increase the robustness of our simulations.

The elliptic reconstruction technique was introduced in [70] and used to address the suboptimality of a posteriori bounds in $\mathrm{L}^\infty(0, T, \mathrm{L}^2)$ obtained via energy arguments. It has proven rather flexible in that bounds of various forms may be used in time-dependent analogues. In [67], gradient recovery estimators are used to construct parabolic a posteriori estimates, while in

[60], the elliptic reconstruction is used in combination with techniques for hyperbolic conservations laws. The energy framework has been extensively developed for a posteriori error analysis of parabolic problems, with residual based estimates derived in [86, 31, 16]. We prefer in this case to use the elliptic reconstruction for its flexibility.

Many works analysing numerical schemes for Richards' equation work with the Kirchoff transformed problem. While this does have the advantage of reducing issues that arise from steep pressure gradients at infiltrating fronts, the transformed variable is not readily interpreted in a physical sense, and due to the fact that the water content has vanishing derivative in saturated soil, it has to be regularised. In this work we prefer to use one of the more standard formulations that arises from a conservation argument, leave the water content function unaltered and regularise the permeability instead.

We also mention here [59] where it is suggested that the van Genuchten curve for hydraulic conductivity is adjusted close to saturation for improved reproduction of experimental results. This would have the effect of lessening the steep gradients that occur and in a way regularise the problem. The regularisation would however not be controllable (it would be fixed for a given soil). We therefore pursue a more flexible approach that allows us to control the amount of regularisation *on the fly* to allow for the closest approximation to the true model that a given discretisation can handle. We also note that this could potentially be applied to other problems, for example the Stefan problem, studied in [81].

6.3 A posteriori analysis of a regularised elliptic problem

To illustrate ideas and provide analytical motivation for later sections, we consider first the following elliptic problem in weak form.

$$
\begin{aligned}
-\nabla \cdot (A\nabla u) &= f \quad \text{in } \Omega \\
u &= 0 \quad \text{on } \partial\Omega.
\end{aligned}
\tag{6.1}
$$

We assume Ω is subdivided into a triangulation satisfying the same assumptions laid out in §2.5.1 and recall the finite element space $\mathcal{V}_h$ from §2.5.2. We assume that $f \in \mathrm{H}^{-1}(\Omega)$ and that $A \in \mathrm{L}^{\infty}(\Omega)$ is uniformly positive definite, that is, there exists a constant $C_A > 0$ such that for all $\boldsymbol{v} \in \mathbb{R}^N$ we have

$$
\boldsymbol{v}^T A \boldsymbol{v} \geqslant C_A |\boldsymbol{v}|^2.
$$

For exposition, let us consider a situation where A is discontinuous and A_ε represents a smoothed coefficient with smoothing parameter $\varepsilon > 0$ that is Lipschitz continuous. We consider a model linear elliptic problem, together with its regularisation, and a finite element approximation to the regularised problem. We seek u, $u_\varepsilon \in \mathrm{H}^1_0(\Omega)$ and $U_\varepsilon \in \mathcal{V}_h$ such that

$$
\langle A\nabla u, \nabla\varphi\rangle = \langle f, \varphi\rangle \quad \forall \varphi \in \mathrm{H}^1_0(\Omega), \tag{6.2}
$$

$$
\langle A_\varepsilon\nabla u_\varepsilon, \nabla\varphi\rangle = \langle f, \varphi\rangle \quad \forall \varphi \in \mathrm{H}^1_0(\Omega), \tag{6.3}
$$

$$
\langle A_\varepsilon\nabla U_\varepsilon, \nabla\Phi\rangle = \langle f, \Phi\rangle \quad \forall \Phi \in \mathcal{V}_h. \tag{6.4}
$$

Remark 6.3.1. *We have in mind that* (6.2) *is the* exact *problem we are trying to solve, but for computational reasons we approximate this with* (6.4). *However, note that* (6.4) *is the approximation of a perturbed problem* (6.3).

Remark 6.3.2. *For later problems, we have in mind a diffusion tensor of the form*

$$
A = A(u) = K(\boldsymbol{x})k(u)I
$$

where I is the identity tensor, k is a non-Lipshitz function such as a model of hydraulic conductivity in a fine-grained soil, and K is a spatially dependent function that represents variations in the domain. For example, for uniform soil, $K \equiv 1$ or for layered soil we could take K piecewise constant. We remark that regularity of the coefficient tensor A is not so crucial in the linear case, but that less regular functions will still be more difficult to represent accurately within a finite element method.

We begin with a standard stability result.

Lemma 6.3.3 (Stability)**.** *Let u be the solution of* (6.2)*. Under the assumptions on A stated above, we have the stability bound*

$$\left\|A^{1/2}\nabla u\right\|_{\mathrm{L}^2(\Omega)} \leqslant \left\|A^{-1/2}\right\|_{\mathrm{L}^\infty(\Omega)} \left\|f\right\|_{\mathrm{H}^{-1}(\Omega)}. \tag{6.5}$$

Proof. To begin, set $\varphi = u$ in (6.2) with and use the $\mathrm{H}^1_0(\Omega)$-$\mathrm{H}^{-1}(\Omega)$ duality splitting to arrive at

$$\begin{aligned}\left\|A^{1/2}\nabla u\right\|^2_{\mathrm{L}^2(\Omega)} &\leqslant \left\|\nabla u\right\|_{\mathrm{L}^2(\Omega)} \left\|f\right\|_{\mathrm{H}^{-1}(\Omega)} \\ &\leqslant \left\|A^{-1/2}\right\|_{\mathrm{L}^\infty(\Omega)} \left\|A^{1/2}\nabla u\right\|_{\mathrm{L}^2(\Omega)} \left\|f\right\|_{\mathrm{H}^{-1}(\Omega)},\end{aligned} \tag{6.6}$$

and the result follows. □

We continue by showing that the difference between the true solution u and the approximate solution to the regularised problem U_ε can be controlled a posteriori by the sum of a standard residual based term and an extra term that depends on the error committed by introducing the approximate diffusion tensor.

Proposition 6.3.4 (A posteriori result for U_ε)**.** *Let u and U_ε solve respectively* (6.2)*, and* (6.4)*. Let $e = u - U_\varepsilon$. Then we have the following a* *posteriori bound.*

$$\left\|A^{1/2}\nabla e\right\|^2_{\mathrm{L}^2(\Omega)} \leqslant C \left\|A^{-1/2}\right\|_{\mathrm{L}^\infty(\Omega)} \sum_{K\in\mathscr{T}} (\eta_K^2 + \gamma_K^2), \tag{6.7}$$

where

$$\begin{aligned}\eta_K^2 &= h_K^2 \left\|f + \nabla\cdot(A_\varepsilon \nabla U_\varepsilon)\right\|^2_{\mathrm{L}^2(K)} + \frac{1}{2} h_K \left\|[\![A_\varepsilon \nabla U_\varepsilon]\!]\right\|^2_{\mathrm{L}^2(\partial K)}\\ \gamma_K^2 &= \left\|(A_\varepsilon - A)\nabla U_\varepsilon\right\|^2_{\mathrm{L}^2(K)}\end{aligned} \tag{6.8}$$

Remark 6.3.5. *The result of proposition 6.3.4 bounds the error $u - U_\varepsilon$ in terms of a standard residual term and a data approximation, or regularisation term. We assume that our regularisation A_ε is such that the latter is controlled by ε in the sense that*

$$\sum_{K\in\mathscr{T}} \gamma_K^2 \leqslant C\varepsilon^\beta, \tag{6.9}$$

and therefore

$$\left\|A^{1/2}\nabla e\right\|^2_{\mathrm{L}^2(\Omega)} \leqslant \mathcal{O}(\varepsilon^\beta + h^2). \tag{6.10}$$

By sending $\varepsilon \to 0$ at an appropriate rate, we can ensure the usual order h convergence for this problem. This is investigated further in the numerical examples below, see figure 6.1.

Proof of proposition 6.3.4. Due to the inconsistency in the discrete problem, we do not have the usual Galerkin orthogonality. Instead, we have

$$\langle A\nabla(u - U_\varepsilon), \nabla\Phi\rangle = \langle (A_\varepsilon - A)\nabla U_\varepsilon, \nabla\Phi\rangle \;\; \forall \Phi \in \mathcal{V}_h. \tag{6.11}$$

We can therefore write

$$
\begin{aligned}
\left\|A^{1/2}\nabla e\right\|^2_{\mathrm{L}^2(\Omega)} &= \langle A\nabla(u-U_\varepsilon), \nabla e\rangle \\
&= \langle A\nabla(u-U_\varepsilon), \nabla(e-\Phi)\rangle + \langle (A_\varepsilon - A)\nabla U_\varepsilon, \nabla\Phi\rangle .
\end{aligned}
\tag{6.12}
$$

Using (6.2), this becomes

$$
\begin{aligned}
\left\|A^{1/2}\nabla e\right\|^2_{\mathrm{L}^2(\Omega)} &= \langle f, e-\Phi\rangle - \langle A\nabla U_\varepsilon, \nabla(e-\Phi)\rangle \\
&\quad + \langle (A_\varepsilon - A)\nabla U_\varepsilon, \nabla\Phi\rangle \\
&= \langle f, e-\Phi\rangle - \langle A_\varepsilon\nabla U_\varepsilon, \nabla(e-\Phi)\rangle \\
&\quad + \langle (A_\varepsilon - A)\nabla U_\varepsilon, \nabla e\rangle .
\end{aligned}
\tag{6.13}
$$

Upon integrating by parts on each element, we obtain

$$
\begin{aligned}
\left\|A^{1/2}\nabla e\right\|^2_{\mathrm{L}^2(\Omega)} = \sum_{K\in\mathscr{T}} &\langle f + \nabla\cdot(A_\varepsilon\nabla U_\varepsilon), e-\Phi\rangle_K - \langle [\![A_\varepsilon\nabla U_\varepsilon]\!], e-\Phi\rangle_{\partial K} \\
&+ \langle (A_\varepsilon - A)\nabla U_\varepsilon, \nabla e\rangle_K
\end{aligned}
\tag{6.14}
$$

An application of the Cauchy-Schwarz inequality gives

$$
\begin{aligned}
\left\|A^{1/2}\nabla e\right\|^2_{\mathrm{L}^2(\Omega)} &\leqslant \sum_{K\in\mathscr{T}} \|f + \nabla\cdot(A_\varepsilon\nabla U_\varepsilon)\|_{\mathrm{L}^2(K)} \|e-\Phi\|_{\mathrm{L}^2(K)} \\
&\quad + \sum_{e\subseteq K} \|[\![A_\varepsilon\nabla U_\varepsilon]\!]\|_{\mathrm{L}^2(e)} \|e-\Phi\|_{\mathrm{L}^2(e)} \\
&\quad + \sum_{K\in\mathscr{T}} \|(A_\varepsilon - A)\nabla U_\varepsilon\|_{\mathrm{L}^2(K)} \|\nabla e\|_{\mathrm{L}^2(K)} \\
&\quad := R_1 + R_2 + R_3.
\end{aligned}
\tag{6.15}
$$

We now set Φ to be the Clément interpolant of e in R_1 and R_2 and bound using theorem 2.39:

$$\begin{aligned} R_1 &\leqslant C_{\text{clem}} \sum_{K\in\mathscr{T}} h_K \left\| f + \nabla \cdot (A_\varepsilon \nabla U_\varepsilon) \right\|_{\mathrm{L}^2(K)} \left\| \nabla e \right\|_{\mathrm{L}^2(K)} \\ &\leqslant C_{\text{clem}} \left(\sum_{K\in\mathscr{T}} h_K^2 \left\| f + \nabla \cdot (A_\varepsilon \nabla U_\varepsilon) \right\|_{\mathrm{L}^2(K)}^2 \right)^{1/2} \left(\sum_{K\in\mathscr{T}} \left\| \nabla e \right\|_{\mathrm{L}^2(K)}^2 \right)^{1/2} \\ &= C_{\text{clem}} \left(\sum_{K\in\mathscr{T}} h_K^2 \left\| f + \nabla \cdot (A_\varepsilon \nabla U_\varepsilon) \right\|_{\mathrm{L}^2(K)}^2 \right)^{1/2} \left\| \nabla e \right\|_{\mathrm{L}^2(\Omega)}, \end{aligned} \tag{6.16}$$

where C_{clem} is the interpolation constant in theorem 2.39 that depends upon the shape regularity of the mesh. Similarly,

$$\begin{aligned} R_2 &\leqslant C_{\text{clem}} \sum_{K\in\mathscr{T}} h_K^{1/2} \sum_{e\subseteq K} \left\| [\![A_\varepsilon \nabla U_\varepsilon]\!] \right\|_{\mathrm{L}^2(e)} \left\| \nabla e \right\|_{\mathrm{L}^2(\hat{K})} \\ &\leqslant C_{\text{clem}} \left(\sum_{K\in\mathscr{T}} h_K \left\| [\![A_\varepsilon \nabla U_\varepsilon]\!] \right\|_{\mathrm{L}^2(\partial K)}^2 \right)^{1/2} \left(\sum_{K\in\mathscr{T}} \left\| \nabla e \right\|_{\mathrm{L}^2(\hat{K})}^2 \right)^{1/2} \\ &\leqslant C_{\text{clem}} C_{\mathscr{T}} \left(\sum_{K\in\mathscr{T}} h_K \left\| [\![A_\varepsilon \nabla U_\varepsilon]\!] \right\|_{\mathrm{L}^2(\partial K)}^2 \right)^{1/2} \left\| \nabla e \right\|_{\mathrm{L}^2(\Omega)}. \end{aligned} \tag{6.17}$$

Here, $C_{\mathscr{T}}$ is a constant that quantifies the overlap of element patches, that is, depending upon the shape regularity of the mesh. For example, on a mesh consisting of uniform squares, $C_{\mathscr{T}} = 9$. The result now follows after bounding R_3 with the Cauchy-Schwarz inequality and noting that due to the coercivity of the problem, we have

$$\left\| \nabla e \right\|_{\mathrm{L}^2(\Omega)} \leqslant \left\| A^{-1/2} \right\|_{\mathrm{L}^\infty(\Omega)} \left\| A^{1/2} \nabla e \right\|_{\mathrm{L}^2(\Omega)}. \tag{6.18}$$

We finally have

$$
\begin{aligned}
\left\|A^{1/2}\nabla e\right\|_{\mathrm{L}^2(\Omega)} \leqslant & C_{\mathrm{clem}} \left\|A^{-1/2}\right\|_{\mathrm{L}^\infty} \left(\sum_{K\in\mathscr{T}} h_K^2 \left\|f + \nabla\cdot(A_\varepsilon \nabla U_\varepsilon)\right\|^2_{\mathrm{L}^2(K)}\right)^{1/2} \\
& + C_{\mathrm{clem}} C_{\mathscr{T}} \left\|A^{-1/2}\right\|_{\mathrm{L}^\infty} \left(\sum_{K\in\mathscr{T}} h_K \left\|[\![A_\varepsilon \nabla U_\varepsilon]\!]\right\|^2_{\mathrm{L}^2(\partial K)}\right)^{1/2} \\
& + \left\|A^{-1/2}\right\|_{\mathrm{L}^\infty} \left(\sum_{K\in\mathscr{T}} \left\|(A_\varepsilon - A)\nabla U_\varepsilon\right\|_{\mathrm{L}^2(K)}\right)^{1/2}
\end{aligned}
\tag{6.19}
$$

□

Remark 6.3.6. *The result of proposition 6.3.4 gives some insight into how the inconsistency affects the numerical approximation. In particular, the difference $A_\varepsilon - A$ is weighted by gradients of the approximate solution. We can therefore expect this term to have a significant effect in problems with large gradients.*

Remark 6.3.7 (Definition of energy norm)**.** *Proposition 6.3.4 gives a bound on the error in an energy norm for (6.2), that is, the norm given by $v \mapsto \left\|A^{1/2}\nabla v\right\|_{\mathrm{L}^2(\Omega)}$. One could also consider the error measured by a different choice of energy, namely $v \mapsto \left\|A_\varepsilon^{1/2}\nabla v\right\|_{\mathrm{L}^2(\Omega)}$. This is the subject of the next proposition. We shall see that using this alternative notion of energy leads to a bound which is not directly computable, and depends on less well behaved constants.*

Proposition 6.3.8 (Alternative error bound)**.** *Under the same hypotheses as proposition 6.3.4, we also have, for $0 < C, C(f, A, A_\varepsilon)$,*

$$
\left\|A_\varepsilon^{1/2}\nabla e\right\|_{\mathrm{L}^2} \leqslant C(f, A, A_\varepsilon) \left\|f\right\|_{\mathrm{H}^{-1}(\Omega)} + C_{\mathscr{T}} C_{clem} \left\|\right\|_{\mathrm{L}^\infty(\Omega)} \sum_{K\in\mathscr{T}} \eta_K^2. \tag{6.20}
$$

where

$$C(f, A, A_\varepsilon) = \left\| A^{-1/2} A_\varepsilon^{-1/2} (A_\varepsilon - A) \right\|_{\mathrm{L}^\infty(\Omega)} \left\| A^{-1/2} \right\|_{\mathrm{L}^\infty(\Omega)} \tag{6.21}$$

Proof. In analogy with proposition 6.3.4, we begin by writing, with $\Phi \in \mathcal{V}$,

$$\begin{aligned} \left\| A_\varepsilon^{1/2} \nabla e \right\|_{\mathrm{L}^2(\Omega)}^2 &= \langle A_\varepsilon \nabla (u - U_\varepsilon), \nabla e \rangle \\ &= \langle A \nabla u - A_\varepsilon \nabla U_\varepsilon, \nabla e \rangle + \langle (A_\varepsilon - A) \nabla u, \nabla e \rangle \\ &= \langle A \nabla u - A_\varepsilon \nabla U_\varepsilon, \nabla (e - \Phi) \rangle + \langle (A_\varepsilon - A) \nabla u, \nabla e \rangle \end{aligned} \tag{6.22}$$

The first term can be bounded analogously as in proposition 6.3.4 to see that

$$\begin{aligned} \langle A \nabla u - A_\varepsilon \nabla U_\varepsilon, \nabla (e - \Phi) \rangle \leqslant \\ \sum_{K \in \mathscr{T}} \langle f + \nabla \cdot (A_\varepsilon \nabla U_\varepsilon), e - \Phi \rangle_K - \langle [\![A_\varepsilon \nabla U_\varepsilon]\!], e - \Phi \rangle_{\partial K} \end{aligned} \tag{6.23}$$

which is precisely the first two terms on the right hand side of (6.14), and is bounded in exactly the same way.

For the second, we can obtain an estimate by using the stability result:

$$\begin{aligned} \langle (A_\varepsilon - A) \nabla u, \nabla e \rangle &= \left\langle A^{-1/2} A_\varepsilon^{-1/2} (A_\varepsilon - A) A^{1/2} \nabla u, A_\varepsilon^{1/2} \nabla e \right\rangle \\ &\leqslant \left\| A^{-1/2} A_\varepsilon^{-1/2} (A_\varepsilon - A) \right\|_{\mathrm{L}^\infty(\Omega)} \left\| A^{1/2} \nabla u \right\|_{\mathrm{L}^2(\Omega)} \left\| A_\varepsilon^{1/2} \nabla e \right\|_{\mathrm{L}^2(\Omega)} \\ &\leqslant \left\| A^{-1/2} A_\varepsilon^{-1/2} (A_\varepsilon - A) \right\|_{\mathrm{L}^\infty(\Omega)} \left\| A^{-1/2} \right\|_{\mathrm{L}^\infty(\Omega)} \left\| f \right\|_{\mathrm{H}^{-1}(\Omega)} \left\| A_\varepsilon^{1/2} \nabla e \right\| \end{aligned} \tag{6.24}$$

We note that this is not directly computable due to the negative Sobolev norm that appears, and must be approximated (as is done for example in [67]). □

6.4 A parabolic model problem

In this section we aim to prove a result analogous to proposition 6.3.4 for evolution problems. We show that, at least in the case where k is dependent on space only, an a posteriori bound that makes use of proposition 6.3.4 can be derived using the elliptic reconstruction technique. In this section, we work in semidiscrete form, that is, discretised in space but not time. This is a common approach when analysing finite element methods for parabolic problems (e.g. [105, 70]) and reduces the technicality for clarity of presentation. Such arguments are not diffucult to extend to the fully discrete case, and the reader is referred to for example [31, 104] for examples of fully discrete analysis.

6.4.1 Error estimates using the elliptic reconstruction

We now consider a parabolic problem and its regularisation analogous to (6.2)-(6.4). We consider a model linear parabolic problem, together with its regularisation, and a finite element approximation to the regularised problem. In weak form, they are given by

$$\langle \partial_t u, \varphi \rangle + \langle k(x) \nabla u, \nabla \varphi \rangle = 0 \quad \forall \varphi \in \mathrm{H}^1_0(\Omega), \tag{6.25}$$

$$\langle \partial_t u_\varepsilon, \varphi \rangle + \langle k_\varepsilon(x) \nabla u_\varepsilon, \nabla \varphi \rangle = 0 \quad \forall \varphi \in \mathrm{H}^1_0(\Omega), \tag{6.26}$$

$$\langle \partial_t U_\varepsilon, \Phi \rangle + \langle k_\varepsilon \nabla U_\varepsilon, \nabla \Phi \rangle = 0 \quad \forall \Phi \in \mathcal{V}_h. \tag{6.27}$$

As in the elliptic case, we remark that (6.25) is *exact*, (6.26) is its regularisation and (6.27) is the finite element approximation of a pertrbed parabolic problem.

Theorem 6.4.1. *Let u be the solution of* (6.25)*, with U_ε the solution of* (6.27)*. Let $e = u - U_\varepsilon$. Then*

$$\int_0^T \left\|k^{1/2}\nabla e\right\|^2 \mathrm{d}t + \|e(T)\|^2_{\mathrm{L}^2(\Omega)} \leqslant \|e(0)\|^2_{\mathrm{L}^2(\Omega)} + 2\int_0^T \sum_{K\in\mathscr{T}} (\bar\eta_K^2 + \gamma_K^2)\,\mathrm{d}t, \tag{6.28}$$

where γ_K is as defined in proposition 6.3.4, and $\bar\eta_K$ is defined to be

$$\bar\eta_K^2 = h_K^2 \left\|\partial_t U_\varepsilon - \nabla\cdot(A_\varepsilon \nabla U_\varepsilon)\right\|^2_{\mathrm{L}^2(K)} \tag{6.29}$$

Before proving theorem 6.4.1, we take the opportunity to introduce key tools in the finite element analysis of parabolic problems.

Definition 6.4.2. *Discrete elliptic operator. For any $\chi \in \mathcal{V}_h$, $\mathfrak{A}_\varepsilon\chi$ is defined to be the element of $\mathcal{V}_h$ such that*

$$-\left\langle \mathfrak{A}_\varepsilon\chi, \Phi\right\rangle = \left\langle k_\varepsilon\nabla\chi, \nabla\Phi\right\rangle \quad \forall\Phi\in\mathcal{V}_h. \tag{6.30}$$

Definition 6.4.3 (Elliptic reconstruction.)**.** *We define the operator $\mathcal{R} : \mathcal{V}_h \to \mathrm{H}^1(\Omega)$ as follows. For an element $\chi \in \mathcal{V}_h$, we define the elliptic reconstruction $\mathcal{R}\chi$ by*

$$\left\langle k\nabla\mathcal{R}\chi, \nabla v\right\rangle = -\left\langle \mathfrak{A}_\varepsilon\chi, v\right\rangle \quad \forall v\in \mathrm{H}^1_0(\Omega). \tag{6.31}$$

Remark 6.4.4 (Relationship between U_ε and its elliptic reconstruction.)**.** *The elliptic reconstruction as defined above allows us to use the a posteriori result given in proposition 6.3.4. To see this, we note that equation (6.31) frames the elliptic reconstruction as the solution to a variational problem with data in $\mathcal{V}_h$. After regularising and discretising this problem in exactly the same way as §6.3, the result is precisely equation (6.30) viewed as a discrete problem for U_ε.*

Proof of theorem 6.4.1. We use the elliptic reconstruction approach. We therefore begin by noting that U_ε solves the following problem in $\mathrm{H}^1_0(\Omega)$:

$$\langle \partial_t U_\varepsilon - \mathfrak{A}_\varepsilon U_\varepsilon, \varphi \rangle = 0 \quad \forall \varphi \in H_0^1(\Omega). \tag{6.32}$$

By the definition of the elliptic reconstruction, we write this as

$$\langle \partial_t U_\varepsilon, \varphi \rangle + \langle k_\varepsilon \nabla \mathcal{R} U_\varepsilon, \nabla \varphi \rangle = 0 \quad \forall \varphi \in H_0^1(\Omega). \tag{6.33}$$

Combined with (6.25), this yields an error equation

$$\langle \partial_t (u - U_\varepsilon), \varphi \rangle + \langle k \nabla (u - \mathcal{R} U_\varepsilon), \nabla \varphi \rangle = 0. \tag{6.34}$$

Now, with $\rho := u - \mathcal{R} U_\varepsilon$ and $\omega := \mathcal{R} U_\varepsilon - U_\varepsilon$, we can choose $\varphi = e = u - U_\varepsilon$ to obtain

$$\langle e_t, e \rangle + \langle k \nabla \rho, \nabla (\rho + \omega) \rangle = 0. \tag{6.35}$$

Now, integrating (6.35) over the interval $[0, T]$ gives

$$\begin{aligned} \frac{1}{2} \left(\|e(T)\|^2_{\mathrm{L}^2(\Omega)} - \|e(0)\|^2_{\mathrm{L}^2(\Omega)} \right) + \int_0^T \left\| k^{1/2} \nabla \rho \right\|^2_{\mathrm{L}^2(\Omega)} \mathrm{d}t = \\ - \int_0^T \langle k \nabla \rho, \nabla \omega \rangle \,\mathrm{d}t. \end{aligned} \tag{6.36}$$

Applying elementary inequalities gives

$$\begin{aligned} \int_0^T \left\| k^{1/2} \nabla \rho \right\|^2 \mathrm{d}t + \|e(T)\|^2_{\mathrm{L}^2(\Omega)} \leqslant \frac{1}{2} \|e(0)\|^2_{\mathrm{L}^2(\Omega)} \\ + \frac{1}{2} \int_0^T \left\| k^{1/2} \nabla \rho \right\|^2_{\mathrm{L}^2(\Omega)} \mathrm{d}t + \frac{1}{2} \int_0^T \left\| k^{1/2} \nabla \omega \right\| \mathrm{d}t \end{aligned} \tag{6.37}$$

i.e.,

$$\int_0^T \left\|k^{1/2}\nabla\rho\right\|^2 \mathrm{d}t + \|e(T)\|^2_{\mathrm{L}^2(\Omega)} \leqslant \|e(0)\|^2_{\mathrm{L}^2(\Omega)} + \int_0^T \left\|k^{1/2}\nabla\omega\right\|^2_{\mathrm{L}^2(\Omega)} \mathrm{d}t. \tag{6.38}$$

Since we have bounded ρ in terms of ω, we may now use the triangle inequality to write

$$\int_0^T \left\|k^{1/2}\nabla e\right\|^2 \mathrm{d}t + \|e(T)\|^2_{\mathrm{L}^2(\Omega)} \leqslant \|e(0)\|^2_{\mathrm{L}^2(\Omega)} + 2\int_0^T \left\|k^{1/2}\nabla\omega\right\|^2_{\mathrm{L}^2(\Omega)} \mathrm{d}t. \tag{6.39}$$

□

Remark 6.4.5. *This becomes a computable bound in light of remark 6.4.4. Indeed, in proposition 6.3.4, the quantity $\left\|k^{1/2}\nabla\omega\right\|^2_{\mathrm{L}^2(\Omega)}$ is bounded a posteriori.*

6.5 Numerical examples - elliptic case

We now investigate the behaviour of the error estimate derived in proposition 6.3.4 using a numerical example. We consider problem (6.2)-(6.4) on $\Omega = (-1,1)^2$ with the following choice for the exact and approximate diffusion tensors:

$$A = \begin{pmatrix} k(2x) & 0 \\ 0 & 1 \end{pmatrix} \tag{6.40}$$

$$A_\varepsilon = \begin{pmatrix} k_\varepsilon(2x) & 0 \\ 0 & 1 \end{pmatrix}, \tag{6.41}$$

where k_ε is the approximation defined in equation (5.38) for the hydraulic conductivity model for clay defined in §4.6.1. We enforce homogeneous Dirichlet boundary conditions on $\partial\Omega$ and set $f \equiv 1$. A plot of the solution is shown in figure 6.1. The rate of convergence of the approximation error $\|k - k_\varepsilon\|_{\mathrm{L}^2([-\varepsilon,0])}$ is numerically determined to be approximately of order $\varepsilon^{0.63}$. We therefore couple regularisation to mesh refinement by selecting $\varepsilon \propto h^{1.6}$ to ensure that the approximation error does not pollute the convergence rate of the finite element method. Indeed, if $\varepsilon \propto h^{1.6}$, then we should have $\|k - k_\varepsilon\|_{\mathrm{L}^2([-\varepsilon,0])} \propto (h^{1.6})^{0.63}$, that is, the data approximation term and residual terms have the same asymptotic order in equation (6.8). See also remark 6.3.5. Results are shown in figure 6.1. This figure shows that the order h convergence is preserved and that, with the correct coupling, the approximation error in A does not adversely affect the convergence rate of the finite element method (figure 6.1b).

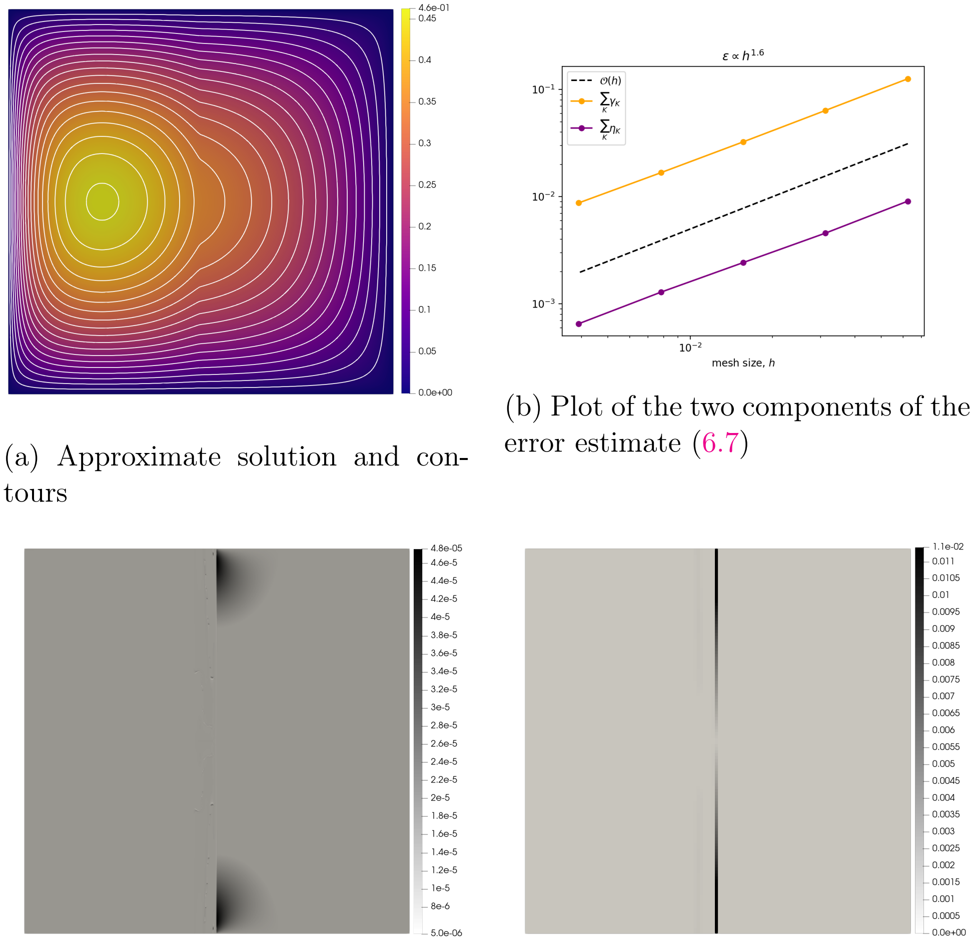
6.6 Numerical examples - nonlinear parabolic case

Motivated by the results of section 6.4, we define the following error indicators.

$$E_{K,n} := \int_{\partial K} [\![k(U^n)\nabla(U^n + h_z)]\!] \tag{6.42}$$

$$\vartheta_n := \left\| k(U^n)\nabla(U^n + z) - k(U^{n-1})\nabla(U^{n-1} + h_z) \right\|_{\mathrm{L}^2(\Omega)}, \tag{6.43}$$

$$\gamma_n := \|(k_\varepsilon(U^n) - k(U^n))\nabla(U^n + h_z)\|_{\mathrm{L}^2(\Omega)}, \tag{6.44}$$



(a) Approximate solution and contours

(b) Plot of the two components of the error estimate (6.7)

(c) Plot of η_k

(d) Plot of γ_k

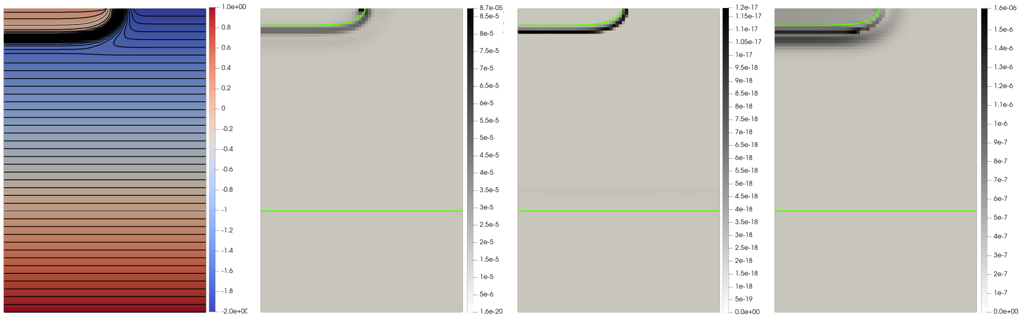
Figure 6.1: Figures 6.1c and 6.1d show the spatial distribution of the error indicators.

where ϑ_n and γ_n are localised to cells in the obvious way. $E_{K,n}$ represents spatial contribution to the error, ϑ the temporal contribution, and γ_n represents the effect of regularising the hydraulic conductivity. We remark that in this case the data approximation term is weighted by $\nabla(U^n + h_z)$. It therefore approximates the difference in flux of water that results by making the approximation in the coefficient k.

As in example 1, §6.5, we couple regularisation to mesh refinement by selecting $\varepsilon \propto h^{1.6}$ and run a sequence of simulations of the aquifer recharge problem in clay soil §5.4.5. The discretisation parameters for the simulations are $\tau = 0.02$ days, and $h = \left(\frac{1}{2}\right)^{i+1}$, $i = 1, 2, 3, 4$ with $\varepsilon = 1.25{*}\left(\frac{1}{2}\right)^{i-1}$. We keep the time step fixed as our primary focus is on capturing spatial phenomena. We also run the simulations on the smaller time interval $t \in [0, 15]$ since the key phenomena in this test case occur earlier (i.e. the initial infiltrating front and the join with the water table). Visualisations of the error indicators (6.42)-(6.44) are shown in figures 6.2b-6.2d. The plots show that the spatial indicator is localised to the wetting front, as well as the singularity at the land surface caused by change of boundary condition, the data indicator lies just ahead of the front where the regularisation takes effect, and the temporal indicator is concentrated ahead of the front where the solution changes most rapidly.

The behaviour of the error indicators as h and ε are decreased is shown in figure 6.3.

The results shown in figure 6.3 present, as expected, a more complex picture than in the linear case. In figures 6.3b and 6.3c we see that the shape of the graph of the indicators γ and E changes as the $\varepsilon \to 0$. This is because smaller ε results in steeper gradients in the solution. We do however observe approximately the expected orders of decrease for both indicators.



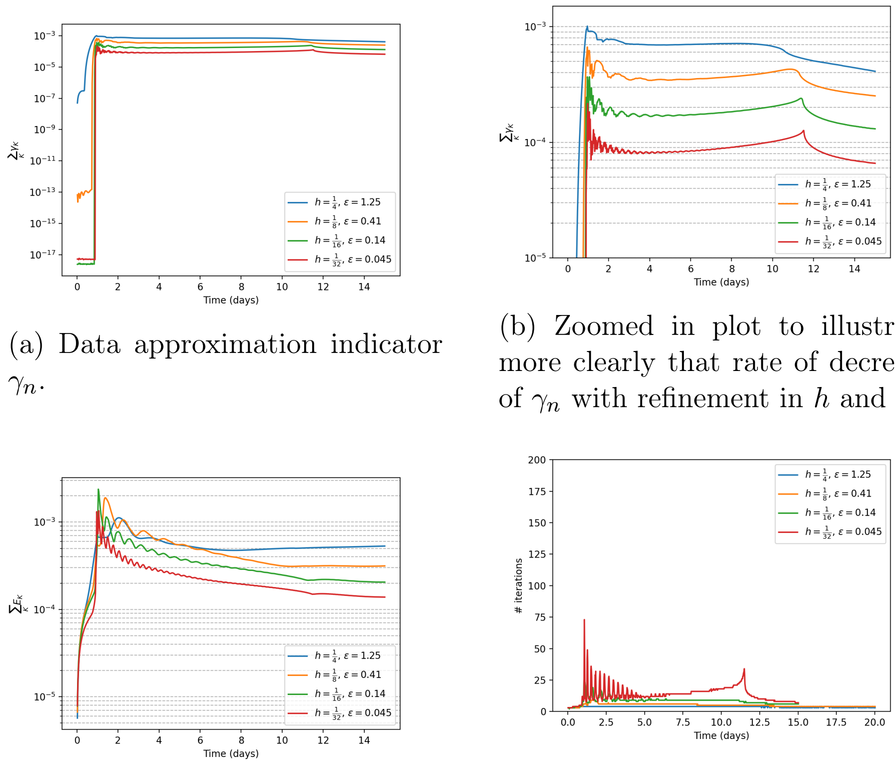
(a) Pressure field (b) $E_{K,n}$ (c) γ_n (d) ϑ_n

Figure 6.2: Spatial distributions of error indicators, $t = 2$ days. Level set $u = 0$ indicated with green line.

6.7 Conclusions & discussion

In this chapter, error estimates were derived for linear elliptic and linear parabolic problems to take into account the approximation error resulting from regularisation of the problem data. In the elliptic case, we note that the correct choice of energy is crucial for obtaining useful bounds, compare propositions 6.3.8 and 6.3.4. In the latter case, the constants depend on the essential suprema of $A^{-1/2}$ and $A_\varepsilon^{-1/2}$, which will blow up in degenerate cases. The resulting bounds provided the appropriate coupling of regularisation with mesh refinement to ensure that asymptotic rates of convergence are maintained despite the approximations made in the coefficients. In addition, when applied to Richards' equation, a stable sequence of approximations to regularised problems was obtained, with the behaviour of the error indicators suggesting that optimal convergence rates are preserved.

We remark that in figure 6.3d we still see a degree of instability in the nonlinear iteration, and that convergence is not guaranteed by our scheme. We saw however in §5 that our regularisation does help the scheme to converge, and the data in figure 6.3 suggests that the data approximation term is not likely to pollute the overall simulation error. Design of a scheme that is



(a) Data approximation indicator γ_n.

(b) Zoomed in plot to illustrate more clearly that rate of decrease of γ_n with refinement in h and ε

(c) Spatial indicator E_n.

(d) Iterations required to reach convergence at each time step.

Figure 6.3: Plots of the indicators and iteration data for the numerical experiment §6.6.

guaranteed to converge for a given regularisation level is a subject for further research.

Chapter 7

Conclusions, discussion and future work

In this thesis, we have explored the application of a posteriori error analysis and adaptive schemes for finite element methods.

In this short chapter, we summarise the main results and discuss future directions for development of the work in this thesis.

7.1 Summary of results

Following a review of literature in chapter 1, we discussed the background material and presented some key known results, most importantly an a posteriori error bound for a finite element method for a linear elliptic problem. The bound, given in theorem 2.6.2 is in $\mathrm{L}^p(\Omega)$ for general p and we made use of duality arguments, a key tool in later chapters.

The first novel results appeared in chapter 3, where a rigorous a posteriori bound analogous to that given in theorem 2.6.2 was proved for the Signorini problem with $p = 4$. This work was in part inspired by new insights from the paper [34] where limitations on dual regularity for the Signorini problem were discussed. A crucial part of the proof is a result on the approximation properties of a new bound-preserving interpolation operator that we proposed specifically for this problem. We proved that the approximation error of this interpolant converges at the optimal rate with respect to the finite element mesh size as well as preserving key inequalities.

This bound was benchmarked using a suite of numerical test problems and its practical performance was evaluated. Adaptive meshes generated using the local error indicators (3.52) as refinement criteria were shown to out-perform uniform grids in terms of lower error and/or improved rates of error reduction in terms of number of degrees of freedom in the mesh (see figures 3.4a and 3.4b respectively).

We also tested on problems that did not have the necessary regularity to apply our theoretical results (§3.8.2 and §3.8.3), since the work in this thesis was motivated by practical applications where such assumptions may not be met, and observed good performance, in the case of §3.8.3, optimal rates of convergence with respect to degrees of freedom were not reached, but rates were improved by using an adaptive scheme, see particularly figure 3.5.

As a final note on this work, we point to the assumption on the discrete contact set, condition $\mathbf{A_h}$ (see remark 3.4.2). This assumption essentially relates to resolution of the contact set by the discrete solution, and is required to prove stability of the dual problem (which depends on the discrete contact set). For our numerical examples, this assumption is rather mild as the solution has simple topology, however we must acknowledge that for more pathological examples, this may require a long pre-asymptotic regime in convergence behaviour until the required resolution is achieved. How this would affect the error estimate is unknown, and is a potential subject of further investigation.

In chapter 4, we moved on to a more realistic seepeage problem related to the Signorini problem via its nonlinear boundary conditions. Study of this problem was motivated by data collected by our industrial partners CEMADEN, as well as the aim of developing our earlier work on the Signorini problem to more realistic problems.

This time we did not have necessary regularity in the dual problem to apply the same methodology as in chapter 3. Instead, we developed an error estimate using the dual weighted residual methodology, our contribution being the extension of the result to a problem with nonlinear diffusion coefficient, and extensive numerical tests and case studies.

Once again, we tested the error estimate with a range of simulations. The estimate was seen to perform very well in the standard seepage problem setup (see §4.7.1 and figure 4.7).

The most interesting numerical examples were the case studies of §4.8, where real data is combined with geolocial information to simulate wells in São Paulo State, Brazil. These test cases were made particularly difficult by the presence of inhomogeneity in the subsurface structure with differences of several orders of magnitude in hydraulic conductivity between layers of soil/ rock. As a result, we found that high levels of local refinement were required to obtain accurate result. In particular, figure 4.11 shows that very high resolution was required before the computation of flux across the seepage face reached a stable value. These test cases also showed some of the limitations of our work. It is expected that measurement error and lack of detail in the subsurface structure will contribute to significant computation errors.

In chapters 5 and 6, we move on to Richards' equation: a time-dependent model of infiltration. Our focus here was to study the instabilities that are observed when simulating Richards' equation with certain difficult soil models (see the numerical results of chapter 5, particularly figure 5.8). In §5.5 we introduce a regularisation for the hydraulic conductivity that can be controlled by a positive parameter ε, with $\varepsilon = 0$ corresponding to the exact model and larger positive values corresponding to reduced steepness at the saturation point. In figure 5.8 we see that regularising the hydraulic conductivity in this manner does indeed stabilise the iterative solver used to solve Richards' equation, and reduce convergence failures. This of course comes with a penalty of being further from the true model, the effect of which is demonstrated in figure 5.5, where we display pressure fields that result from different values of ε.

The discussion in chapter 5 was rather heuristic and qualitative, so in chapter 6, we aim to quantify the effect of our regularisation on numerical errors, and quantify the relative convergence rates of discretisation and regularisation error. To this end, we develop an a posteriori error bound for a linear elliptic problem in proposition 6.3.4 that takes account of the error induced by approximating the coefficient. The result is the augmentation of a residual error bound with a term including the data error weighted by the gradient of the numerical solution. This bound is further developed into a bound for the parabolic version using the elliptic reconstruction (see theorem 6.4.1).

The error estimates developed in chapter 6 are benchmarked under uniform grid refinement and demonstrate the expected rates of convergence.

7.2 Further work

As previously noted, more accurate problem data is required for accurate simulation of the subsurface well problems studied in §4.8. It is a subject of ongoing research to extract as much useful information as possible from data provided by resistivity methods for subsurface measurement, explored in [6]. Further work could also make use of efficient numerical routines such as the adaptive schemes of chapter 4 as the forward model in an inverse problem to infer subsurace structure. Since forward models must be solved many times, efficiency is cricual, and adaptive methods come into their own.

In chapter 6, analysis was provided only for a linear problem. Further research could investigate the analysis of Richards' equation or a similar nonlinear problem, including the possible extension of the elliptic reconstruction technique to parabolic problems with nonlinearity in the time derivative.

The combination of local adaptive mesh refinement and regularisation, possibly combined with adaptivity in the time step is an interesting avenue of future research. In this case, one must be careful to ensure that the regularisation parameter does not become too small for the current mesh to handle. In this thesis, we provided an improvement in iteration stability using regularisation, but an ideal outcome would be a full space-time adaptive algorithm that could provide efficient solution of Richards' equation in challenging scenarios, and guarantee convergence of nonlinear iterative solvers.

Bibliography

[1] R. A. Adams and J. J. Fournier. *Sobolev spaces*. Elsevier, 2003.

[2] M. Ainsworth and J. T. Oden. *A posteriori error estimation in finite element analysis*. Vol. 37. John Wiley & Sons, 2011.

[3] H. W. Alt and S. Luckhaus. "Quasilinear elliptic-parabolic differential equations". In: *Mathematische Zeitschrift* 183.3 (1983), pp. 311–341.

[4] T. Apel and S. Nicaise. "Regularity of the solution of the scalar Signorini problem in polygonal domains". In: *Results in Mathematics* 75.2 (2020), pp. 1–15.

[5] D. Arndt et al. "The `deal.II` Library, Version 9.1". In: *Journal of Numerical Mathematics* (2019). accepted. DOI: 10.1515/jnma-2019-0064. URL: https://dealii.org/deal91-preprint.pdf.

[6] B. Ashby, T. Pryer, A. Lukyanov, and C. Bortolozo. *Towards an operational landslide prediction mechanism.* Springer (forthcoming), 2022.

[7] B. Ashby, C. Bortolozo, A. Lukyanov, and T. Pryer. "Adaptive modelling of variably saturated seepage problems". In: *The Quarterly Journal of Mechanics and Applied Mathematics* 74.1 (2021), pp. 55–81.

[8] W. Bangerth and R. Rannacher. *Adaptive finite element methods for differential equations*. Birkhäuser, 2013.

[9] W. Bangerth and R. Rannacher. “Finite element approximation of the acoustic wave equation: Error control and mesh adaptation”. In: *East West Journal of Numerical Mathematics* 7.4 (1999), pp. 263–282.

[10] R. E. Bank. “Hierarchical bases and the finite element method”. In: *Acta numerica* 5 (1996), pp. 1–43.

[11] G. R. Barrenechea, V. John, and P. Knobloch. “Finite element methods respecting the discrete maximum principle for convection-diffusion equations”. In: *arXiv preprint arXiv:2204.07480* (2022).

[12] M. Bause and P. Knabner. “Computation of variably saturated subsurface flow by adaptive mixed hybrid finite element methods”. In: *Advances in Water Resources* 27.6 (2004), pp. 565–581. ISSN: 0309-1708. DOI: https://doi.org/10.1016/j.advwatres.2004.03.005. URL: http://www.sciencedirect.com/science/article/pii/S0309170804000600.

[13] J. Bear. *Dynamics of fluids in porous media*. Dover, 1988. ISBN: 0486656756.

[14] R. Becker and R. Rannacher. “An optimal control approach to a posteriori error estimation in finite element methods”. In: *Acta numerica* 10 (2001), pp. 1–102.

[15] R. Becker and R. Rannacher. *Weighted a posteriori error control in FE methods*. IWR, 1996.

[16] A. Bergam, C. Bernardi, and Z. Mghazli. “A posteriori analysis of the finite element discretization of some parabolic equations”. In: *Mathematics of computation* 74.251 (2005), pp. 1117–1138.

[17] H. Blum and F. Suttmeier. “An adaptive finite element discretisation for a simplified Signorini problem”. In: *Calcolo* 37.2 (2000), pp. 65–77. ISSN: 1126-5434. DOI: 10.1007/s100920070008. URL: https://doi.org/10.1007/s100920070008.

[18] D. Braess. “A posteriori error estimators for obstacle problems–another look”. In: *Numerische Mathematik* 101.3 (2005), pp. 415–421.

[19] K. Brenner, D. Hilhorst, and H. C. V. Do. “A gradient scheme for the discretization of Richards equation”. In: *Finite Volumes for Complex Applications VII-Elliptic, Parabolic and Hyperbolic Problems*. Springer, 2014, pp. 537–545.

[20] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. Springer Science & Business Media, 2008. ISBN: 978-0-387-75933-3. DOI: 10.1007/978-0-387-75934-0. URL: http://linkinghub.elsevier.com/retrieve/pii/S0898122103900617%7B%5C%%7D5Cnhttp://link.springer.com/10.1007/978-0-387-75934-0.

[21] H. Brezis. “Sur une nouvelle formulation du problème de l’écoulement à travers une digue”. In: *CR Acad. Sci. Paris* 287 (1978), pp. 711–714.

[22] H. Brézis. *Functional analysis. Theory and applications.(Analyse fonctionnelle. Théorie et applications.). Collection Mathématiques Appliquées pour la Maîtrise*. 1994.

[23] F. Brezzi and G. Sacchi. “A finite element approximation of a variational inequality related to hydraulics”. In: *Calcolo* 13.3 (1976), pp. 257–273.

[24] F. Brezzi, W. W. Hager, and P.-A. Raviart. “Error estimates for the finite element solution of variational inequalities”. In: *Numerische Mathematik* 28.4 (1977), pp. 431–443.

[25] R. Brooks. “Corey, Hydraulic Properties of Porous Media”. In: *Colorado University, Fort Collins* (1964).

[26] I. Bubuška and M. Vogelius. “Feedback and adaptive finite element solution of one-dimensional boundary value problems”. In: *Numerische Mathematik* 44.1 (1984), pp. 75–102.

[27] E. Buckingham. “Studies on the movement of soil moisture”. In: *US Dept. Agic. Bur. Soils Bull.* 38 (1907), pp. 28–36.

[28] M. Camporese, C. Paniconi, M. Putti, and S. Orlandini. “Surface-subsurface flow modeling with path-based runoff routing, boundary condition-based coupling, and assimilation of multisource observation data”. In: *Water Resources Research* 46.2 (2010).

[29] C. Carstensen and R. Verfürth. “Edge residuals dominate a posteriori error estimates for low order finite element methods”. In: *SIAM journal on numerical analysis* 36.5 (1999), pp. 1571–1587.

[30] M. A. Celia, E. T. Bouloutas, and R. L. Zarba. “A general mass-conservative numerical solution for the unsaturated flow equation”. In: *Water resources research* 26.7 (1990), pp. 1483–1496.

[31] Z. Chen and J. Feng. “An adaptive finite element algorithm with reliable and efficient error control for linear parabolic problems”. In: *Mathematics of computation* 73.247 (2004), pp. 1167–1193.

[32] Z. Chen and R. H. Nochetto. “Residual type a posteriori error estimates for elliptic obstacle problems”. In: *Numerische Mathematik* 84.4 (Feb. 2000), pp. 527–548. ISSN: 0945-3245. DOI: 10.1007/s002110050009. URL: https://doi.org/10.1007/s002110050009.

[33] Z. Chen and R. H. Nochetto. “Residual type a posteriori error estimates for elliptic obstacle problems”. In: *Numerische Mathematik* 84.4 (2000), pp. 527–548.

[34] C. Christof and C. Haubner. “Finite element error estimates in non-energy norms for the two-dimensional scalar Signorini problem”. In: *Numerische Mathematik* 145.3 (2020), pp. 513–551.

[35] P. G. Ciarlet. *The finite element method for elliptic problems.* SIAM, 2002.

[36] P. Clément. “Approximation by finite element functions using local regularization”. In: *Revue française d’automatique, informatique, recherche opérationnelle. Analyse numérique* 9.R2 (1975), pp. 77–84.

[37] K. A. Cliffe, J. Collis, and P. Houston. “Goal-oriented a posteriori error estimation for the travel time functional in porous media flows”. In: *SIAM Journal on Scientific Computing* 37.2 (2015), B127–B152.

[38] R. L. Cooley. “Some new procedures for numerical solution of variably saturated flow problems”. In: *Water Resources Research* 19.5 (1983), pp. 1271–1285.

[39] M. Darbandi, S. Torabi, M. Saadat, Y. Daghighi, and D. Jarrahbashi. “A moving-mesh finite-volume method to solve free-surface seepage problem in arbitrary geometries”. In: *International Journal for Numerical and Analytical Methods in Geomechanics* 31.14 (2007), pp. 1609–1629.

[40] A. Dedner, J. Giesselmann, T. Pryer, and J. K. Ryan. “Residual estimates for post-processors in elliptic problems”. In: *Journal of Scientific Computing* 88.2 (2021), pp. 1–28.

[41] A. Demlow and E. H. Georgoulis. “Pointwise a posteriori error control for discontinuous Galerkin methods for elliptic problems”. In: *SIAM Journal on Numerical Analysis* 50.5 (2012), pp. 2159–2181.

[42] P. Deuflhard, P. Leinen, and H. Yserentant. “Concepts of an adaptive hierarchical finite element code”. In: *IMPACT of Computing in Science and Engineering* 1.1 (1989), pp. 3–35.

[43] R. A. DeVore. “Nonlinear approximation”. In: *Acta numerica* 7 (1998), pp. 51–150.

[44] W. Dörfler. “A Convergent Adaptive Algorithm for Poisson’s Equation”. In: *SIAM J. Numer. Anal.* 33.3 (1996), pp. 1106–1124. DOI: 10.1137/0733054. URL: https://doi.org/10.1137/0733054.

[45] K. Eriksson. “An adaptive finite element method with efficient maximum norm error control for elliptic problems”. In: *Mathematical Models and Methods in Applied Sciences* 4.03 (1994), pp. 313–329.

[46] K. Eriksson and C. Johnson. “Adaptive Finite Element Methods for Parabolic Problems I: A Linear Model Problem”. In: *SIAM J. Numer. Anal.* 28.1 (1991), pp. 43–77. ISSN: 00361429. URL: http://www.jstor.org/stable/2157933.

[47] A. Ern and J. Guermond. *Finite Elements I: Approximation and interpolation*. Springer, 2021.

[48] A. Ern and J.-L. Guermond. *Theory and practice of finite elements*. Vol. 159. Springer, 2004.

[49] L. C. Evans. *Partial differential equations*. Vol. 19. American Mathematical Soc., 2010.

[50] R. Eymard, M. Gutnic, and D. Hilhorst. “The finite volume method for Richards equation”. In: *Computational Geosciences* 3.3-4 (1999), pp. 259–294.

[51] R. S. Falk. “Error estimates for the approximation of a class of variational inequalities”. In: *Mathematics of Computation* 28.128 (1974), pp. 963–971.

[52] P. Grisvard. *Elliptic problems in nonsmooth domains*. SIAM, 1985.

[53] R. Hartmann and P. Houston. “Adaptive discontinuous Galerkin finite element methods for the compressible Euler equations”. In: *Journal of Computational Physics* 183.2 (2002), pp. 508–532.

[54] R. Hartmann. “Adaptive Finite Element Methods for the Compressible Euler Equations”. PhD thesis. University of Heidelberg, 2002.

[55] R. Haverkamp, M. Vauclin, J. Touma, P. Wierenga, and G. Vachaud. “A comparison of numerical simulation models for one-dimensional infiltration 1”. In: *Soil Science Society of America Journal* 41.2 (1977), pp. 285–294.

[56] A. Hawkins-Daarud, K. G. van der Zee, and J. Tinsley Oden. “Numerical simulation of a thermodynamically consistent four-species tumor growth model”. In: *International journal for numerical methods in biomedical engineering* 28.1 (2012), pp. 3–24.

[57] P. Hild and S. Nicaise. “A posteriori error estimations of residual type for Signorini’s problem”. In: *Numerische Mathematik* 101.3 (2005), pp. 523–549.

[58] R. H. Hoppe and R. Kornhuber. “Adaptive multilevel methods for obstacle problems”. In: *SIAM journal on numerical analysis* 31.2 (1994), pp. 301–323.

[59] O. Ippisch, H. J. Vogel, and P. Bastian. “Validity limits for the van Genuchten–Mualem model and implications for parameter estimation and numerical simulation”. In: *Advances in water resources* 29.12 (2006), pp. 1780–1789.

[60] J. Jackaman and T. Pryer. “Conservative Galerkin methods for dispersive Hamiltonian problems”. In: *Calcolo* 58.3 (2021), pp. 1–36.

[61] W. Jäger and J. Kačur. “Solution of porous medium type systems by linear approximation schemes”. In: *Numerische Mathematik* 60.1 (1991), pp. 407–427.

[62] M. J. Kazemzadeh-Parsi and F. Daneshmand. “Unconfined seepage analysis in earth dams using smoothed fixed grid finite element method”. In: *International Journal for Numerical and Analytical Methods in Geomechanics* 36.6 (2012), pp. 780–797.

[63] D. W. Kelly, J. P. De S. R. Gago, O. C. Zienkiewicz, and I. Babuška. “A posteriori error analysis and adaptive processes in the finite element method: Part I–Error Analysis”. In: *Int. J. Num. Meth. Engrg.* 19 (1983), pp. 1593–1619.

[64] D. Kinderlehrer and G. Stampacchia. *An introduction to variational inequalities and their applications*. Vol. 31. SIAM, 1980.

[65] D. Kinderlehrer and G. Stampacchia. *An introduction to variational inequalities and their applications*. SIAM, 2000.

[66] R. Krause, A. Veeser, and M. Walloth. “An efficient and reliable residual-type a posteriori error estimator for the Signorini problem”. In: *Numerische Mathematik* 130.1 (2015), pp. 151–197.

[67] O. Lakkis and T. Pryer. “Gradient recovery in adaptive finite-element methods for parabolic problems”. In: *IMA Journal of Numerical Analysis* 32.1 (2012), pp. 246–278.

[68] J.-L. Lions and G. Stampacchia. “Variational inequalities”. In: *Communications on pure and applied mathematics* 20.3 (1967), pp. 493–519.

[69] F. List and F. A. Radu. “A study on iterative methods for solving Richards’ equation”. In: *Comput. Geosci.* 20.2 (2016), pp. 341–353. ISSN: 1420-0597. DOI: 10.1007/s10596-016-9566-3. URL: http://link.springer.com/10.1007/s10596-016-9566-3.

[70] C. Makridakis and R. H. Nochetto. “Elliptic Reconstruction and a Posteriori Error Estimates for Parabolic Problems”. In: *SIAM J. Numer. Anal.* 41.4 (2004), pp. 1585–1594. ISSN: 00361429. URL: http://www.jstor.org/stable/4101184.

[71] E. Milakis and L. Silvestre. “Regularity for the nonlinear Signorini problem”. In: *Advances in Mathematics* 217.3 (2008), pp. 1301–1312.

[72] U. Mosco. “Error estimates for some variational inequalities”. In: *Mathematical Aspects of Finite Element Methods*. Springer, 1977, pp. 224–236.

[73] U. Mosco and G. Strang. “One-sided approximation and variational inequalities”. In: *Bulletin of the American Mathematical Society* 80.2 (1974), pp. 308–312.

[74] U. Mosco and G. Strang. “One-sided approximation and variational inequalities”. In: *Bulletin of the American Mathematical Society* 80.2 (1974), pp. 308–312.

[75] Y. Mualem. “A new model for predicting the hydraulic conductivity of unsaturated porous media”. In: *Water resources research* 12.3 (1976), pp. 513–522.

[76] S. P. Neuman. “Saturated-unsaturated seepage by finite elements.” In: *Journal of the Hydraulics Division., PROC., ASCE.* (1973), pp. 2233–2250.

[77] J. Nitsche. “L^∞-convergence of finite element approximations”. In: *Mathematical aspects of finite element methods*. Springer, 1977, pp. 261–274.

[78] R. H. Nochetto, A. Veeser, and M. Verani. “A safeguarded dual weighted residual method”. In: *IMA Journal of Numerical Analysis* 29.1 (2008), pp. 126–140.

[79] R. Nochetto, A. Schmidt, and C. Verdi. “A posteriori error estimation and adaptivity for degenerate parabolic problems”. In: *Mathematics of computation* 69.229 (2000), pp. 1–24.

[80] R. H. Nochetto. “Pointwise a posteriori error estimates for elliptic problems on highly graded meshes”. In: *Mathematics of computation* 64.209 (1995), pp. 1–22.

[81] R. H. Nochetto, M. Paolini, and C. Verdi. “An adaptive finite element method for two-phase Stefan problems in two space dimensions. II: Implementation and numerical experiments”. In: *SIAM journal on scientific and statistical computing* 12.5 (1991), pp. 1207–1244.

[82] R. H. Nochetto, K. G. Siebert, and A. Veeser. “Pointwise a posteriori error control for elliptic obstacle problems”. In: *Numerische Mathematik* 95.1 (July 2003), pp. 163–195. ISSN: 0945-3245. DOI: 10.1007/s00211-002-0411-3. URL: https://doi.org/10.1007/s00211-002-0411-3.

[83] J. T. Oden and N. Kikuchi. “Theory of variational inequalities with applications to problems of flow through porous media”. In: *International Journal of Engineering Science* 18.10 (1980), pp. 1173–1284.

[84] C. Paniconi and M. Putti. “A comparison of Picard and Newton iteration in the numerical solution of multidimensional variably saturated flow problems”. In: *Water Resources Research* 30.12 (1994), pp. 3357–3374.

[85] H. Q. Pham, D. G. Fredlund, and S. L. Barbour. “A study of hysteresis models for soil-water characteristic curves”. In: *Canadian Geotechnical Journal* 42.6 (2005), pp. 1548–1568.

[86] M. Picasso. “Adaptive finite elements for a linear parabolic problem”. In: *Computer Methods in Applied Mechanics and Engineering* 167.3-4 (1998), pp. 223–237.

[87] R. Rannacher. “Adaptive finite element methods in flow computations”. In: *Recent Advances in Adaptive Computation. Contemporary Mathematics* 383 (2004), pp. 183–176.

[88] L. A. Richards. “Capillary conduction of liquids through porous media”. In: *Phys. 1* 1.May (1931), pp. 318–333.

[89] L. A. Richards. “The usefulness of capillary potential to soil moisture and plant investigators”. In: *J. Ag. Res.* 37.12 (1928), pp. 719–742.

[90] J. J. Rulon, R. Rodway, and R. A. Freeze. “The development of multiple seepage faces on layered slopes”. In: *Water Resources Research* 21.11 (1985), pp. 1625–1636.

[91] F. Scarpini and M. A. Vivaldi. “Error estimates for the approximation of some unilateral problems”. In: *RAIRO. Analyse numérique* 11.2 (1977), pp. 197–208.

[92] A. Schmidt and K. Siebert. *Design of adaptive finite element software.* 2005.

[93] L. R. Scott and S. Zhang. “Finite element interpolation of nonsmooth functions satisfying boundary conditions”. In: *Mathematics of computation* 54.190 (1990), pp. 483–493.

[94] C. Scudeler, C. Paniconi, D. Pasetto, and M. Putti. “Examination of the seepage face boundary condition in subsurface and coupled surface/subsurface hydrological models”. In: *Water Resources Research* 53.3 (2017), pp. 1799–1819.

[95] F. I. Siddiqui and S. B. A. B. S. Osman. “Simple and multiple regression models for relationship between electrical resistivity and various soil properties for soil characterization”. In: *Environ. Earth Sci.* 70.1 (Sept. 2013), pp. 259–267. ISSN: 1866-6299. DOI: 10.1007/s12665-012-2122-0. URL: https://doi.org/10.1007/s12665-012-2122-0.

[96] G. Şimşek, X. Wu, K. van der Zee, and E. van Brummelen. “Duality-based two-level error estimation for time-dependent PDEs: Application to linear and nonlinear parabolic equations”. In: *Computer Methods in Applied Mechanics and Engineering* 288 (2015), pp. 83–109.

[97] M. Slodicka. “A robust and efficient linearization scheme for doubly nonlinear and degenerate parabolic problems arising in flow in porous media”. In: *SIAM Journal on Scientific Computing* 23.5 (2002), pp. 1593–1614.

[98] S. L. Sobolev. *Applications of functional analysis in mathematical physics*. Vol. 7. American Mathematical Society, 1963.

[99] P. Solin, D. Andrs, J. Cerveny, and M. Simko. “PDE-independent adaptive hp-FEM based on hierarchic extension of finite element spaces”. In: *Journal of computational and applied mathematics* 233.12 (2010), pp. 3086–3094.

[100] O. Steinbach, B. Wohlmuth, and L. Wunderlich. “Trace and flux a priori error estimates in finite-element approximations of Signorni-type problems”. In: *IMA Journal of Numerical Analysis* 36.3 (July 2015), pp. 1072–1095. ISSN: 0272-4979. DOI: 10.1093/imanum/drv039. eprint: https://academic.oup.com/imajna/article-pdf/36/3/1072/6767945/drv039.pdf. URL: https://doi.org/10.1093/imanum/drv039.

[101] G. Strang. “One-sided approximation and plate bending”. In: *Computing Methods in Applied Sciences and Engineering Part 1*. Springer, 1974, pp. 140–155.

[102] G. Strang. “The dimension of piecewise polynomial spaces, and one-sided approximation”. In: *Conference on the Numerical Solution of Differential Equations*. Springer. 1974, pp. 144–152.

[103] F.-T. Suttmeier. *Numerical solution of variational inequalities by adaptive finite elements*. Springer, 2008.

[104] O. J. Sutton. “Long-time $L^\infty(L^2)$ a posteriori error estimates for fully discrete parabolic problems”. In: *IMA Journal of Numerical analysis* 40.1 (2020), pp. 498–529.

[105] V. Thomée. *Galerkin finite element methods for parabolic problems*. Vol. 25. Springer Science & Business Media, 2007.

[106] M. Van Genuchten. “A Closed-form Equation for Predicting the Hydraulic Conductivity of Unsaturated Soils”. In: *Soil Sci. Soc. Am. J.* 44 (1980), pp. 892–898. DOI: 10.2136/sssaj1980.03615995004400050002x.

[107] M. Van Genuchten and D. Nielsen. "On describing and predicting the hydraulic properties". In: *Annales Geophysicae* 3.5 (1985), pp. 615–628.

[108] A. Veeser. "Approximating gradients with continuous piecewise polynomial functions". In: *Foundations of Computational Mathematics* 16.3 (2016), pp. 723–750.

[109] A. Veeser. "Positivity preserving gradient approximation with linear finite elements". In: *Computational Methods in Applied Mathematics* 19.2 (2019), pp. 295–310.

[110] R. Verfürth. *A posteriori error estimation techniques for finite element methods*. OUP Oxford, 2013.

[111] T. Vogel, M. Van Genuchten, and M. Cislerova. "Effect of the shape of the soil hydraulic functions near saturation on variably-saturated flow predictions". In: *Advances in water resources* 24.2 (2000), pp. 133–144.

[112] M. Walloth. "A reliable, efficient and localized error estimator for a discontinuous Galerkin method for the Signorini problem". In: *Applied Numerical Mathematics* 135 (2019), pp. 276–296.

[113] T. K. Weber, W. Durner, T. Streck, and E. Diamantopoulos. "A modular framework for modeling unsaturated soil hydraulic properties over the full moisture range". In: *Water Resources Research* 55.6 (2019), pp. 4994–5011.

[114] A. Weiss and B. I. Wohlmuth. "A posteriori error estimator for obstacle problems". In: *SIAM Journal on Scientific Computing* 32.5 (2010), pp. 2627–2658.

[115] O. Wilderotter. "An adaptive finite element method for singular parabolic equations". In: *Numerische Mathematik* 96.2 (Dec. 2003), pp. 377–399. ISSN: 0945-3245. DOI: 10.1007/s00211-003-0463-z. URL: https://doi.org/10.1007/s00211-003-0463-z.

[116] C. Zhang. “Adaptive finite element methods for variational inequalities: Theory and applications in finance”. PhD thesis. 2007.

[117] J. Zhang, Q. Xu, and Z. Chen. “Seepage analysis based on the unified unsaturated soil theory”. In: *Mechanics Research Communications* 28.1 (2001), pp. 107–112.

[118] H. Zheng, D. Liu, C. Lee, and L. Tham. “A new formulation of Signorini’s type for seepage problems with free surfaces”. In: *International Journal for Numerical Methods in Engineering* 64.1 (2005), pp. 1–16.