The Pennsylvania State University
The Graduate School

ANALYSIS AND ROBUST PRECONDITIONING FOR NUMERICAL IMPLEMENTATIONS OF RICHARDS' EQUATION IN GROUNDWATER FLOW

A Dissertation in
Mathematics
by
Juan Batista

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

December 2019

The dissertation of Juan Batista was reviewed and approved* by the following:

Ludmil Zikatanov
Professor of Mathematics
Dissertation Co-Advisor
Co-Chair of Committee

Anna Mazzucato
Professor of Mathematics
Dissertation Co-Advisor
Co-Chair of Committee

Xiantao Li
Professor of Mathematics

Corina Drapaca
Professor of Earth and Engineering Sciences

Mark Levi
Professor of Mathematics
Department Head

*Signatures are on file in the Graduate School.

Abstract

This thesis serves as a mathematical and numerical exploration of Richards' equation, a quasilinear partial differential equation modeling the flow of nearly incompressible fluid through unsaturated porous media, with degeneracies at full saturation. This physical model can be seen as a reduction of a full two phase (wetting/air phase) flow model, where the air phase is assumed to have constant atmospheric pressure. In this case, the air pressure only affects the pressure of the wetting phase via the hydraulic conductivity of the porous matrix through capillary action, which is in turn modeled as a function of saturation of the water phase. We discuss various formulations of Richards' equation, and popular models for the water content θ and hydraulic conductivity K as functions of the pressure head. For the ubiquitous VGM model, we describe some analytic properties of the physical parameters.

Due to the nonlinear nature of the problem, closed-form solutions do not exist, except in some special cases. As such, numerical treatment is required to approximate solutions for physically relevant problems. In this thesis we consider several linearization schemes used to treat the nonlinearities, including the Picard, Newton-Raphson, modified Picard, and the L-scheme. For a time-continuous Picard linerization, we were able to prove that under a technical assumption on the behavior of the nonlinearities, the sequence of solutions to the linearized problem is a contractive sequence, thus guaranteeing convergence of the iterates to the solution of the nonlinear problem. The need for control of K and θ in the analysis is confirmed by various numerical results reported in the literature in which the Picard linearization of Richards' equation fail to converge due to the nonlinearities.

For the resulting sequence of parabolic problems, we discretize with the implicit Euler method in time, and a mixed finite element discretization in space (lowest order Raviart-Thomas elements). For the efficient solution of the resulting linear systems during the iterations we introduce a combined preconditioner: an inexact Uzawa iteration paired with an auxiliary space preconditioner using the standard linear continuous Lagrange finite element space. We prove that the preconditioned system has a uniformly bounded condition number. The combined preconditioner for the symmetric linearizations is robust with respect to discretization parameters, and jumps in the conductivity, though the convergence theory of these linear schemes to a weak solution for problems with layered media is a more complicated matter, and was not the focus of this work. We present several numerical tests veryfying the theoretical results. Additionally, we present numerical results for the nonsymmetric linear problems arising from the Newton-Raphson linearization, and in some cases we observe that the preconditioners are robust with respect to discretization parameters and the nonlinear physical parameters K and θ.

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my two advisors, Ludmil Zikatanov and Anna Mazzucato, for their unending patience, guidance, and the uncountable hours they spent working with me on my research, I could not have done this without them. I would also like to thank our collaborators at Tufts university, Xiaozhe Hu and James Adler, for many productive discussions and guidance on the particulars regarding implementation, in particular on getting our preconditioner working in HAZmath. I would also like to thank my committee members for their time. Finally, I would also like to thank my friends and family for their invaluable emotional support through the ups and downs of graduate student life, especially my mother and father, Edna and Roberto Batista, my best friend, Sergio Rodriguez, and my girlfriend, Zhiyao Jia.

This material is based upon work supported by the NSF under Award No. DMS-1720114 and DMS-1819157 (P.I. Ludmil Zikatanov), and by Award No. DMS-1615457 and DMS-1909103 (P.I. Anna Mazzucato). Any opinions, findings, and conclusions or recommendations expressed in this publication are those of the author, and do not necessarily reflect the views of the NSF.

Chapter 1
Introduction

Richards' equation is one of many nonlinear models of the flow of an incompressible fluid through a porous medium. This model, first introduced by Lorenzo Richards in 1931, is used extensively in the physical sciences to model fluid flow through unsaturated porous materials. The equations in their various forms were originally intended to be used for predicting and analysing the flow of water through various soils, and while most applications are still hydrology-related, such as modeling aquifer recharge rates due to rainwater [2], the effects of carbon injection into deep groundwater aquifers [3], and the tracking of contaminants flowing through soil [4], the model has also found use for more general purpose porous media applications, such as enhanced oil recovery [5], the flow of groundwater through root-soil systems [6], and even diaper mechanics [7].

This thesis consists of two parts, the first of which serves as an introduction to the problem. In the first chapter, we discuss the derivation of Richards flow in groundwater as a simplification of a two-phase flow model, and discuss the work of Forsyth [8], who investigated the viability of the Richards simplification as opposed to the full two phase model. We introduce the various forms of Richards' equation, and discuss various parameter models for the nonlinear hydraulic conductivity K and water content θ as functions of pressure head Ψ, that have been used historically. We focus on the Van Genuchten-Mualem (VGM) model for K and θ, and discuss some analytic properties of these nonlinear functions. The second chapter serves as an introduction to the mathematical question of well-posedness of the constitutive Cauchy problem, which has been shown to have a unique weak solution given some mild assumptions on the initial and boundary data. In particular, Alt and Luckhaus [9] employ a non-standard strategy to prove existence of weak solutions to a general quasilinear elliptic-parabolic PDE that contains Richards' equation, and chapter two serves as a brief exposition of their method for the sake of completeness for the thesis. The question of

uniqueness of weak solutions was resolved in [9] for a specific class of solutions, and for specific forms of θ and K (after application of the Kirchoff transformation, as expanded upon in section 2.1.2). In a later work, F. Otto [10] was able to prove uniqueness of weak solutions without these restrictions, so we summarize some of these results as well.

In chapter three, we introduce various linearization schemes used in the literature to yield a sequence of linear problems that can be discretized in time and space to be solved. One of these linearization schemes is a time-continuous Picard iteration, for which we are able to give sufficient conditions on K and θ in order to prove that the linearized operator is contractive, with a time-independent rate, implying the convergence of this Picard iteration to the weak solution of the fully nonlinear problem. We also focus on some of the experiments of Paniconi and Putti [11], whose findings on robustness of the Picard linearization match with the intuition that we developed in our proof, that low order linearizations will only converge globally if K and θ are mild enough. We also survey various numerical papers that work with two linearizations, the modified Picard [12] and L-scheme [13] methods; in particular, Slodička [14] was able to show that after a Backward Euler discretization in time, the L-scheme linearization for each resulting elliptic problem can be interpreted as a contractive operator that yields the solution to the corresponding nonlinear variational problem for large enough L-scheme iterations, with a rate dependent on the minimal value of K, the relaxation parameter L, the timestep τ, and geometric properties of the domain, and is able to show that a piecewise-linear-in-time interpolation of these nonlinear elliptic problems converges weakly to the weak solution of the nonlinear problem.

Part two focuses on numerical implementation for the problem. In chapter four, we first discretize the problem in time using Backward Euler, and discretize in space with the finite element method. We focus on mixed finite element discretizations, which have some prevalence in the numerical literature of Richards' equation ([15], [16], [15], [17], [18]). We also consider a monotonicity-preserving, edge-based formulation of the standard $P1$-Lagrange finite element discretization, known as Edge Averaged Finite Elements [19] (EAFE). We also briefly mention results regarding the stability of these finite element discretizations, relying on a proof by Radu et al [20] that guarantees stability of the mixed finite element discretization for Holder continuous θ, with error estimates dependent on the Holder exponent $\alpha \in (0, 1]$. In a previous work [15], they also show that stability of the mixed finite element discretization after applying an integration technique pioneered in [5] implies stability of the $P1$-Lagrange formulation, with the solution converging with the same order of accuracy.

The rest of chapter four describes basic properties of two Krylov-subspace based iterative

solvers, Generalized Min Res and Conjugate Gradient, and also discusses multigrid, which we used as both preconditioner and solver in several parts of our numerical schemes. Finally, chapter five focuses on the preconditioner for the saddle point systems that are generated after fully discretizing with mixed finite elements. This preconditioner is composed of two main steps: an algorithm for preconditioning the indefinite saddle point system which we refer to as the Schur iteration, and an auxiliary space preconditioner for the approximate Schur complement solve done in the pressure correction solve in the Schur iteration. For symmetric linearizations (all the ones we consider but Newton-Raphson), we are able to prove the uniformity of this combined preconditioner with respect to discretization parameters, and are able to show this uniformity via numerical experiments. We also show some numerical experiments where our scheme also uniformly preconditions Newton-Raphson linearized problems, in situations where the nonsymmetric part is mild. This preconditioner is novel for Richards' equation, and its robustness with respect to K and θ for the more popular symmetric linearizations makes it a viable addition to many black-box Richards simulators used in various popular codes used to simulate unsaturated groundwater flow. Another important feature of this preconditioner is its scalability, as the number of degrees of freedom in the full saddle point system is much larger than that of the approximate Schur complement solve in each Schur iteration, which is in turn has about $d!$ times more unkowns than the problem discretized in the nodal auxiliary space, with d being the spatial dimension of the finite elements considered. As such, the auxiliary space preconditioner achieves a reduction in problem size of $\frac{1}{6}$ for the $3D$ simulations we consider, and for problems with more regular initial data and K and θ, the theoretical result we prove also applies to space-time elements, which would imply a problem-size reduction of $\frac{1}{24}$. We then close with some conclusions, and directions for this work to be continued.

The rest of this chapter focuses on the derivation and formulation of Richards' equation. In section 1.1, we discuss the modeling of unsaturated groundwater flow as a full two phase model, in which one can measure the changes in pressure of the air and water phase. We show Richards' equation as a simplification of the two phase model, where the air pressure is taken as constant, and nonlinear effects on water compressibility and fluid viscosity are ignored. The rest of that section is a review of the work of Forsyth et al [8], in comparing the accuracy and performance of the one phase model versus the two phase model on various simulations. Section 1.2 defines the various forms of Richards' equation that are used when modeling unsaturated groundwater flow, namely the pure pressure head formulation, the mixed pressure head-water content formulation, and the pure saturation formulation. We

focus on the mixed pressure head-water content formulation, as it is in conservation form, which guarantees global mass conservation in linearization schemes used for simulations. Section 1.3 gives a brief overview of some models used for the water content, $\theta(\Psi)$, and the hydraulic conductivity, $K(\theta(\Psi))$, as functions of pressure head Ψ. We formulate the VGM [21] model, and discuss some analytic properties of these parameters. Finally, section 1.4 rounds out the chapter by showing some figures of these parameters as functions of pressure head from $\Psi = -100$cm to $\Psi = 0$cm for various materials.

1.1 Modeling unsaturated groundwater flow

The flow in unsaturated soils can be modeled as a two-phase flow of immiscible fluids, namely the air and water phases. The physical processes that give rise to this kind of flow are infiltration of surface water through the upper layers of soil which enriches the soil moisture, and subsurface flow through soils which are partially filled with air. The interaction of roots of plants with this flow, and the advection and dispersion of fertilizers and pesticides within the unsaturated zone make this model of considerable interest to soil scientists, agronomists, and irrigation engineers. Unsaturated flow also captures the interest of environmental engineers whose main concern would be predicting the infiltration and subsequent distribution of contaminants and pollutants from industrial processes, including seepage into underground streams, and for petroleum engineers, in particular in the cases of underground reservoirs where immiscible fluid flows are also encountered, with the fluids being water, oil, and gas.

Multiphase flow is governed by a system of coupled mass balance laws, one for each phase. In the case of modeling the flow of water into unsaturated soil, we have two fluid phases: air(which we denote with subscript a) and water(which we denote with subscript w). This gives the following system of equations:

$$\partial_t(\phi\theta_w\rho_w) + \operatorname{div}(\rho_w V_w) = q_w,$$
$$\partial_t(\phi\theta_a\rho_a) + \operatorname{div}(\rho_a V_a) = q_a,$$

with Darcy velocity of each phase l being proportional to the pressure gradient, and negative sign indicating the tendency of fluid to flow from higher pressure values to lower ones:

$$V_l = -\boldsymbol{\kappa}\lambda_l(\nabla P_l + \rho_l g e_z).$$

Here, θ_l is the saturation of phase l, P_l is its pressure, ρ_l is the mass density of phase l, $\boldsymbol{\kappa}$ is

the permeability tensor, $\lambda_l = \frac{k_l}{\mu_l}$, k_{rl} is the relative permeability of phase l, μ_l is the viscosity of phase l, and ge_z is the gravitational acceleration vector in the positive z direction. q_l are the source terms for phase l.

To link the phase pressures, one can use an experimentally determined capillary pressure, P_{caw}:

$$P_a = P_w + P_{caw}(S_w). \tag{1.1}$$

If the air phase pressure is not assumed to be constant, we have

$$\theta_w + \theta_a = 1,$$

which allows us to eliminate the air saturation from the system.

Typically, the following models are assumed for the density of air and water phases, and porosity of the medium:

$$\begin{aligned}
\rho_a &= \rho_{a0} P_a / P_{a0} \\
\rho_w &= \rho_{w0}(1 + c_w(P_w - P_{w0})) \\
\phi &= \phi_0(1 + c_m(P_w - P_{w0})),
\end{aligned}$$

where c_m is the compressibility of the soil, c_w is the compressibility of the water, and ρ_{l0} is the mass density of phase l at base pressures P_{w0} and P_{a0}, respectively. Typically, the compressibility coefficients are very small (on the order of 10^{-7} kPa), and so we will ignore their effects, so that all but the air density becomes constant. Then,

$$\begin{aligned}
&\phi \rho_w \partial_t \theta_w - \operatorname{div}\left(K_w(\theta_w)(\nabla P_w + \rho_w g e_z)\right) = q_w, \\
&\phi \frac{\rho_{a0}}{P_{a0}} \partial_t \left[(1 - \theta_w) P_a\right] - \operatorname{div}\left(K_a(\theta_w)(\nabla P_a + \frac{\rho_{a0}}{P_{a0}} P_a g e_z)\right) = q_a.
\end{aligned}$$

Here we have consolidated the phase permeabilities and viscosities into the conductivity terms, $K_l = \boldsymbol{\kappa} \lambda_l(\theta_l)$.

In the full two phase formulation, the relative conductivities for both water and air are determined from field experiments, as nonlinear functions of water saturation.

1.1.1 Constant air pressure (Richards) approximation

In the case of groundwater flow, it is usual to assume that the dynamics of air flow play an insignificant role in determining water movement and storage in the unsaturated zone, i.e,

for flows near the surface, one assumes the air pressure to be atmospheric (constant). This eliminates the second equation entirely, and further simplifies the water pressure equation, as one can use (1.1) to solve $\theta_w = \theta(P_w)$, by inverting P_{cwa}. Typically, the water pressure is replaced with the pressure head, which is a more direct experimental measure. To relate the two, the standard relation is $\Psi = \frac{P_w}{\rho_w g}$. Thus, the system of equations can be reduced to the single equation,

$$C\partial_t\theta(\Psi) - \operatorname{div}\left(K(\theta(\Psi))(\nabla\Psi + e_z)\right) = q_w.$$

In doing so, we have significantly reduced the number of unknowns to solve for, at the cost of introducing harsher nonlinearities in the elliptic term.

1.1.1.1 Simplified example: linear flow in one dimension

The full two phase model has developed increasing popularity to model subsurface groundwater flow for certain situations. Some authors have even questioned the assumption of a passive air phase ([22], [23]).

As was discussed in [22], by assuming one dimensional flow along an infinite depth column with no gravity or source/sink terms, and that the pressure in air changes are small, then the two phase system can be approximated by a nonlinear hyperbolic system in water saturation θ_w, where the fluxes V_l are constant:

$$\partial_t(\phi\theta_w) = K(V_a + V_w)\partial_x(f_w),$$

and the fractional flow curve f_w is used to formulate a condition for physically admissable solutions,

$$f_w = \frac{k_{rw}(\theta_w)/\mu_{rw}}{\frac{k_{rw}(\theta_w)}{\mu_{rw}} + \frac{k_{ra}(\theta_w)}{\mu_{ra}}}.$$

When considering initial profiles for this simplified system of the form

$$\theta_w = \begin{cases} 1, & x \leq 0, \\ 0, & x \geq 0, \end{cases}$$

an entropy condition can be formulated that requires physically admissable solutions to obey a convexity constraint with respect to f_w, namely that the line connecting the state on the right to the state on the left must lie above f_w. In cases of f_w where this is valid, the shock

of unit height at $x = 0$ propagates to the right. However, in one of the cases considered ($k_{rl} = \theta_l^4$), the chord intersects at a height less than one, hence the system forms a rarefaction fan with values $\theta_s < \theta_w < 1$, follwed by a shock of height θ_s.

The simplified single phase flow model, on the other hand, cannot capture this rarefaction, as it completely ignores the effects of air pressure, hence only unit height shocks would propagate, with no rarefaction.

1.1.1.2 Behavior in higher dimensions

The model problem above, while indicative of potential discrepancies between two phase flow and the one phase simplification, is only one-dimensional, which permits use of the simplifying assumption of constant total fluid velocity due to the scalar nature of V_a and V_w, and hence requires solution of only one equation for both approaches. This simplification is not possible in more than one dimension.

To consider the multidimensional case, Forsyth tested the validity and performance of two phase flow versus the one phase simplification in various numerical experiments ([8], [4], [24]). We wish to highlight two multidimensional tests done in [8]: a standard dam seepage problem (steady state), where the goal is to determine whether the height of the seepage point (the highest point where the dam wall is fully saturated on the opposite side of the water front) differs when using two phase flow versus one phase. In this test, a linear approximation of the capillary pressure was used,

$$P_{caw}(\theta_w) = P_c'(1 - \theta_w).$$

In the two phase formulation, $P_c' = 0$. For the single phase formulation, $P_c' = 1$kpa was used, and the simulations predict the same seepage height for both cases. As the nonlinearities for the two phase formulation considered are much milder than the one phase formulation, the number of linear iterations required to solve the two phase formulation are much fewer than those of the one phase model.

The second problem considered is a time dependent variant of a similar problem to the first, with constant air pressure boundary conditions for the two phase formulation on all but one side with water, and for the one phase approximation, all sides exposed to air as zero flow boundaries. In this case, the metric used to determine the accuracy of of either phase model was the fraction of the available pore volume of the porous medium that is filled with water after one day passes in the simulation, comparing with an exact analytic solution.

See [8], figures 2 and 3 for pictures showing the problem set up for both of these examples.

Both simulations were run for one phase and two phase flow, using the relations $k_{rl} = \theta_l^n$, with $n = 1, 2, 4$. The simulations indicate that the fraction of water occupied pore volumes are roughly the same for both for $n = 1$ and 2, (which match the condition of admissable solutions as mentioned in the previous section), but for $n = 4$, the two vary significantly, as was predicted in the heuristic analysis of the previous section, and both formulations are equally difficult to solve.

What these examples show is that, in many practical situations, the assumption of a passive air phase will give results very similar to a full two-phase formulation with far fewer unknowns to solve for, at the cost of generating harder linear problems to solve. However, if the air relative permeability is very flat near $\theta_a = 0$, then the one-phase formulation cannot accurately represent the full dynamics of the system, and the two-phase formulation is recommended.

One potential difficulty in using the two phase model is that numerical schemes can have problems with convergence to physically admissable solutions, as the basic analysis above indicates. In particular, [25] and [26] show that upwind schemes must be aligned according to saturation fronts, with proper time and space stepping procedures done to guarantee convergence to the physically correct solution for the two-phase model. On the other hand, as the one phase system is a parabolic approximation, numerical convergence to the wrong physical solution has never been observed in practice; what one sees using central weighting as opposed to upwind weighting is the appearance of non-physical oscillations instead, that can be eliminated using smaller discretization in space, or using special upwinding schemes proposed to develop monotone discretizations ([27], [24], [25], [12]); this point will be expanded upon in chapter 3.

1.2 Various forms of Richards' equation

As shown above, Richards' equation can be interpreted as a simplification of the standard two-phase flow formulation for a gas and water phase in a porous medium where the pressure gradient required to drive flow of the gas phase is ignored due to the large mobility contrast between the water and gas phases.

Analytic closed form solutions for the one-dimensional Richards' equation have been derived for a few specialized forms of the constitutive relations describing the soil-water retention and the unsaturated hydraulic conductivity functions ([28], [29], [30]). However,

these solutions are not generally applicable because either the functional forms are dissimilar from widely used constitutive relations that represent real soils, and/or the solutions impose strict requirements on the initial and boundary conditions, so most practical situations require a numerical solution in two or three dimensions, depending on the problem and complexity of the flow situation. Despite the fact that the first reasonably complete conservative numerical solution method was published in the early 1990s [12], the numerical solution of Richards' equation remains computationally expensive, and in certain circumstances, unreliable. A universally robust and accurate solution methodology has not yet been identified that is applicable across the full range of soils, initial and boundary conditions found in practice. Existing solution codes have been modified over the years in an attempt to increase their robustness; however, as the theoretical analysis regarding the existence of solutions given sufficiently regular Cauchy data can only guarantee relatively low regularity of solutions unless strong assumptions on θ and K are made, numerical methods can fail to demonstrate reliable convergence behavior in practice, especially for higher order linearizations.

There are three standard forms of the equation that are considered, in which the primary variable changes depending on the case being considered:

There is the purely pressure head-based form,

$$\begin{cases} C(\Psi)\partial_t \Psi - \operatorname{div}(K(\Psi)\nabla(\Psi + z)) = f(\Psi), & (x,t) \in \Omega \times (0,T], \\ \Psi(x,0) = \Psi_0(x), & (x,t) \in \Omega \times \{0\}, \\ \Psi(x,t) = \Psi_D(x,t), & (x,t) \in \Gamma_D \times (0,T], \\ K(\Psi)\nabla(\Psi + z) \cdot \nu = 0, & (x,t) \in \Gamma_N \times (0,T]. \end{cases} \tag{1.2}$$

The pure saturation-based form,

$$\begin{cases} \partial_t \theta - \operatorname{div}(D(\theta)\nabla\theta) = f(\theta), & (x,t) \in \Omega \times (0,T], \\ \theta(x,0) = \theta_0(x), & (x,t) \in \Omega \times \{0\}, \\ \theta(x,t) = \theta_D(x,t), & (x,t) \in \Gamma_D \times (0,T], \\ D(\theta)\nabla(\theta) \cdot \nu = 0, & (x,t) \in \Gamma_N \times (0,T]. \end{cases} \tag{1.3}$$

Finally, there is the mixed saturation-head form

$$\begin{cases} \partial_t \theta(\Psi) - \operatorname{div}(K(\theta(\Psi))\nabla(\Psi + z)) = f(\theta(\Psi)), & (x,t) \in \Omega \times (0,T], \\ \theta(\Psi(x,0)) = \theta(\Psi_0(x)), & (x,t) \in \Omega \times \{0\}, \\ \Psi(x,t) = \Psi_D(x,t), & (x,t) \in \Gamma_D \times (0,T], \\ K(\theta(\Psi))\nabla(\Psi + z) \cdot \nu = 0, & (x,t) \in \Gamma_N \times (0,T]. \end{cases} \tag{1.4}$$

Both forms (1.2) and (1.4) are valid for the entire regime of possible saturation cases, while the pure saturation-based form is only relevant for purely unsaturated problems. We note here that these may be equivalent for classical solutions where appropriate use of the chain rule and other simplifications can bring you from one formulation to another, but in the case of weak solutions where those derivatives may not be defined, these formulations are not equivalent. That being said, the head-based form, being an explicit chain rule calculation, allows for simple expansion of the storage coefficient $C(\Psi)$,

$$C(\Psi) = S_s \theta_a(\Psi) + c(\Psi),$$

where S_s is the specific storage and θ_a is the saturation of the aqueous phase; this facilitates models of contaminant transport or non-aqueous phase materials ([4], [31]).

However, this extendability comes at a price: namely, the scheme is not mass conservative. Celia et al. [12] show numerically that, due to the equation not being in conservation form, discretizations of (1.2) with low order (Backward Euler) time discretizations suffer mass discrepancies, as the strong nonlinearities of θ imply that low order discretizations in time of the chain rule do not approximate the time derivative well. To combat this, higher order time integration can be used. Tocci et al. [32] use a method of lines approach, in which the problem is discretized in space using finite differences, and integrated in time using an adaptive multistep method in time to deliver high fidelity-in-time solutions of (1.2).

The pure saturation form enjoys perfect mass conservation on the discrete level, and the time derivative being linear significantly simplifies analysis. However, the diffusion coefficient $D(\theta) \sim K(\theta)/C(\theta) \to \infty$ as $\theta \to 1$, since $C(\theta) = \frac{\partial \theta}{\partial \Psi} \to 0$. As such, the problem fully degenerates in an asymptotically divergent manner near full saturation. Further, for physically realistic formulations of K and θ, $D(\theta) \to 0$ near fully dry media, thus degenerating the problem for $\theta = 0$ as well. There have been some methods proposed to deal with these degeneracies. In [33], Pop derives error estimates for a time discretization method of the

saturation form, which he allows to fully degenerate, by applying the Kirchoff transformation $\theta(\Psi) \to b(u)$ (described in section 2.1.2 of this chapter), and considering a perturbed form of D as follows:

$$D_\epsilon(\theta) = \begin{cases} \epsilon, & 0 \leq \theta \leq C\epsilon^{1/(\alpha-1)}, \\ \frac{1}{\epsilon^r}, & 1 - C\epsilon^{r/\beta} \leq \theta \leq 1, \\ D'(\theta), & otherwise. \end{cases}$$

He is then able to prove stability of a Backward Euler discretization, and show numerical convergence using the standard VGM model, with $n \geq 3$ (described below). Another difficulty with the primary variable being the water content is that the water content is a continuous variable only in homogeneous soils, and soils are seldom homogeneous over significant length scales in most practical applications. In the case of layered soils, the water content is discontinuous across layer interfaces because of unique unsaturated capillary head relations in the different soil layers [34]. Rather, the pressure head Ψ is continuous, and it is better to write the Richards equation with pressure head as the dependent variable and evaluate the moisture content in terms of Ψ.

The most popular of these three forms (and the one we focus on for this work) is the mixed saturation-head form (1.4). This form enjoys perfect mass conservation on the discrete level, without the extreme degeneracy of the diffusion coefficient near full saturation. The problem is still degenerate, however, and switches type from parabolic to elliptic in the fully saturated regime.

From the analytic point of view, a significant contributor to the difficulty of establishing well-posedness of the problem is the presence of the nonlinear water content $\theta(\Psi)$ in the parabolic term. This water content is in general a strongly nonlinear function of the pressure head, and this nonlinearity makes proving existence and uniqueness of solutions to the problem a considerable challenge [9], [10]. Furthermore, for certain physically realistic parametrizations, the hydraulic conductivity $K(\theta)$ can develop steep spatial gradients near saturation fronts, which can significantly impact convergence of nonlinear iterations near full saturation; this point will be expanded upon in chapter 3. As such, a significant portion of the literature on numerical methods is focused on finding linearizations for Richards' equation that are robust with respect to these poorly behaved parameters.

1.3 Parameter models

Accurately modeling the change in hydraulic conductivity K as a function of pressure head for unsaturated media has proven a difficult challenge for hydrologists to tackle. Some of the many reasons for this are the lack of homogeneity in micropore structure, and that capillary action effectively changes the macropore structure by obstructing pore networks and forming air pockets that are hard for infiltrating fluids to penetrate [35], [36]. This is reflected by the practical difficulty in measuring water retention as a function of pressure head, or the even greater difficulty in measuring the effect of a change in pressure head on conductivity for test media [37], [2], [38]. To tackle this issue, there have been many proposed models to represent K as a function of Ψ, and the easier to measure relation of K as a function of water content θ, with the water content dependent on Ψ.

In what follows, we discuss various K and θ models used in the literature, with a focus on the Van Genuchten-Mualem model that is the most widely used.

1.3.1 K models

One popular group of parameter relations have K depending on Ψ or θ as some sort of power law dependence, such as the models first proposed by Kozeny [39]:

$$K_{rel} = K/K_{sat} = \theta_N^{\alpha}.$$

Here θ_N is the normalized saturation level of a medium, whose water content varies from a residual water content θ_R to a maximum water content θ_S,

$$\theta_N(\Psi) = \frac{\theta(\Psi) - \theta_R}{\theta_S - \theta_R}.$$

K_{sat} is the maximal conductivity at full saturation, and α is a parameter that can be tuned to fit observed saturation curves of a given porous medium.

Averjanov [35], who derived a simplified model in which a pore could be modelled as a capillary tube in which there is a concentric air tube restricting the flow of water to the outer annulus of the tube, proposed $\alpha = 3.5$ for most media. This particular formulation was widely used in the past due to its ease of implementation.

The other popular group of relations couple K and θ using integral dependence. In the

context of petroleum engineering, the Burdine equation [40] is widely used,

$$K_{rel}(\theta - \theta_R) = \theta_N^2 \frac{\left[\int_{\theta=\theta_R}^{\theta-\theta_R} \mathrm{d}\theta/\Psi^2\right]}{\left[\int_{\theta=\theta_R}^{\theta_S} \mathrm{d}\theta/\Psi^2\right]},$$

while for soil scientists, the initial choice was a simple quadrature of a more general form, which is a variant of the model introduced by Childs and Collis-George [37]:

$$K_{rel}(\theta - \theta_R) = \theta_N^\beta \frac{\left[\int_{\theta=\theta_R}^{\theta-\theta_R} \mathrm{d}\theta/\Psi^2\right]}{\left[\int_{\theta=\theta_R}^{\theta_S} \mathrm{d}\theta/\Psi^2\right]}. \tag{1.5}$$

The coefficient β was suggested to be 0, 4/3, and 1 by various authors, but Mualem [41] was one of the first to numerically verify using experimental data from 45 different soil samples that a choice of $\beta = 1/2$ minimized the error between the model K and measured K, which he verified using a deviation measure

$$D = \left[(1 - \theta_{N\,\min})^{-1} \int_{\theta_{N\,\min}}^{1} \left[\ln(K_{meas}(\theta_N)) - \ln(K_{model}(\theta_N))\right]^2 \mathrm{d}\theta_N\right]^{1/2},$$

that can be interpreted as an average of the orders of magnitude difference of K_{meas} and K_{model} over the range of $\theta_{N\,\min}$ values.

With this measure of error, Mualem found that the average D value across all 45 soils with $\beta = 1/2$ was 0.97, with standard deviation 0.82, as opposed to 1.49 (std dev 1.64) for Averjanov's model, and 1.17 (std dev 0.88) for the CCG model with $\beta = 4/3$.

1.3.2 θ models

Models of the water content with respect to changes in pressure head proved to be a much more tractable exercise, as measuring this experimentally was more direct. Among the different models suggested, one of the most popular considered is the Brooks-Corey model [42]:

$$\theta(\Psi) = \left[\frac{\Psi_b}{\Psi}\right]^\lambda,$$

with Ψ_b being a threshold negative "bubblng pressure", at which point there is a discontinuity in the derivative of the water content- pressure head model. Brooks and Corey were able to show using data from over 40 different materials that their model could predict the

measured soil-water retention curve to high accuracy, though near full saturation and near this experimentally discovered minimal bubbling pressure, the predictions lose fidelity.

Another popular model is the Van Genuchten model [21],

$$\theta_N = \frac{1}{(1 + (-\alpha\Psi)^n)^{1-1/n}},$$

where $\alpha > 0$ and $n > 1$ are physically determined parameters that can be determined using a log-slope technique from the graph of the soil-water retention curves. The benefit of this model is the smoothness of the functions involved, particularly at the initially dry ($\theta_N \to 0$) and nearly saturated ($\theta_N \to 1$) regimes; this is expanded upon in the next section.

In [21], Van Genuchten verified his formulation of water content with experimental data for many materials, with a focus on Hygiene sandstone, Touchet silt loam, silt loam, Beit Netofa clay, and Guelph loam (wetting and drying phases). For all but the Beit Netofa clay, the model proposed and the experimental data match closely; the discrepancy in the former soil was explained as being due to the poor estimate on the residual water content for that particular material.

1.3.3 VGM model

Van Genuchten incorporated his model for the water content with the Mualem model for conductivity as function of water content, deriving the standard Van Genuchten-Mualem (VGM) model for K and θ:

$$\theta(\Psi) = \begin{cases} \theta_R + (\theta_S - \theta_R)\left[1 + (-\alpha\Psi)^n\right]^{\frac{1}{n}-1}, & \Psi < 0, \\ \theta_S, & \Psi \geq 0, \end{cases} \tag{1.6}$$

$$K(\theta) = \begin{cases} K_S\theta_N(\Psi)^{\frac{1}{2}}\left[1 - \left(1 - \theta_N(\Psi)^{\frac{n}{n-1}}\right)^{\frac{n-1}{n}}\right]^2, & \Psi < 0, \\ K_S, & \Psi \geq 0. \end{cases} \tag{1.7}$$

θ_R is the residual water content, θ_S and K_S are the maximal water content and hydraulic conductivites, resp. at full saturation, and the parameters n and α are parameters that are gleaned via a log-slope technique used on the saturation-head retention curves. n increases as pore size distribution becomes more uniform, and α gives some measure of the slope of θ at an inflection point in the S-curve that is generated by the model [21]. The function θ_N is

the normalized water content,

$$\theta_N(\Psi) = \frac{\theta(\Psi) - \theta_R}{\theta_S - \theta_R}.$$

As the smoothness of both functions depends on n, we wanted to investigate the smoothness of these functions with respect to this parameter.

Below are the first two derivatives of $\theta_N(\Psi)$:

$$\theta'_N(\Psi) = \alpha(n-1)\frac{(-\alpha\Psi)^{n-1}}{\left[1+(-\alpha\Psi)^n\right]^{2-\frac{1}{n}}},$$
$$\theta''_N(\Psi) = \alpha^2(n-1)\frac{(-\alpha\Psi)^{n-2}}{\left[1+(-\alpha\Psi)^n\right]^{2-\frac{1}{n}}}\left[\frac{(-\alpha\Psi)^n}{1+(-\alpha\Psi)^n}(2n-1)-(n-1)\right].$$

Thus, the function is strictly monotone increasing for $\Psi < 0$, and continuously differentiable for all admissable n, with derivative asymptotics given by the following,

$$\theta'_N(\epsilon) = \mathcal{O}(\epsilon^{n-1}), \quad \epsilon \to 0, \tag{1.8}$$
$$\theta'_N(\epsilon) = \mathcal{O}(\epsilon^{-n}), \quad \epsilon \to \infty. \tag{1.9}$$

The asymptotics for θ''_N show that θ'_N is continuous at $\Psi = 0$ for $n > 2$:

$$\theta''_N(\epsilon) = \mathcal{O}(\epsilon^{n-2}), \quad \epsilon \to 0, \tag{1.10}$$
$$\theta''_N(\epsilon) = \mathcal{O}(\epsilon^{-(n+1)}), \quad \epsilon \to \infty. \tag{1.11}$$

The bracketed term in θ''_N implies that there is an inflection point in the graph of the function for some $\Psi < 0$, at which point θ'_N attains a maximum. This inflection point is used by Van Genuchten in his log-slope technique to determine n. Hence, θ_N is Lipschitz continuous, with Lipschitz constant being the value of θ'_N at that inflection point, implying that Newton schemes should perform well for any admissable n and α, though for $1 < n < 2$, the unboundedness of the second derivative as $\Psi \to 0$ can foreshadow potential problems with a Newton linearization.

Elementary calculations show that the maximum value of θ'_N (i.e, the Lipschitz constant L_θ) happens when $(-\alpha\Psi)^n = \frac{n-1}{n}$, with value

$$L_\theta = \max_{\Psi} \theta'_N(\Psi) = \alpha\frac{n(n-1)}{2n-1}\left[\frac{n-1}{2n-1}\right]^{1-\frac{1}{n}}. \tag{1.12}$$

Writing $n = 1 + \epsilon$, its a straightforward matter to check that

$$L_\theta < \frac{\alpha\epsilon}{2},$$

which implies that as $\epsilon \to 0$, $L_\theta \to 0$. As $n \to \infty$, clearly $L_\theta \to \frac{\alpha n}{4}$. In this sense, θ as function of Ψ varies less rapidly as $n \to 1$, which is important when considering the properties of K as a function of Ψ.

Given the more complex expression involving K, we only analyze the first derivative of K with respect to θ_N:

$$K'(\theta_N) = \xi(\theta_N)\left(\frac{\theta_N^{-1/2}}{2}\xi(\theta_N) + 2\theta_N^{\frac{1}{2}}\xi'(\theta_N)\right),$$

Where $\xi(\theta_N) = \left[1 - (1 - \theta_N^{\frac{n}{n-1}})^{\frac{n-1}{n}}\right]$, and $\xi'(\theta_N) = \theta_N^{\frac{1}{n-1}}(1 - \theta_N^{\frac{n}{n-1}})^{-1/n}$. A chain of nested inequalities can be used to show that, because $0 \leq \theta_N \leq 1$ and $n > 1$, $\xi(\theta_N) \leq \theta_N^{\frac{n}{n-1}}$, and

$$\xi'(\theta_N) \geq \frac{\theta_N^{\frac{1}{n-1}}}{1 - \theta_N^{\frac{1}{n-1}}}.$$

Thus, we get the following bounding asymptotics for K' approaching relatively dry and nearly saturated conditions:

$$K'(\epsilon) \leq \mathcal{O}(\epsilon^{-\frac{1}{2}+\frac{2n}{n-1}}), \quad \epsilon \to 0 \tag{1.13}$$

$$K'(\epsilon) \geq \mathcal{O}\left(\frac{\epsilon^{\min(1,\frac{1}{n-1})}}{1 - \epsilon^{\min(1,\frac{1}{n-1})}}\right), \quad \epsilon \to 1. \tag{1.14}$$

Thus, the derivative K approaches the dry case smoothly, but approaches the fully saturated case sharply, as $\xi'(1) \to \infty$, most notably for media with n approaching 1. However, if one considers the full derivative,

$$\frac{\mathrm{d}K}{\mathrm{d}\Psi} = K'(\theta_N(\Psi))\theta_N'(\Psi),$$

and the following property of θ_N,

$$\theta_N^{\frac{n}{n-1}} = \frac{1}{1 + (-(\alpha\Psi)^n}, \tag{1.15}$$

one can show the following:

Lemma 1.3.1 (Analytic properties of $K(\Psi)$)**.** *For the VGM model, if $n > 2$, then $K(\Psi)$ is Lipschitz throughout the entire domain of definition. If $n = 2$, the function is Lipschitz, but has discontinuous derivative at $\Psi = 0$. For $1 < n < 2$, the function has unbounded derivative as $\Psi \to 0$ from the left.*

Proof. Expanding the derivative of K as a function of Ψ and using the chain rule and (1.15) to simplify, one can deduce

$$\frac{\partial K}{\partial \Psi} \leq \alpha(n-1)\left[\frac{1}{2}(-\alpha\Psi)^{n-1} + 2(-\alpha\Psi)^{n-2}\left(1+(-\alpha\Psi)^n\right)^{\frac{1}{2}-\frac{1}{n}}\right],$$

which clearly shows how n determines the analytic properties of $K(\Psi)$. □

This corroborates with the many observations in the literature on the numerical challenge of modeling variably saturated-unsaturated media, as steep gradients in the conductivity near saturation fronts with certain discretizations have been shown to lead to false convergence when modeling both the water and air phase (i.e two phase flow) [4], [27] and the appearance of non-physical oscillations (i.e, a lack of monotonicity) near saturation fronts for Richards' equation [24], [27], [43], [12].

1.4 Plots of K_r and θ for different soils

The primary challenge for numerical considerations is that the conductivity of unsaturated soils tends to vary by several orders of magnitude from being dry to being fully saturated; particularly for initially dry soils, the conductivity approaches 0 for most media. In this sense, the nonlinear effect of capillary action drives the behavior of numerical approximations. This change from minimal to maximal relative conductivity is not only large, but occurs very rapidly near full saturation. For certain soil compositions that we considered in our numerical tests, the standard VGM models described below predicted that the relative conductivity changes by as much as 10 orders of magnitude from fully saturated ($\Psi = 0$cm) to $\Psi = -100$cm, which is a small change in pressure head compared to standard simulations that feature the pressure head changing by 100s of cm in dry media [44, 45], or even on the order of 1000s of cm in the testing done by Mualem and Van Genuchten to validate their models.

To illustrate this, several graphs of VGM K_r and θ_N are presented. Note that in all but the Beit Netofa clay example, $K_r(\Psi)$ falls by at least four orders of magnitude from $\Psi = 0$ to $\Psi = -10$, and fall from 7 to 10 orders of magnitude by $\Psi = -100$. Such high contrast K tend to make simulations very computationally expensive.

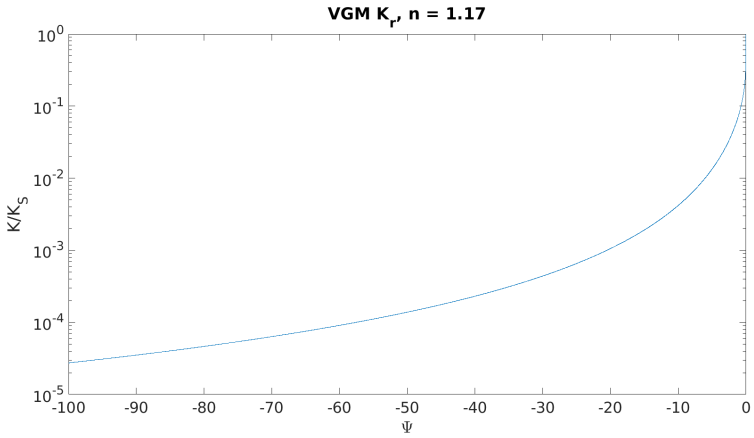

Figure 1.1. VGM $K_r(\Psi)$, Beit Netofa clay ($\alpha = 0.152$, $n = 1.17$, $K_S = 8.2 \times 10^{-4}$).

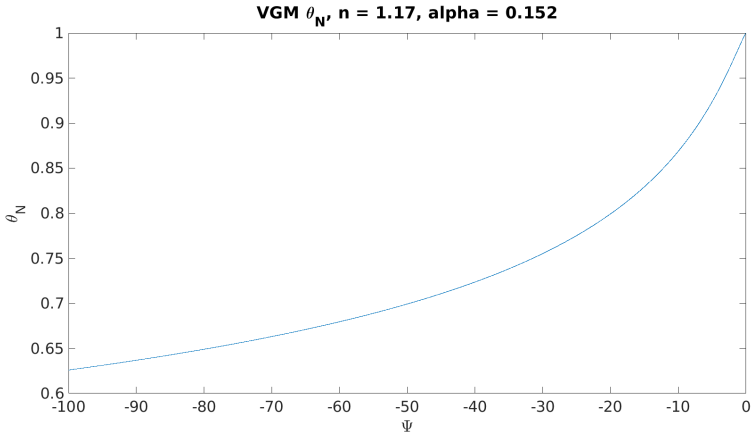

Figure 1.2. VGM $\theta_N(\Psi)$, Beit Netofa clay ($\alpha = 0.152$, $n = 1.17$, $K_S = 8.2 \times 10^{-4}$).

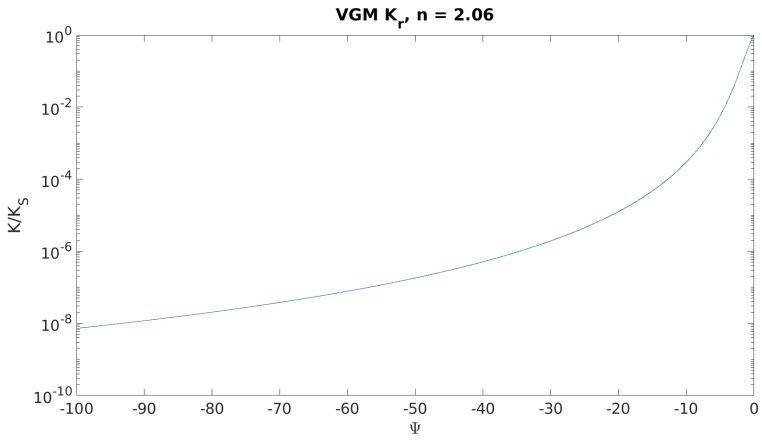

Figure 1.3. VGM $K_r(\Psi)$, silt loam ($\alpha = 0.423$, $n = 2.06$, $K_S = 5 \times 10^{-2}$).

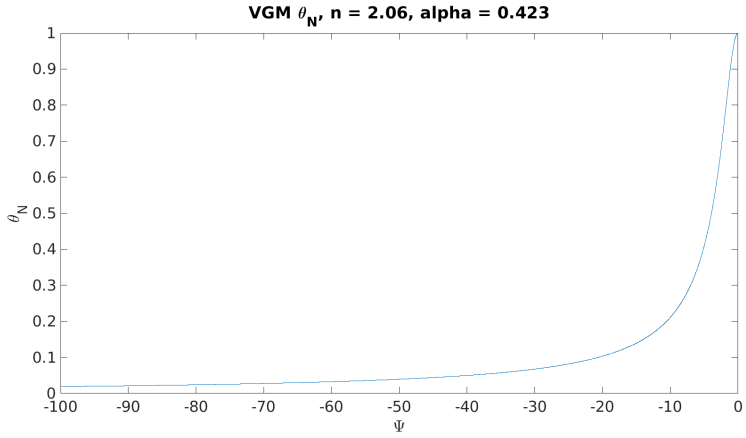

Figure 1.4. VGM $\theta_N(\Psi)$, silt loam ($\alpha = 0.423$, $n = 2.06$, $K_S = 5 \times 10^{-2}$).

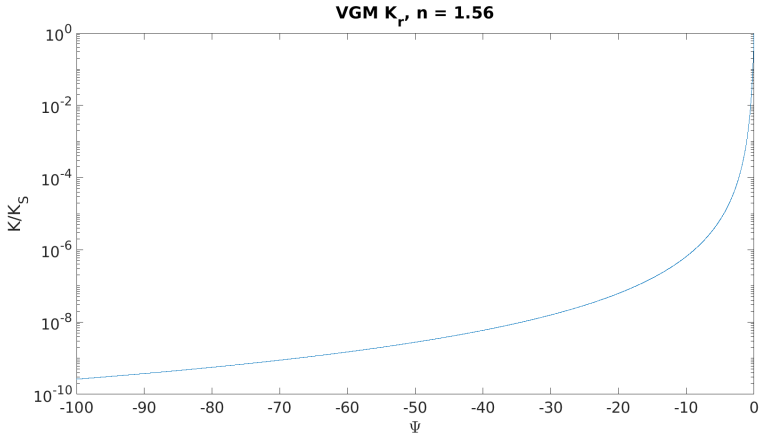

Figure 1.5. VGM $K_r(\Psi)$, loam soil ($\alpha = 3.6$, $n = 1.56$, $K_S = 2.5 \times 10^{-1}$).

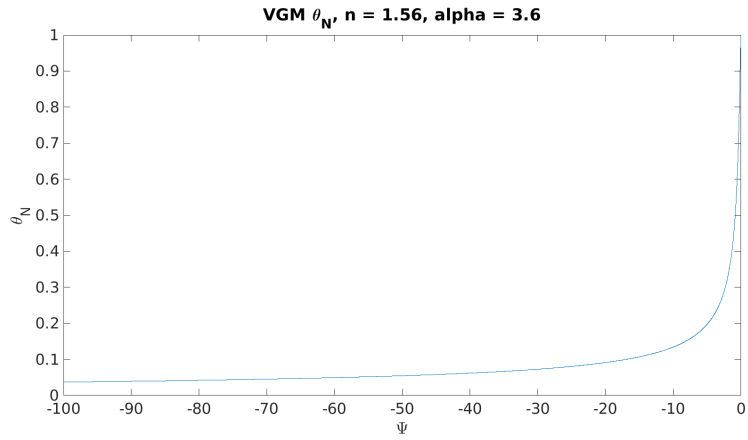

Figure 1.6. VGM $\theta_N(\Psi)$, loam soil ($\alpha = 3.6$, $n = 1.56$, $K_S = 2.5 \times 10^{-1}$).

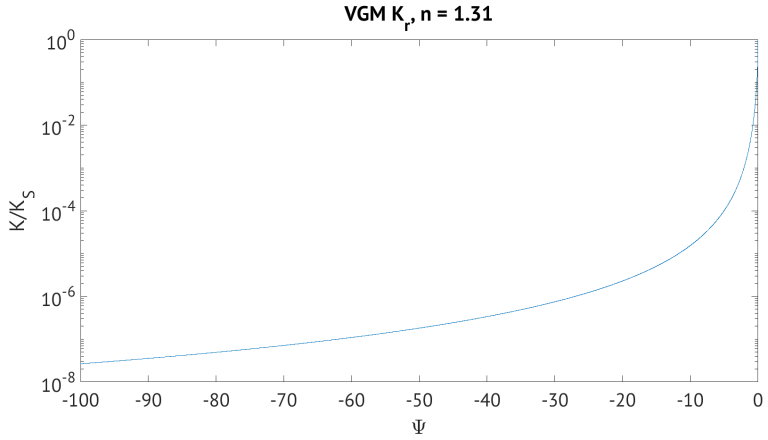

Figure 1.7. VGM $K_r(\Psi)$, clay loam ($\alpha = 1.9$, $n = 1.31$, $K_S = 6.2 \times 10^{-2}$).

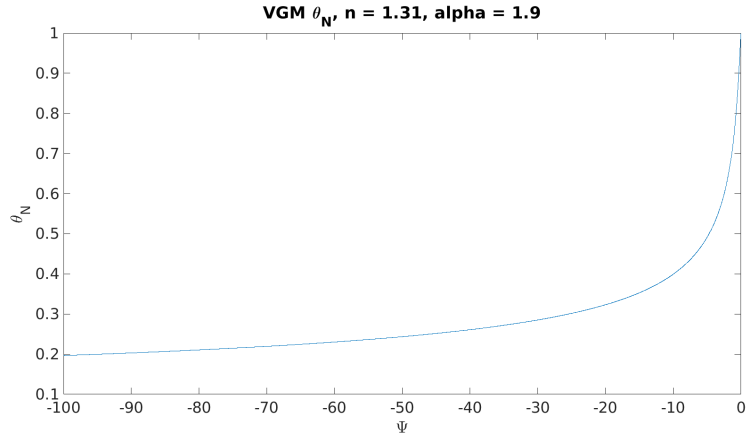

Figure 1.8. VGM $\theta_N(\Psi)$, clay loam ($\alpha = 1.9$, $n = 1.31$, $K_S = 6.2 \times 10^{-2}$).

Chapter 2
Well-posedness of Richards' equation

Here we survey the results on existence and uniqueness as described in the works [9] and [10]. Given the quasilinearity and degenerate nature of the mixed pressure head-water content form of Richards equation, proving existence and uniqueness of weak solutions, even in the case of a Lipschitz boundary and smooth initial data, is challenging. The main source of difficulty lies in writing estimates for the time derivative term: θ, while weakly monotone increasing, is also nonlinear, so because it is in the time derivative, the standard Gronwall approach for existence of weak solutions must be adapted to accomodate this nonlinearity, and uniqueness must be derived by other means. Due to this nonlinearity, stability in the sense of Hadamard is also difficult to prove; as such, we rely on results regarding the stability of the finite element approximations we consider as proved by Radu et al. [20] to imply stability of solutions of the PDE to small changes of the initial and boundary data; see chapter 4 for a brief overview of such results.

Another complication is that the parabolic term is also degenerate, as the water content term becomes constant at full saturation. This happens for mixed unsaturated-saturated scenarios, such as the infiltration of a saturation front into unsaturated media. The results discussed here regarding existence and uniqueness of a weak solution to (1.4) described in this chapter accomodate these scenarios.

The outline of this chapter is as follows: in section 2.1, we introduce the standard notation that are used for energy estimates, and also define the Kirchoff transformation that is applied to (1.4). Section 2.2 discusses the assumptions on the Cauchy data, and on the nonlinear terms $b(u)$ and $K(b(u))$; here we also define the Legendre transformation of $b(u)$, $B(u)$, and motivate its role in the proofs of Alt and Luckhaus [9] and Otto [10]. Section 2.3 discusses

the strategies employed by Alt and Luckhaus [9] used to prove the existence of weak solutions to the PDE, and uniqueness of the weak solution as proved by Alt and Luckhaus for certain classes of the nonlinearities K and θ, and in general by Otto [10] to the problem.

2.1 Preliminaries and definitions

To prove existence and uniqueness of solutions to (1.4), we first need to define the weak formulation of the Cauchy problem.

2.1.1 Sobolev spaces and weak derivatives

Before establishing a priori estimates, we first define the spaces of functions that have finite energy, in the Lebesgue sense. Given a domain $\Omega \subset \mathbb{R}^d$ and a finite time interval $[0, T]$, we define the reflexive Hilbert space $L^2(\Omega)$ as the space of functions $u(x) : \Omega \to \mathbb{R}^{d_2}$ ($d_2 = 1, 2, 3$) endowed with inner product

$$(u, v)_{L^2(\Omega)} := \int_\Omega u \cdot v \, \mathrm{d}x.$$

This inner product induces the following norm and categorization of elements in the space:

$$L^2(\Omega) = \{u(x) : \Omega \to \mathbb{R}^{d_2} : \quad ||u||_{L^2(\Omega)} := \left(\int_\Omega |u|^2 \, \mathrm{d}x \right)^{1/2} < \infty\}.$$

Sobolev spaces are a natural generalization of L^2 for functions with weak (distributional) partial derivatives; as such, solutions of PDE fall quite naturally into these spaces of functions defined almost everywhere, hence standard tools of functional analysis apply. For our purposes, we define the particular self-adjoint Sobolev spaces

$$W^{k,2}(\Omega) = H^k(\Omega) := \{u \in L^2(\Omega) : \left(\sum_{|\alpha| \leq k} ||D^\alpha u||^2 \right)^{1/2} < \infty\},$$

where the differential operator D^α is written in the standard multi-index notation, and these derivatives are to be interpreted in the sense of distributions. We refer to [46] for an in-depth discussion of weak derivatives and Sobolev space theory.

Finally, to incorporate time dependence for parabolic problems, we consider the spaces of functions in time taking values in a Banach space X, $L^p(0, T; X)$ as the the set of functions

$u : [0, T] \to X$ with

$$||u||_{L^p(0,T;X)} := \begin{cases} \left(\int_0^T ||u(t)||_X^p \right)^{1/p} \, \mathrm{d}t < \infty, & 1 \leq p < \infty \\ \operatorname{ess\,sup}_{0 \leq t \leq T} ||u(t)||_X < \infty, & p = \infty. \end{cases}$$

In this context, weak derivatives, and therefore Sobolev spaces, can be defined. We define the Sobolev space $H^1(0, T; X)$ to be the space of functions $u \in L^2(0, T; X)$ such that u' exists in the weak sense and belongs to $L^2(0, T; X)$. The norm of this space is given by

$$||u||_{H^1(0,T;X)} := \begin{cases} \left(\int_0^T ||u(t)||_X^p + ||u'(t)||_X^p \right)^{1/p} \, \mathrm{d}t, & 1 \leq p < \infty \\ \operatorname{ess\,sup}_{0 \leq t \leq T} \left(||u(t)||_X + ||u'(t)||_X \right), & p = \infty. \end{cases}$$

Typically, the spaces X that we concern ourselves with are themselves Sobolev spaces, of functions that vary in the spatial variable. This perspective facilitates tracking the energy of weak solutions in PDE that vary in space and time, and in particular, parabolic problems such as Richards' equation.

Some useful properties of spaces that vary in time are recorded below; their proof can be found in [46].

Theorem 1 ([46] Ch. 5.9, Thm 2)**.** *Let $u \in H^1(0, T; V)$ for some $1 \leq n < \infty$. Then*

1. *$u \in C([0, T]; V)$, the space of continuous functions $u : [0, T] \to V$ with*

$$\max_{0 \leq t \leq T} ||u(t)||_V < \infty$$

   (after possibly being redefined on a set of measure zero), and

2. *$u(t) = u(r) + \int_r^t u'(s)\, ds$ for all $0 \leq r \leq t \leq T$.*

3. *Furthermore, the estimate*

$$\max_{0 \leq t \leq T} ||u(t)||_V \leq C||u||_{H^1(0,T;V)}$$

   holds, with the constant C depending only on T.

Theorem 2 ([46] Ch. 5.9, Thm 3)**.** *Suppose $u \in L^2(0, T; H_0^1(\Omega))$, with $u' \in L^2(0, T; H^{-1}(\Omega))$. Then*

1. *The norm $||u||_{L^2(\Omega)}$ is continuous in time, i.e, $u \in C([0,T]; L^2(\Omega))$, (after possibly being redefined on a set of measure zero).*

2. *The mapping*

$$t \to ||u(t)||^2_{L^2(\Omega)}$$

is absolutely continuous, with

$$\frac{d}{dt}||u(t)||^2_{L^2(\Omega)} = 2(u(t), u'(t))$$

for a.e. $0 \leq t \leq T$.

3. *Furthermore, the estimate*

$$\max_{0\leq t\leq T} ||u(t)||_{L^2(\Omega)} \leq C(||u||_{L^2(0,T;H^1_0(\Omega))} + ||u'||_{L^2(0,T;H^{-1}(\Omega))})$$

holds, with the constant C only depending on T.

This second result is especially useful when deriving energy estimates for parabolic PDE with a standard parabolic term, as the second item allows us to "pull out" the time derivative from the integrand, which allows us to treat the parabolic part as a time derivative of the energy of the solution. In our case however, the presence of the nonlinearity $b(u)$ will not allow us to do this so easily, hence the need to bound the energy in time via non-standard methods. This also complicates the standard Gronwall inequality estimates used to prove uniqueness of these weak solutions to the problem.

To properly describe the well-posedness of Richards' equation, we must first describe the notion of weak solutions to a boundary/initial value problem. To illustrate this, let us define the weak formulation for the canonical example for parabolic PDEs, the heat equation. We define some bounded domain $\Omega \subset \mathbb{R}^d$, for a fixed time interval $[0,T]$, with homogeneous Dirichlet and Neumann data defined on the boundary of the domain $\partial\Omega = \Gamma_D \cup \Gamma_N$, and an initial condition for $t = 0$. The heat equation with these conditions has solutions u that satisfy the following:

$$\begin{cases} \partial_t u - \Delta u = f, & (x,t) \in \Omega \times (0,T], \\ u(x,0) = u_0(x), & x \in \Omega, \\ u(x,t)|_{\Gamma_D} = 0, & (x,t) \in \Gamma_D \times (0,T], \\ \frac{\partial u}{\partial \nu} = 0, & (x,t) \in \Gamma_N \times (0,T]. \end{cases} \tag{2.1}$$

In the particular case when $\Gamma_D = \partial\Omega$, one can define the space of solutions as

$$V = L^2(0, T; H_0^1(\Omega)),$$

with

$$H_0^1(\Omega) = \{u \in H^1(\Omega) : u|_{\partial\Omega} = 0\}.$$

The weak formulation of this problem is derived by multiplying the PDE with a suitable test function $v \in V$, and integrating in space and time to get the weak formulation, namely $u \in L^2(0, T; H_0^1(\Omega))$ with $u' \in L^2(0, T; H^{-1}(\Omega))$ is a weak solution of the initial boundary value problem (2.1) if the following variational equality is satisfied for any $v \in V$ with $v(T) = 0$:

1. The following holds for all $v \in L^2(0, T; H_0^1(\Omega))$ with $v(T) = 0$:

$$\int_0^T \langle \partial_t u, v \rangle \, \mathrm{d}t + \int_0^T \int_\Omega (u - u_0) \partial_t v \, \mathrm{d}x \, \mathrm{d}t = 0,$$

   with $\langle \cdot, \cdot \rangle$ being the duality pairing between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$.

2. u satisfies

$$\int_0^T \langle \partial_t u, v \rangle \, \mathrm{d}t + \int_0^T \int_\Omega \nabla u \cdot \nabla v \, \mathrm{d}x \, \mathrm{d}t = \int_0^T \int_\Omega f v \, \mathrm{d}x \, \mathrm{d}t$$

   for each $v \in L^2(0, T; H_0^1(\Omega))$.

In this sense, the weak formulation of a problem allows for a larger class of solutions than exists in the classical sense, as the derivatives of admissable functions can be defined almost everywhere in space, and in the sense of distributions in time, increasing the classes of admissable solutions to the PDE.

2.1.2 Kirchoff transformation of Richards' equation

To work with the nonlinearities present in (1.4), one common technique is to use the Kirchoff integral transformation technique [47] applied to Richards' equation, by defining the transformed unknown,

$$u(\Psi) = \int_0^\Psi K(\theta(s)) \, \mathrm{d}s. \tag{2.2}$$

This function has the following properties:

1. u preserves the sign of Ψ, and is 0 when Ψ is 0.

2. u is strictly increasing in Ψ, hence it is invertible.

3. The Kirchoff transformation of $\theta(\Psi)$, which we denote as

$$b(u(\Psi)) = \theta(\Psi(u)) = K^{-1}\left(\frac{\partial u}{\partial \Psi}\right),$$

   is monotone and continuous, hence there exists some C^1 convex function Φ with $\nabla\Phi = b$. Note that we can assume $b(0) = 0$, by considering $b(u) = b(u(\Psi)) - 1$, so that

$$ub(u) \geq 0, \quad u \in \mathbb{R}.$$

4. The PDE after applying the Kirchoff transformation becomes

$$\partial_t b(u) - \operatorname{div}(\nabla u + K(b(u))e_z) = f(b(u)),$$

   hence the ellipcit part becomes semi-linear, simplifying the analysis.

The full Kirchoff transformed problem becomes

$$\begin{cases} \partial_t b(u) - \operatorname{div}(\nabla u + K(b(u))e_z) = f(b(u)), & (x,t) \in \Omega \times (0,T], \\ b(u(x,0)) = b(u_0(x)), & t = 0, \\ u(x,t) = u_D(x,t), & (x,t) \in \Gamma_D \times (0,T], \\ (\nabla u + K(b(u))e_z) \cdot \nu = 0, & (x,t) \in \Gamma_N \times (0,T]. \end{cases} \tag{2.3}$$

As the invertibility of the transform implies that any results proved for the Kirchoff transformed problem should transfer back to the untransformed problem, (1.4), Alt and Luckhaus [9], Otto [10], and Radu et al. [20] work with (2.3) for their analytic results. For our numerical treatment, we will use (1.4), as transforming the equation, solving, then transforming back numerically can be cumbersome, and due to the transform being invertible pointwise, we can assume that any well-posedness results for (2.3) also apply to (1.4). Numerical work with the Kirchoff-transformed problem has been developed, with promising results; see Berninger et al. [48] for more details on this approach.

2.2 Assumptions on the Data

We assume that $\Omega \subset \mathbb{R}^n$ is open, bounded, and with Lipschitz boundary, $0 < T < \infty$.

Alt and Luckhaus [9] assume that $b(u)$ is merely continuous and weakly monotone, hence defines the gradient of a convex continuous function Φ, for their existence proof, and uniqueness was proven by Otto [10] under these same conditions.

However, analysis of the VGM K and θ as expanded upon in section 1.3 show that $\theta(\Psi)$ is Lipschitz continuous; i.e,

$$\exists L_\theta > 0 : \quad \theta(p_1) - \theta(p_2) \leq L_\theta(p_1 - p_2), \quad p_1 \geq p_2. \tag{2.4}$$

As such, it follows that $b(u)$, the Kirchoff-transformed $\theta(\Psi)$, is also Lipschitz continuous, with constant L_b. In the analysis of the Picard method introduced in chapter 3, we assume that

$$0 < K_{\min} \leq K(b(u)) \leq K_{\max}, \tag{2.5}$$

so that the Kirchoff transformation is invertible for all the domain, and $n > 2$ so that K is Lipschitz with respect to Ψ with constant K_b; this greatly simplifies the analysis of our scheme. In this sense, we trade the generality necessary for VGM formulations of K with $1 < n \leq 2$, and problems with b being discontinuous, such as Stefan problems [49] for greater control on the nonlinearities, and hence, our energy estimates, though it should be noted that for many practical examples, these conditions still apply [45].

We also assume that the boundary data is in $L^2(0, T; H^{1/2}(\partial\Omega))$, where $H^{1/2}(\partial\Omega)$ is in the trace of $H^1(\Omega)$, and that the initial data is in $L^2(\Omega)$.

2.2.1 Legendre Transformation of $b(u)$

In order to deal with the parabolic term, we consider the Legendre transform of $b(u)$ as introduced by [9], which due to the convexity of the primitive of $b(u)$, has the following form:

$$B(u) = \int_0^u (b(u) - b(s))\,\mathrm{d}s. \tag{2.6}$$

Some useful properties to note of $B(u)$ are:

1. $B(u) \geq 0$ for all $u \in \mathbb{R}$, with equality only for $u = 0$.

2. $$(b(u) - b(u_0))u_0 \leq B(u) - B(u_0) \leq (b(u) - b(u_0))u$$

for all $u, u_0 \in \mathbb{R}$; in particular, for every $u \in \mathbb{R}$,

$$B(u) \leq ub(u).$$

3. $|b(u)| \leq \delta B(u) + \sup_{|\sigma| \leq \frac{1}{\delta}} |b(\sigma)|, \quad \delta \in \mathbb{R}^+$.

Arguably the most important feature of $B(u)$ is motivated by the following formal calculation. If we were to assume $\frac{db}{du}$ existed in the classical sense, then

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}B(u(t)) &= \frac{\mathrm{d}}{\mathrm{d}t}\left[\int_0^{u(t)} b(u) - b(s)\,\mathrm{d}s\right] \\ &= \frac{\mathrm{d}u}{\mathrm{d}t}b(u) + u\frac{\mathrm{d}b(u(t))}{\mathrm{d}t} - b(u)\frac{\mathrm{d}u}{\mathrm{d}t} \\ &= \partial_t(b(u))u.\end{aligned}$$

In other words, $B(u)$ is the primitive in time of the product $\partial_t(b(u))u$, which is the parabolic term that needs to be controlled in energy estimates of the weak form of (2.3).

This motivates the following Lemma proved by Alt and Luckhaus (here we look at the simplified version where $u_D(x,t) = 0$).

Lemma 2.2.1 ([9], Lemma 1.5)**.** *Assume a solution u fulfills the definition 1 of a weak solution to* (2.3)*. Then $B(u) \in L^\infty(0,T;L^1(\Omega))$, and for almost all $t \in [0,T]$ the following formula holds:*

$$\int_\Omega B(u(t)) - B(u_0)\,dx = \int_0^t \int_\Omega \partial_t b(u) u\,dx\,d\tau.$$

2.3 Existence and uniqueness of solutions

We can define the analogous weak formulation for (1.4), where for the sake of simplicity in exposition, we assume the full Dirichlet problem $\Gamma_D = \partial\Omega$, though it should be noted that Alt and Luckhaus [9] considered the general problem with mixed boundary conditions:

Definition 1 (Nonlinear Weak formulation)**.** *We call $u \in L^2(0,T;H_0^1(\Omega))$ a weak solution of the initial boundary value problem* (2.3) *if the following two properties are fulfilled:*

1. *$b(u) \in L^\infty(0,T;L^1(\Omega))$ and $\partial_t b(u) \in L^2(0,T;H^{-1}(\Omega))$ with initial value $b_0 = b(u_0)$; i.e, the following holds for all $v \in L^2(0,T;H_0^1(\Omega)) \cap W^{1,1}(0,T;L^\infty(\Omega))$ with $v(T) = 0$:*

$$\int_0^T \langle \partial_t b(u), v\rangle \; dt + \int_0^T \int_\Omega (b(u) - b_0)\partial_t v \; dx \; dt = 0,$$

with $\langle \cdot, \cdot \rangle$ being the duality pairing between $H_0^1(\Omega)$ and $H^{-1}(\Omega)$.

2. *u satisfies*

$$\int_0^T \langle \partial_t b(u), v\rangle \; dt + \int_0^T \int_\Omega (\nabla u + K(b(u))e_u) \cdot \nabla v \; dx \; dt = \int_0^T \int_\Omega f(b(u))v \; dx \; dt$$

for each $v \in L^2(0,T;H_0^1(\Omega))$.

Theorem 3 ([9], Thm 1.7)**.** *Suppose the data satisfy the assumptions above, and assume that $\partial_t u^D \in L^1(0,T;L^\infty(\Omega))$. Then there is a weak solution.*

Proof. We wanted to give a rough sketch of the proof in Alt and Luckhaus [9] for the sake of completeness, and because our proof of convergence for the time-continuous Picard iteration we consider in chapter 3 uses some similar arguments. The strategy is as follows: discretize the parabolic term in time using backward Euler,

$$\partial_t^{-h} b(u) = \frac{1}{h}(b(u(t)) - b(u(t-h))),$$

to arrive at elliptic problems, which can be discretized in space using the standard Galerkin approach. For each time interval, this yields finite dimensional problems of the form for a.e $t \in (0,T)$, where

$$u_{hm}(x,t) = \sum_{k=1}^{m} \alpha_{hmk}(t) e_k(x),$$

with $h = T/(m+1)$, and the functions e_k are the first m elements in a basis in $L^2(\Omega)$ that is also orthogonal in $H^1(\Omega)$, $V_m = \text{span}\,(\{e_k(x)\}_{k=1}^m)$, with coefficients in time, $\alpha_{hmk} \in L^\infty((0,T))$. This fully discretized problem generates a finite dimensional system with continuous coefficients,

$$\phi_{hm} : \mathbb{R}^m \to \mathbb{R}^m,$$

with unknowns $\alpha = (\alpha_{hm1}, \alpha_{hm2}, \ldots, \alpha_{hmm})$. By the convexity of B in u, one can show that $\phi_{hm}(\alpha)$ has a zero for small enough h, implying the existence of the fully discrete solution $u_{hm}(t)$.

Then, standard energy estimates for the elliptic part and source terms apply. One caveat, though, is that because of the presence of $b(u)$ in the parabolic term, the standard technique of applying Theorem 2 to interchange the limits of the integral in space and derivative in time of u_t cannot be done, and so the sign of $\int_0^T \int_\Omega b(u_{hm})u_{hm}\,\mathrm{d}x\,\mathrm{d}t$ cannot be established. To compensate for this, Alt and Luckhaus use Lemma 2.2.1 to work with this integral, and define an alternative energy:

$$\sup_{0\leq t\leq T}\int_\Omega B(u_{hm}(t))\,\mathrm{d}x + ||u_{hm}||^2_{L^2(0,T;H^1_0(\Omega))} \leq C.$$

This is sufficient to provide a uniform bound of the sequence of discrete solutions in a precompact space, so that there exists some subsequence $(h_i, m_i), i \to \infty$ with $u_{h_i,m_i} \rightharpoonup u \in L^2(0,T;H^1_0(\Omega))$.

All that remained to show was that this subsequent limit is actually a weak solution to the problem. As the elliptic part can be dealt with via standard estimates, the only nontrivial part of this step was showing that $B(u_{hm}) \to B(u)$ weakly. This is accomplished by using a step function that is piecewise constant in time, and deriving an estimate of the form,

$$\int_\Omega \int_0^{T-kh} \left(b(u_{hm}(\tau+kh)) - b(u_{hm}(\tau))\right)\left(u_{hm}(\tau+kh) - u_{hm}(\tau)\right)\,\mathrm{d}\tau\,\mathrm{d}x \leq Ckh, \tag{2.7}$$

With C being independent of h and m. Since u_{hm} is piecewise constant in time, this estimate is also satisfied if kh is replaced by any positive number. Given this bound, Alt and Luckhaus then show that b and B converge pointwise a.e, via the following Lemmata:

Lemma 2.3.1 ([9], Lemma 1.8)**.** *If two mappings v_1 and v_2 in $H^1(\Omega)$ satisfy the estimates*

$$||v_i||_{H^1(\Omega)} \leq M,\ ||B(v_i)||_{L^1(\Omega)} \leq M, \quad i = 1,2,$$

and

$$\int_\Omega \left(b(v_2) - b(v_1)\right)\left(v_2 - v_1\right) \leq \delta,$$

then

$$\int_\Omega |b(v_2) - b(v_1)| \leq \omega_M(\delta),$$

with continuous functions ω_M satisfying $\omega_M(0) = 0$.

Lemma 2.3.2 ([9], Lemma 1.9)**.** *Suppose u_ϵ converge weakly in $L^2(0,T;H^1(\Omega))$ to u with*

estimates

$$\frac{1}{h}\int_0^{T-h}\int_\Omega \left(b(u_{hm}(t+h)) - b(u_{hm}(t))\right)\left(u_{hm}(t+h) - u_{hm}(t)\right)\, dx\, dt \leq C,$$

and

$$\int_\Omega B(u_\epsilon(t)) \leq C \quad \textit{for } 0 < t < T.$$

Then $b(u_\epsilon) \to b(u)$ *in* $L^1((0,T)\times\Omega)$ *and* $B(u_\epsilon) \to B(u)$ *a.e.*

Using some estimates from the proof of Lemma 2.2.1, they then bound the parabolic term of the energy estimate with test function $v_\ell \to u$ from below in terms of differences of $B(u)$'s, i.e,

$$\begin{aligned}
\int_0^t \langle \partial_t^{-h} b(u_{hm}), u_{hm} - v_{hm}\rangle \,\mathrm{d}\tau &\geq \frac{1}{h}\int_{t-h}^t\int_\Omega B(u_{hm})\,\mathrm{d}x\,\mathrm{d}\tau - \int_\Omega B(u_h^0)\,\mathrm{d}x \\
&\quad - \left(\frac{1}{h}\int_{t-h}^t\int_\Omega B(u)\,\mathrm{d}x\,\mathrm{d}\tau - \int_\Omega B(u_0)\,\mathrm{d}x\right) \\
&= \frac{1}{h}\int_{t-h}^t\int_\Omega \left(B(u_{hm}) - B(u)\right)\,\mathrm{d}x\,\mathrm{d}\tau - \int_\Omega B(u_{hm}^0) - B(u_0)\,\mathrm{d}x \\
&= \frac{1}{h}\int_{t-h}^t\int_\Omega \left(B(u_{hm}) - B(u)\right)\,\mathrm{d}x\,\mathrm{d}\tau.
\end{aligned}$$

Combining this estimate with the standard estimates for the elliptic and source terms, they get a Gronwall inequality in B, i.e,

$$\begin{aligned}
&\frac{1}{h}\int_{t-h}^t\int_\Omega \left(B(u_{hm}) - B(u)\right)\,\mathrm{d}x\,\mathrm{d}\tau + \int_0^t\int_\Omega |\nabla(u_{hm} - v_{hm})|^2\,\mathrm{d}x\,\mathrm{d}\tau \\
&\leq \int_0^t \langle \partial_t^{-h} b(u_{hm}), u_{hm} - v_{hm}\rangle\,\mathrm{d}\tau + \int_0^t\int_\Omega |\nabla(u_{hm} - v_{hm})|^2\,\mathrm{d}x\,\mathrm{d}\tau \\
&\leq C\int_0^t\int_\Omega B(u_{hm}) - B(u)\,\mathrm{d}x\,\mathrm{d}\tau + o(1),
\end{aligned}$$

with $o(1)$ being the standard Landau notation of terms that go to 0 in the continuous limit.

Then an application of Gronwall's inequality applied to the function

$$\phi = \limsup_{h\to 0, m\to\infty} \int_\Omega B(u_{hm}(t)) - B(u(t))\,\mathrm{d}x$$

shows that it becomes non-negative in the limit, and yields

$$\nabla u_{hm} \to \nabla u$$

strongly in $L^2((0,t)\times\Omega)$, for $t<T$. This then gives the convergence of the elliptic part and the source term almost everywhere, hence weakly in $L^2((0,t)\times\Omega)$, which proves that the weak limit u is a weak solution. □

To ensure that each linear problem solved on each step of the Picard linearization proposed in section 3.2 was well-posed, all of these calculations were carefully checked and verified to make sure the logic carried through, but for the sake of the reader, we do not include these computations here.

Alt and Luckhaus [9] were able to prove uniqueness in the case of b being Lipchitz and a semilinear elliptic part satisfying certain constraints:

Theorem 4 ([9], Theorem 2.4)**.** *Suppose b is a Lipschits continuous monotone increasing function, and the elliptic part*

$$a(t,x,b(z),p) = A(t,x)p + e(b(z)),$$

where $A(t,x)$ is a symmetric matrix and measurable in t and x such that for some $\alpha>0$,

$$A-\alpha I \quad \textit{and} \quad A+\alpha\partial_t A$$

are positive definite. Moreover assume that

$$|e(b(z_2))-e(b(z_1))|^2 + |f(b(z_2))-f(b(z_1))|^2 \le C(b(z_2)-b(z_1))(z_2-z_1).$$

Then there is at most one weak solution.

Clearly, the Kirchoff-transformed problem with Lipschitz K and b satisfies the conditions required, giving the desired uniqueness.

Without these assumptions, Alt and Luckhaus are only able to prove uniqueness by assuming the integrability of the time derivative of the difference of sub and super solutions u_1 and u_2. Uniqueness was later proven without this assumption by Otto [10], who used the theory of sub and super solutions of parabolic PDE to consider the difference of weak sub- and super- solutions. His goal was to prove that the set on which this difference was greater than 0 had 0 measure. The main idea of this proof of Otto is that, while the derivative of $b(u)$ doesn't have to be a function, the monotonicity of b can be used to define a nonnegative Borel measure with no atoms. This, combined with the clever use of a convex, two times differentiable auxiliary function η with other properties that he specifies in the paper, gives

the following approximate chain rule for any nonnegative test function $\gamma \in C_0^\infty((0,T)\times\mathbb{R}^n)$, where $v^0 \in H^1(\Omega)$, with $B(v^0) \in L^1(\Omega)$:

$$\langle \partial_t b(u), \eta'(u-v^0)\gamma\rangle = -\int_{\Omega\times[0,T)}\int_{v^0}^{u} \eta'(\zeta - v^0)\, b(\mathrm{d}\zeta)\partial_t\gamma.$$

This allows the time derivative to be moved to the test function, so that one can start estimating the differences $b(u_-) - b(u_+)$.

Since the test functions γ are regular, Otto further estimates the differences in time using a technique pioneered by Kružkov, introduced to estimate first order quasilinear conservation laws [50], wherein mollifiers are convolved with the differences in question to generate smoothness in time to estimate the differences of the sub and super solution b terms in time only.

This essentially splits the differences into five pieces (two each for differences in the two independent time variables, and one involving the difference $a(\nabla u_-, b(u_+)) - a(\nabla u_+, b(u_+)))$ that can each be controlled using the various properties of the elliptic term $a(\nabla u, b(u))$. From this and by assuming

$$f(w_1) - f(w_2) \le L(w_2 - w_1), \quad w_1 > w_2,$$

Otto is able to get a contraction statement of the form

$$\int_\Omega (b(u_-) - b(u_+))^+ \le e^{Lt}\int_\Omega (b_-^0 - b_+^0)^+,$$

with $(f)^+ = f$ only when $f \ge 0$, and 0 otherwise, and $b^0_{-/+}$ being the initial values of $b(u_-)/b(u_+)$.

Chapter 3
Linearization Schemes for Richards' Equation

In Richards' equation, the physical properties of the medium are represented as nonlinear coefficients, which introduce a number of challenges. The most salient of these difficulties is rapid changes in capillary head induced by the rapid change in K and θ, with the potential of blow-up in the derivative for certain materials near full saturation, as was expanded upon in Chapter 1. These coefficients can also be discontinuous for non-homogeneous porous media. Such strong nonlinearities significantly impact the efficiency of most standard nonlinear solvers, particularly when using low-order-in-time implicit schemes to solve the fully discretized system. It is for this reason that a significant portion of the numerical literature for Richards' equation focuses on the effectiveness of specific linearization techniques for solving (1.4) numerically.

Section 3.1 provides a brief overview of the most basic linearization algorithm that is typically used to compute numerical solutions of nonlinear PDE. The rest of the chapter surveys the effectiveness and features of several linearization schemes, namely the Picard iteration in 3.2, Newton-Raphson in 3.3, modified Picard in 3.4, and L-scheme in 3.5. All of these schemes can be interpreted as different combinations of zeroth and first order linearizations of (1.4), with the exception of the L-scheme [14], which is a relaxation scheme of the popular modified Picard [12] method. In section 3.2, we seek to understand the convergence properties of zeroth order linearizations by analyzing a Picard-like iteration scheme on the fully continuous problem, and we are able to prove that if both θ and K are Lipschitz continuous, with small Lipschitz constant, the Picard iteration yields a contractive sequence of solutions that converge to the unique weak solution proved to exist by Alt and Luckhaus [9] and Otto [10].

3.1 Common linearizations

In order to solve any nonlinear problem of the form

$$A(u) = f,$$

the typical procedure involves an iterative process,

1. Start with guess solution u^j,
2. Compute residual $R^j = f - A(u^j)$,
3. Solve linear problem $J(u^j)u_\epsilon = R^j$, with $J(u^j)$ being an approximation of the derivative of $A(u)$ or the value of A at the previous iteration,
4. update: $u^{j+1} = u^j + u_\epsilon$.

This process is iterated until some criterion is reached, with this criterion typically being related to the size of either the residual or the correction (in other words, either $||u_\epsilon^j|| < \epsilon$ or $||R^j|| < \epsilon$ in some norm $||\cdot||$).

The choice of approximation to the nonlinear $A(u)$ can vary, but two of the most popular are either $J(u_j) = A(u_j)$ (Picard iteration), or $J(u_j) = A'(u_j)$ (Newton-Raphson). Given the presence of two nonlinearities in (1.4), and the problems associated with these as mentioned above and in chapter 1, these two linearizations, and combinations of the two, have been proposed and tested extensively in the literature. In this section, we will define some of the more popular ones, and develop some insight into their strengths and weaknesses.

We first split the time interval $[0, T]$ into N equal size time intervals with length τ. The index n indicates the value of a function at $t = n\tau$, i.e,

$$f^n = f(x, n\tau), \quad \theta^n = \theta(\Psi(x, n\tau)),$$

etc.

Multiplying each term of the time discretized (1.4) by a test function v in a suitable solution space V and integrating over Ω, we get the following nonlinear, implicit variational problem in space for each time step t_n:

$$(\theta^n, v) - \tau(\mathrm{div}(K^n \nabla(\Psi^n + z)), v) = \tau(f^n, v) + (\theta^{n-1}, v), \quad v \in V. \tag{3.1}$$

For what follows, we drop the n time index, and introduce the j index for nonlinear iterations. We also consolidate all terms on the right hand side of (3.1) into one term, F. Then for each nonlinear iteration, we solve a variational problem

$$(\theta^{j+1}, v) - \tau \left(\text{div}(K^{j+1} \nabla(\Psi^{j+1} + z)), v \right) = F - (\theta^j, v) + \tau(\text{div}(K^j \nabla(\Psi^j + z)), v) = R^j, \quad v \in V_0, \tag{3.2}$$

where θ^{j+1} and K^{j+1} are formed using Taylor approximations of K and θ of either zeroth or first order centered at Ψ^j, multiplied by a correction Ψ_ϵ, and V_0 is a space that enforces 0 boundary conditions, i.e, in the case of discretizing in space with $P1$-Lagrange finite elements,

$$V_0 = \{v \in H^1(\Omega) : v|_{\Gamma_D} = 0, \ \frac{\partial v}{\partial n}|_{\Gamma_N} = 0\} \subset H_0^1(\Omega).$$

We then discretize the problem in space using a particular choice of V_0, solve the subsequent linear system for Ψ_ϵ and update,

$$\Psi^{j+1} = \Psi^j + \Psi_\epsilon.$$

3.2 Zeroth order (Picard) linearizations

The zeroth order linearization of (3.2) substitutes the nonlinear coefficients with their values at the previous iteration:

$$(\theta^j \Psi_\epsilon, v) - \tau(\text{div}(K^j \nabla \Psi_\epsilon), v) = R^j. \tag{3.3}$$

The Picard iteration is the simplest linearization considered. It leads to a symmetric problem being solved on each iteration, with computationally inexpensive stiffness and mass matrices to be constructed; this leads to linear problems that are both easier to discretize and construct, and to solve.

3.2.1 Continuous Picard linearization

We investigated the properties of the Picard linearization of (2.3) on the continuous level, as estimates on when numerical schemes using the Picard iteration can converge appears to be lacking in the literature. We were able to prove the convergence of the Picard iteration on (2.3) in a way that gives insight into the nature of convergence of zeroth order linear schemes for this problem.

We first define the particular continuous level Picard iteration of (2.3) we considered. We wanted to work with a Picard linearization that kept the parabolic term implicit, as the proofs of Alt and Luckhaus and Otto as outlined in chapter 2 indicate that ignoring the time derivative would over-simplify the dynamics of the Picard iteration.

For the sake of simplicity, we consider a bounded domain Ω with Lipschitz boundary, and consider the homogeneous Kirchoff-transformed Dirichlet problem; i.e, $\Gamma_D = \partial\Omega$, $u_D(x,t) = 0$, and $f = 0$. We can define a weak formulation analogous to Definition 1, where the linearized problem to solve for each iterate is as follows:

$$\int_0^T \left\langle \partial_t b(u^k), v \right\rangle \,\mathrm{d}t + \int_0^T \int_\Omega \nabla u^k \cdot \nabla v \,\mathrm{d}x \,\mathrm{d}t = \int_0^T \int_\Omega -K(b(u^{k-1}))e_u \cdot \nabla v \,\mathrm{d}x \,\mathrm{d}t, \tag{3.4}$$

$k = 1, 2, \ldots$ For the sake of clarity in the following exposition, we drop the k superscript in 3.4 in the following discussion.

3.2.2 Contraction for the Picard iteration for VGM K, θ

Theorem 5 (Contraction of the continuous in time Picard iteration)**.** *Suppose K and θ are Lipschitz continuous in their arguments, and their Lipschitz constants satisfy*

$$L_K L_b < 1. \tag{3.5}$$

Then the Picard iteration as defined in (3.4) *is a contraction.*

Proof. An immediate result of the Lipschitz continuity of b and K is that each problem in the Picard linearization scheme defined above satisfies the conditions required for the uniqueness result proven by Alt and Luckhaus, Theorem 4. The existence of weak solutions also follows from their argument, hence each Picard iteration has a unique weak solution u^k. The stability results for FE discretizations proven by [20] also apply for each Picard iterate.

Guaranteed a well-posed problem on each iteration, we wanted to show that the iterates $\{u^k\}$ form a global contraction by first proving it inductively from the first time step for small enough time, with a convergence rate independent of timestep size. Due to this independence, the same argument can then be bootstrapped on each interval, to get a global contraction. To do this, we can use the uniform bound used in the existence proof in [9], namely we assume inductively that for $j \leq k-1$,

$$||u^j||_{L^\infty(0,T;H_0^1(\Omega))} \leq M,$$

with M independent of j. Note that by definition of the nonlinear problem, this is true of u_0. Now, if both b and K are Lipschitz, then these conditions for $j = k-1$ imply the same for u^k.

Now, we divide the interval $[0,T] = \cup_{i=1}^{N}[t_{i-1}, t_i]$, $t_0 = 0$, $t_N = T$, each of length $\tau = T/N$, and take $\xi \in L^2(0,T; H_0^1(\Omega))$ to be piecewise constant in time on each interval, defined as $\xi(t) = u^k(t_n) - u^{k-1}(t_n)$ on $[t_n, t_{n+1})$. Then for the parabolic term, we can integrate the difference of the Picard iterates, i.e,

$$\begin{aligned}
&\int_0^T \langle \partial_t \left(b(u^k) - b(u^{k-1}) \right), \xi \rangle \, \mathrm{d}t = \sum_{j=0}^{N-1} \int_{t_j}^{t_{j+1}} \langle \partial_t \left(b(u^k) - b(u^{k-1}) \right), \xi \rangle \, \mathrm{d}t \\
&= - \sum_{j=0}^{N-1} \int_{t_j}^{t_{j+1}} \langle \left(b(u^k) - b(u^{k-1}) \right), \partial_t \xi \rangle \, \mathrm{d}t \\
&+ \sum_{j=0}^{N-1} \int_\Omega \left(b(u^k(t_{j+1})) - b(u^{k-1}(t_{j+1})) \right) \left(u^k(t_{j+1}) - u^k(t_{j+1}) \right) \, \mathrm{d}x \\
&- \sum_{j=0}^{N-1} \int_\Omega \left(b(u^k(t_j)) - b(u^{k-1}(t_j)) \right) \left(u^k(t_j) - u^k(t_j) \right) \, \mathrm{d}x \\
&= \int_\Omega \left(b(u^k(t_{j+1})) - b(u^{k-1}(t_{j+1})) \right) \left(u^k(t_T) - u^k(t_T) \right) \, \mathrm{d}x \\
&- \int_\Omega \left(b(u^k(t_{j+1})) - b(u^{k-1}(t_{j+1})) \right) \left(u^k(0) - u^k(0) \right) \, \mathrm{d}x,
\end{aligned}$$

due to the sum teloscoping for all $j = 1, \ldots, N-1$, and the piecewise constant nature of ξ in time. Note also that as $u^j(0) = u_0$ for all j, the last term also vanishes, leaving only the term in T. This same argument can be applied by replacing T with a.e $t \in (0,T)$, to give

$$\int_0^t \langle \partial_t \left(b(u^k)(s) - b(u^{k-1})(s) \right), \xi \rangle \, \mathrm{d}s = \int_\Omega \left(b(u^k(t_{j+1})) - b(u^{k-1}(t_{j+1})) \right) \left(u^k(0) - u^k(0) \right) \, \mathrm{d}x, \tag{3.6}$$

for almost every t.

Introducing ξ as defined above into (3.4), and taking the difference of (3.4) for u^k and u^{k-1}, we get for a.a. t,

$$\begin{aligned}
&\int_0^t \langle \partial_t \left(b(u^k) - b(u^{k-1}) \right), \xi \rangle \, \mathrm{d}s + \int_0^t \int_\Omega \nabla \left(u^k(s) - u^{k-1}(s) \right) \cdot \nabla \xi \, \mathrm{d}x \, \mathrm{d}s \\
&\leq \left| \int_0^t \int_\Omega \nabla \left[K(b(u^{k-1}(s))) - K(b(u^{k-2}(s))) \right] e_z \cdot \nabla \xi \, \mathrm{d}x \, \mathrm{d}s \right|.
\end{aligned}$$

Using (3.6) and the definition of ξ, we get, for a.a. t,

$$\begin{aligned}
&\int_\Omega \left(b(u^k) - b(u^{k-1})\right)\left(u^k(t) - u^{k-1}(t)\right)\,\mathrm{d}x \\
&+ \sum_{j=0}^{N-1} \int_{t_j}^{t_{j+1}} \int_\Omega \left[\nabla\left(u^k(s) - u^{k-1}(s)\right) \cdot \nabla\left(u^k(t_j) - u^{k-1}(t_j)\right)\right]\,\mathrm{d}x\,\mathrm{d}s \\
&\leq \left| \sum_{j=0}^{N-1} \int_{t_j}^{t_{j+1}} \int_\Omega \nabla\left[K(b(u^{k-1}(s))) - K(b(u^{k-2}(s)))\right] e_z \cdot \nabla\left(u^k(t_j) - u^{k-1}(t_j)\right)\,\mathrm{d}x\,\mathrm{d}s \right|.
\end{aligned}$$

Note that, although u^k may not be defined at t_j, by density we can approximate it with a simple function that is defined there. Given this, approximating ξ with some sequence of simple functions $v_\ell^k \in L^2([0,T], H_0^1(\Omega))$ that are piecewise constant on each (t_j, t_{j+1}), repeating the same argument with v_ℓ^k, and taking the difference of the two equations gives us that for ℓ large enough, the elliptic term becomes

$$\int_0^t \int_\Omega |\nabla u^k(s) - \nabla u^{k-1}(s)|^2\,\mathrm{d}x\,\mathrm{d}s + o(1).$$

To control the K term, we first apply Cauchy inequality:

$$\begin{aligned}
&\left| \int_0^t \int_\Omega \nabla\left[K(b(u^{k-1}(s))) - K(b(u^{k-2}(s)))\right] e_z \cdot \nabla v_\ell^k(s)\,\mathrm{d}x\,\mathrm{d}s \right| \\
&\leq \frac{1}{2}||v_\ell^k||^2_{L^2(0,T;H_0^1(\Omega))} + \frac{1}{2}\int_0^t \int_\Omega |\nabla K(b(u^{k-1}(s))) - \nabla K(b(u^{k-2})(s))|^2\,\mathrm{d}x\,\mathrm{d}s.
\end{aligned}$$

We can absorb the first term into the elliptic term on the left hand side to yield

$$\begin{aligned}
&\int_0^t \langle \partial_t \left(b(u^k) - b(u^{k-1})\right), b_\ell^k(t)\rangle\,\mathrm{d}s + \frac{1}{2}\int_0^t \int_\Omega |\nabla u^k(s) - \nabla u^{k-1}(s)|^2\,\mathrm{d}x\,\mathrm{d}s + o(1) \\
&\leq \frac{L_b^2 L_K^2}{2}||u^{k-1} - u^{k-2}||^2_{L^2(0,T;H_0^1(\Omega))}.
\end{aligned}$$

Passing now to a subsequence ℓ_n that allows the convergence of $v_\ell^k \to u^k - u^{k-1}$ pointwise a.e, we get that, for large enough n, the $o(1)$ terms become nonnegative, hence we can drop them from the inequality. This gives us the contraction

$$||u^k - u^{k-1}||^2_{L^2(0,t;H_0^1(\Omega))} \leq (L_K L_b)^2 ||u^{k-1} - u^{k-2}||^2_{L^2(0,t;H_0^1(\Omega))}. \tag{3.7}$$

In this sense, we must have that the contraction only holds if (3.5) holds. In this case, we get a contraction on the interval $[0, t_0]$, with t_0 being small enough that $||u_1 - u_0||_{L^2(0,T;H_0^1(\Omega))} < 1$.

As (3.5) is true independent of the time t, then we can bootstrap to the interval $[t_0, 2t_0]$ and repeat the argument, and continue in such a manner that we can cover the entire interval $[0, T]$. □

Remark 3.2.1. This proof gives an indication that the convergence of schemes that treat both K and b with zeroth order linearizations are affected more by the analytic properties of these functions than anything else. Indeed, the rather strong condition (3.5) implies such simplistic treatment of both nonlinearities might lead to a failure of convergence in zeroth order linearization scheme for certain problems, which corresponds well with the survey of Paniconi and Putti's [11] as described in what follows.

3.2.3 Picard iteration in the numerical literature

To analyze the effectiveness of the Picard iteration as compared to the Newton iteration, Paniconi and Putti [11] considered eight different tests of solving the h-based Richards equation (1.2) with second order (Crank-Nicolson) discretization in time with small time steps to ensure global mass balance.

In the first test, they consider a fully 3D steady state problem in which the domain Ω is a columns of soil with length L_z, for various choices of L_z. In this case $K_r = e^{\Psi}$, with zero flow conditions on the sides, and the pressure being set to 0 at depth L_z.

The Picard scheme for this problem showed difficulty with convergence as L_z was increased; indeed, at the highest value ($L_z = 30$ m), the Picard scheme failed to converge at all, and for the second highest ($L_z = 20$ m), the error couldn't breach 10^{-10}. after over 140 iterations. The Newton method showed no such sensitivity to L_z, consistently converging in $7 - 8$ iterations. In this case, they surmised that the gravity effects were not being captured well enough by 3.3, as the zeroth order approximation of the flux term wasn't capturing the effect gravity had on the change in values of K.

On the other hand, working with Picard was better for certain problems. Take for instance, test 2T from the same paper. In this test, the authors considered two dimensional transient flow in an unsaturated soil slab. The characteristic feature of this test case is the occurrence of convergence difficulties several time steps into the simulation, in response to the buildup of a sharp moisture front near the inflow boundary at $x = 0$, $6 \leq z \leq 10$. The problem was run to a simulated T of 5 days, at which point the problem reached steady state. They allowed for variable time step sizes, where the step size could be increased or decreased depending on how many iterations were needed to converge for the next time step. They also incorporated

a back stepping procedure that would be used when a linear scheme failed to converge on a certain time step; this was used more frequently for the Newton scheme than the Picard, hence the total number of time steps needing to be solved being greater for Newton in general. In this problem, a variant of the VGM formulation was used, with $n = 4$. The authors consider the number of linear solves needed for each of 10 different strategies, including the effect of mass lumping, simple Backward Euler time stepping, finer space approximations, and fixing the time steps.

In the case of initially steep saturation fronts with strong nonlinearities, the Picard iteration, though requiring more iterations per timestep in some cases, tended to require less time steps to fully simulate the transient problem; this is due to the fact that the steep saturation fronts present at the beginning of the simulation due to the relatively dry soil initial condition generated linear problems for the Newton iterations that were difficult to solve without restricting the timestep size. Paniconi and Putti [11] consider two other problems with steep saturation fronts that feature similar behavior, though they do mention that near the end of these iterations (i.e, approaching the equilibrium state), the Newton iterations outperform the Picard iterations considerably, as the algorithm used to adjust time steps increased the final time steps of the Newton iteration to several orders of magnitude larger than the Picard final time steps.

3.3 First order (Newton-Raphson) linearization

The full Newton linearization of (3.2) can be seen as approximating both nonlinearities with a first order expansion centered at Ψ^j:

$$\theta^{j+1} \approx \theta^j + \partial\theta^j \Psi_\epsilon,$$

$$K^{j+1}\nabla(\Psi^{j+1} + z) \approx K^j \nabla(\Psi^j + z) + \left(\partial K^j \nabla(\Psi^j + z) + K^j \nabla(\cdot)\right) \Psi_\epsilon.$$

Substituting these into (3.2), we get the variational problem

$$(\partial\theta^j \Psi_\epsilon, v) - \tau(\mathrm{div}(\partial K^j \nabla(\Psi^j + z)\Psi_\epsilon + K^j \nabla\Psi_\epsilon), v) = R^j. \tag{3.8}$$

In particular, this method is desirable because, provided a good initial guess and locally

Lipschitz nonlinearities, convergence of the Newton method is quadratic; i.e,

$$||u_{k+1} - u_k|| \leq C||u_k - u_{k-1}||^2,$$

meaning that when the derivative of K and θ are bounded, the Newton iteration will significantly reduce the number of linear solves needed to converge within a set tolerance. This of course, tends to come at the increased cost of construction of the fully discrete system, and in the particular case of this problem, the addition of a nonsymmetric term, $(\partial K^j \nabla(\Psi^j + z)\Psi_\epsilon, \nabla v)$. Nonsymmetric problems are in general harder to solve than symmetric problems, and the presence of the derivative of the conductivity in the nonsymmetric part implies that for certain materials, on any given linear solve there will be portions of the domain where the Jacobian of K will dominate K and vice versa, which will greatly increase the computational cost of each linear solve.

The full Newton method for (1.4) has been tested extensively in the literature ([17], [31], [32], [24], [44]). In particular, much of the focus is on working with problems that feature initially dry media on which a saturation front is imposed, typically as a boundary condition. In this case, the VGM K relations dictate large spatial gradients in K near the saturation front for materials with sufficiently non-homogeneous micropore structure ($n < 2$).

As an emblematic test of this type, consider example 1 in [44]. In this problem, Lehmann and Ackerer consider a column of soil with length $L = 30$ cm, and impose initial and boundary conditions $\Psi(z, 0) = -1000$ cm, $\Psi(0, t) = -75$ cm, $\Psi(L, t) = -1000$ cm. The VGM parameters were $K_s = 9.22 \times 10^{-3}$ cm/s, $\theta_s = 0.368$, $\theta_r = 0.102$, $\alpha = 0.0335$, $n = 2$. The time-transient (1.4) was simulated until a final time $T = 6$ hours. The space and time step discretization parameters were $\Delta z = 0.25$cm, and an algorithm allowed varying Δt that could change adaptively depending on how many iterations were required to solve for the previous time step. To compare the approximate solutions, a dense grid solution Ψ^* with uniform grid $\Delta z = 0.1$cm, and fixed timestep of $\Delta t = 0.1$s was computed.

In this test, the authors observe that the Newton scheme is superior to the Modified Picard iteration in nearly every metric. This again makes sense, considering that the modified Picard linearization doesn't take the gradient of K into account.

The next test Lehmann and Ackerer [44] consider monitors the effect of heterogeneity in the material. They consider a three-layered problem in one dimension, with relatively wet initial conditions $\Psi(z, 0) = -100$cm, a flux condition at $z = 0$ that varied in time, relating to rainfall and water evaporation, and a free drainage condition at the lower boundary. The first and third layers had VGM parameters $K_s = 541$ cm/day, $\theta_s = 0.3658$, $\theta_r = 0.0286$,

$\alpha = 0.0280$, $n = 2.239$ (Berino loamy fine sand), and second layer $K_s = 513.1$ cm/day, $\theta_s = 0.4686$, $\theta_r = 0.1060$, $\alpha = 0.0104$, $n = 1.3954$ (Glendale clay loam). The simulation was run for 60 days, with variable time steps.

As their simulations show, this is an example where the modified Picard iteration outperforms the Newton iteration in certain aspects. However, when they changed the initial condition ($\Psi(z, 0) = -10000$cm), the oppsite was true– the Newton method outperformed the modified Picard iteration by as much as 8 times fewer number of timesteps and iterations, with roughly the same solution accuracy.

One final thing the authors note is that in both the initially wet and dry problems, creating a hybrid algorithm where one starts with modified Picard iterations, and then switches to Newton after some criteria is reached (such as some combination of the residual and correction norms being less than some tolerance) consistently leads to the lowest CPU time and number of iterations.

Finally, as the examples of these two different papers shows, the effectiveness of the Newton scheme appears to depend on whether one uses the head-based (1.2) or the mixed saturation-head based (1.4) formulations. In [51], various numerical experiments into initially dry layered media in one dimension show that Newton solves of the head based formulation (1.2) restricts the time step size significantly (up to four orders of magnitude smaller than the water based formulation) to ensure convergence for the initial modeling of saturation fronts in transient problems, even for layered soils where the water content can be discontinuous across layers. However, due to the degeneracy present for the water content based RE (1.3) at full saturation, pure water content based formulations become singular for mixed saturated-unsaturated problems. Kirkland et al. suggest a transformed RE that will change the primary unknown dependent on which regime the soil is in at a point in space and time [52].

Forsyth et. al [24] further develop this idea by proposing a simple algorithm to switch variables dependent on water saturation, where if the saturation at a point is over a set tolerance tol_f, the scheme uses the pressure head based form, while if the saturation is less than another tolerance tol_b, then the water content is used as primary variable. To avoid complications in computing the Jacobian analytically with variable switching, they suggest numerical computation of the Jacobian using finite differences. They combine this with upstream weighting of physical parameters to obtain a monotone discretization and observe fast convergence for several mixed unsaturated-saturated examples.

3.4 Modified Picard linearization

The modified Picard iteration, introduced by Celia et al [43] and later refined in [12], modifies the Newton iteration by taking the first order expansion of θ centered at θ^j, and zeroth order expansion of K centered at j, yielding the problem

$$(\partial\theta^j\Psi_\epsilon, v) - \tau(\text{div}(K^j\nabla\Psi_\epsilon), v) = R^j. \tag{3.9}$$

The reason for doing this is motivated by mass conservation: by doing a first order expansion in the water content term, the source term automatically corrects water mass balance for each iteration. This can be seen on the discrete level: summing up all of the water content source terms of form $\theta^{j-1} - \theta^{n-1}$ at each node in a first order lumped mass nodal finite element discretization or a finite difference discretization will show that the mass balances cancel out completely on the interior of the domain, with the only potential mass imbalances at the boundary being dealt with by boundary conditions [12]. Further, without the nonsymmetric Jacobian contribution for the conductivity, discretizations yield symmetric linear systems that are much easier to solve than the Newton systems. The experiments of Paniconi and Putti [11] and Lehmann and Ackerer [44] confirm that for any test where the total number of linear solves for modified Picard is roughly the same as those for Newton, the Newton scheme takes much more CPU time; this is partially due to the added cost of constructing the Jacobian, but is mostly due to the added time required to solve the resulting nonsymmetric linear system.

Due to the perfect discrete mass conservation and ease of solving the resulting linear systems, modified Picard has become the standard numerical approach for Richards' equation. It is the method that is used in many production codes, including the USDA Hydrus-1D Richards' equation solver that is used to simulate large scale groundwater flow problems in one dimension [53].

This is not to say that the modified Picard iteration is the best for all situations; as explained in the previous section, when simulating scenarios featuring infiltration into dry soils, the Newton method for the mixed form (1.4) that features the gradients of the conductivity is superior in the number of iterations per timestep, so much so that the Newton scheme consistently solved in less CPU time than modified Picard. However, due to the sensitivity of Newton convergence to initial guess, robustness is a common issue that has arisen in much of the work with Newton based approaches. In addition to the difficulties associated with poor initial guesses, the addition of steep gradients may also lead to overcorrection of

iterates that the less conductivity-sensitive modified Picard scheme doesn't struggle with, as was shown in the work of Lehmann and Ackerer [44]. Even worse, we showed in chapter 1 that for sufficiently irregular micropore structure ($1 < n \leq 2$) the conductivity gradient is either discontinuous at $\Psi = 0$, or even blows up. When this happens, Newton-Raphson is no longer guaranteed to converge locally. Miller, et al. [45] attempt to tackle this issue by approximating the nonlinearities near these points with various interpolations, concluding that cubic spline interpolations near $\Psi = 0$ and using integral approximations of the Jacobian with high order quadrature compensated for these problems to an extent. In light of these issues, modified Picard, which features none of these added complications, is a desirable alternative.

To deal with slow convergence of modified Picard in cases with steep conductivity gradients, modified Picard with Anderson acceleration has been suggested [54]. This acceleration technique can be regarded as a nonlinear version of the GMRES method for iterative solution of linear systems [55]. The intuition of the authors was that, even with the additional cost of solving the least squares problem in the previous iterates, the Newton method would still be slower, due to the additional computational costs that come with solving a nonsymmetric problem on each step. To test this hypothesis, they considered various tests, the first one establishing the proper choice of the minimal residual tolerance and number of previous solutions to store for a variety of VGM parameters, finding that $m = 5$ and $tol = 1 \times 10^{-3}$ were the best choices. They then ran comparisons of Newton's method with variable correction step and line search optimization, and an optimized form of Anderson accelerated modified Picard with $m = 5$ on a 2D steady state infiltration problem with 18 different sets of VGM parameters, on a 1024×1024 grid. The results showed that the modified Picard iteration was by far the most robust with respect to different parameters, and while the Newton method was fastest on about half of the problems, the accelerated modified Picard took no more than 350% as much CPU time to solve, and even beat out the Newton method in about 35% of the test problems.

For a problem of infiltration into initially dry media simulated twice two different timescales, Anderson accelerated modified Picard was clearly more robust, as it worked for both time step sizes, while Newton didn't converge at the 1024 mesh size for the larger time step. When Newton did work, however, it tended to solve roughly as fast as Anderson accelerated modified Picard.

Thus while modified Picard can have problems converging when K gradients are large, there are methods to help combat this, making modified Picard a strong choice for linearization.

3.5 L-scheme linearization

Finally, the L-scheme is a relaxation method for the Modified Picard iteration, first introduced in [14], which simplifies (3.9) even further by replacing the Jacobian term $\partial\theta^j$ with some constant upper bound, $\infty > L_\theta \geq \sup_\Psi |\theta'(\Psi)|$:

$$(L_\theta \Psi_\epsilon, v) - \tau(\text{div}(K^j \nabla \Psi_\epsilon), v) = R^j. \tag{3.10}$$

The purpose of this modification of the modified Picard iteration is to add robustness to the modified Picard iteration for saturated-unsaturated problems. In [14], Slodička was able to show convergence of the L-scheme on each iteration from arbitrary initial guess data, by assuming that θ and K are Lipschitz in their arguments, and $0 < K_{\min} \leq K \leq K_{\max}$.

This is accomplished by showing that on each fixed timestep, the iterative solutions $\{\Psi_j\}$ yielded from solving (3.10) form a Cauchy sequence in $L^2(\Omega)$, approaching the solution to the nonlinear elliptic problem (3.1) for the current timestep, Ψ^n. The rate of convergence is given by

$$||\Psi^j - \Psi^n||_{L^2(\Omega)} \leq \left(1 - \frac{\tau\lambda}{L + \tau\lambda}\right)^j ||\Psi^0 - \Psi^n||_{L^2(\Omega)},$$

where $0 < \lambda = \lambda(\Omega, K_{\min})$ is the same for all of the nonlinear elliptic problems (3.1). This implies that the the L-scheme linearizes the nonlinear elliptic system by forming a contractive map that yields convergence for any $\tau > 0$ on each timestep, though it should be noted that the number of iterations to reach a desired tolerance is dependent on the choice of initial guess Ψ^0, on how large the time step is, and implicitly on $K_{\min}$ and the geometry of the domain Ω. Because of this, convergence, while guaranteed, has the potential to be very slow, and tends to work better for larger time steps.

In order to circumvent this potential slow down, List and Radu suggest using the L-scheme as a stabilizer for Newton-Raphson [13]; by using the L-scheme a few times and supplying the result as the first guess for a Newton method, this would significantly increase the robustness of Newton-Raphson and provide the speed up necessary for cases with steep K gradients.

More specifically, the nonlinear iteration described in [13] starts with L-scheme relaxation, and then after a criterion of the form

$$||\Psi^j - \Psi^{j-1}|| \leq \delta_a + \delta_r ||\Psi^j||$$

is reached for some specified $\delta_a, \delta_r > 0$ (or after some fixed number of L-scheme iterations), would switch to Newton-Raphson.

The authors test their method on two problems that feature mixed saturated-unsaturated conditions in two dimensions. In the first problem, they solve a two dimensional problem with injection and extraction in the vadose zone $\Omega_{vad} = (0,1) \times (-3/4, 0)$ above the groundwater zone $\Omega_{gw} = (0,1) \times (-1, -3/4]$. For boundary conditions and initial conditions, they take a fixed pressure $\Psi_{vad} = -3$ on the surface $(0,1) \times \{0\}$, and no flow conditions on the remaining sides of the square. Initial pressure profile is given as $\Psi^0 = \Psi_{vad}$ on Ω_{vad}, and a linear scaling pressure, $\Psi^0 = -(z + 3/4)$ on Ω_{gw}.

List and Radu [13] tested Newton, modified Picard, L-scheme for two choices of L, one slightly larger than $L_\theta = \sup_\Psi |\theta'(\Psi)|$, and one closer to $L_\theta/2$, mixed Picard/Newton (as done in [44]), and mixed L-scheme/Newton with the two choices of L (with switching parameters $\delta_a = 2$, $\delta_r = 0$) to solve the first time step. To discretize the system, they applied backward Euler with various choices of τ, and discretize in space using $P1$-Lagrange finite elements, with a uniform mesh with sizes $\Delta x = \Delta z = \frac{1}{10}, \frac{1}{20}, \frac{1}{30}, \cdots, \frac{1}{80}$.

The results of their tests showed that the two L-schemes always converged, though the one with the undershot value did better for all time steps; this robustness was lacking for the (modified) Picard iteration and especially the Newton method, which didn't converge at all by itself, except on the smallest time step size. However, whenever the Picard iteration or Newton iteration converged, it converged in significantly fewer iterations than either L-scheme, with those methods finishing in less time despite the decreased cost in not having to compute the Jacobians for θ and K. However, the slow convergence of the L-scheme was overcome by combining it with Newton's method after a few iterations of the L-scheme as a relaxation. This method always worked well, and led to much faster convergence, only barely being beat by the Picard/Newton combination for the smallest timestep.

For the second problem in [13], they consider a time transient problem with a boundary condition on the top that varies from being negative in time to becoming positive, and groundwater pressure on the bottom third of the domain $\Omega = (0,2) \times (0,3)$. They test with the VGM parameters for silt loam and Beit netofa clay (see chapter 1), and show that for the silt loam case, the L-scheme again perform the worst in terms of number of iterations, but when combined with the Newton scheme (which performed the best wrt number of iterations), they improve considerably, even beating out the Newton scheme on overall CPU time, mostly due to the good condition of the L-scheme systems and the reduced cost in construction. For the Beit Netofa clay parameters, due to steep gradients in K, the Newton

method again has the least number of iterations, but also takes the most time to compute, as the condition number of the systems being solved explode, mostly due to the discontinuity of the K gradients at the saturation fronts. The L-schemes are insensitive to this, and thus finish their computations in less time; indeed, this is the one case where the L-scheme without Newton outperforms the combination of the two, mostly due to the poor conditioning of the Newton linear systems.

One potential class of problems that the L-scheme may struggle with is in dealing with heterogeneous problems, such as layered soil examples, as the water content is discontinuous across the layers' boundaries, implying its derivative could become unbounded, sending appropriate L to infinity. There appears to be no mention of this in the literature involving the L-scheme, though preliminary results with layered media with saturation fronts in initially wet and dry layered as discussed in [44] indicate that the L-scheme may very well still perform well, as modified Picard appears to work better than Newton for relatively wet initial conditions.

Chapter 4
Numerical implementation

In this chapter, we describe our discretization strategy for the linearizations discussed in the previous chapter. In time, we always use fully implicit Backward Euler, and finite elements in space. Section 4.1 defines the two finite element discretizations we employ, the mixed finite element scheme [56] and the Edge Average Finite Element (EAFE) [19] scheme. Mixed finite elements enjoy local mass conservation on the continuous level, and is the only finite element scheme to guarantee perfect mass conservation without the need for post processing schemes. EAFE is an alternate formulation of the standard first order Lagrange finite elements which reorders the bilinear form from sums of products of values at vertices in a mesh to products of differences of values along edges in a mesh. Due to this, discretizing with EAFE guarantees monotonicity, in the sense of the resulting linear systems having positive diagonal entries and negative off-diagonal entries, with the negative sum of the off diagonals being less than the diagonal value; this property, known as the M-matrix property, guarantees that there will be no non-physical oscillations in the solution. This is typically not possible to guarantee without special techniques, such as mass-lumping [12] and upwinding [27], and that EAFE provides this automatically, provided the mesh satisfy certain mild geometric constraints, is very desirable. In our work we fully discretize with the mixed finite elements, and use the EAFE formulation as an auxiliary discretization in the preconditioner we propose in Chapter 5. We give stability results in the form of error estimates for each discretization. Section 4.2 serves as a survey of two of the linear iterative solvers we use for the resulting systems of linear equations, namely the conjugate Gradient method for symmetric positive definite problems, and generalized minimal residual for nonsymmetric problems. Section 4.3 serves as a survey of multigrid, which we use as both solver and preconditioner.

4.1 Finite element discretizations

To discretize any of the linearized variational problems as described in chapter 3, we use finite elements in space. Finite elements, as defined in Ciarlet [57] can be viewed as a triple $(\mathcal{T}_h, \mathcal{V}_h, \Sigma_h)$, where $\mathcal{T}_h$ is some polygonal approximation of $\Omega \subset \mathbb{R}^d$, $\mathcal{V}_h$ is a finite dimensional space of functions defined on $\mathcal{T}_h$, and Σ_h is a basis for the dual of $\mathcal{V}_h$; i.e, a set of functionals

$$\{\sigma_i\}_{i=1}^{|\mathcal{V}_h|} : \quad \mathcal{V}_h \to \mathbb{R},$$

with the property that the map

$$p \ni \mathcal{V}_h \to \Big(\sigma_1(p), \sigma_2(p), \ldots, \sigma_{|\mathcal{V}_h|}(p)\Big) \in \mathbb{R}^{|\mathcal{V}_h|}$$

is bijective. These functionals are typically called the nodal basis, or more generally, the degrees of freedom of the system.

Using finite elements, one can then approximate (3.1) with a finite dimensional form, namely, the problem of determining $\Psi_h^n \in V_h$ such that, for all $v_h \in V_h$,

$$(\theta^n, v_h) - \tau(\mathrm{div}(K^n \nabla(\Psi_h^n + z)), v_h) = \tau(f^n, v_h) + (\theta^n, v_h). \tag{4.1}$$

For our model problems, we approximate Ω with a tesselation of d-dimensional simplices of uniform size h, $\mathcal{T}_h = \cup_{i=1}^{n_T} T_i$. For the finite dimensional subspaces V_h, we chose the standard RT0-$P0$ mixed finite elements, and work with a particular variational form of the standard $P1$-Lagrange finite elements that we use to precondition a part of the resulting saddle-point problem.

In this section, we define these two different finite element discretizations of the form (4.1) and discuss several pertinent aspects of these approximations, namely

1. Are the finite elements stable for the problem we consider? i.e, are the resulting finite dimensional problems well-posed?

2. With what truncation order do our finite element solutions approximate the true solution?

3. What is the number of degrees of freedom of the resulting finite dimensional problem?

4. What physical properties do the finite element bestow upon the approximate solutions

(i.e, do they conserve mass, can we ensure that the approximate solution is free of non-physical osctillations)?

4.1.1 Edge average finite element discretization

Arguably the most popular finite element discretization of (4.1) uses the standard $P1$-Lagrange finite elements, where

$$V_h = \{v_h \in H_0^1(\mathcal{T}_h) : v_h|_{T\in\mathcal{T}_h} \in \mathbb{P}^1\},$$

with $\mathbb{P}^1$ being the space of linear polynomials, and the degrees of freedom are the values of the function at the NV vertices $\{x_i\}_{i=1}^{NV}$ of triangulation $\mathcal{T}_h$, with functional basis

$$\sigma_i(p) = p(x_i).$$

The induced basis on V_h is then the set of linear tent shape functions $\{\varphi_i\}_{i=1}^{NV}$, where

$$\varphi_i(x) = \begin{cases} 1, & x = x_i \\ 0, & x = x_j, j \neq i. \end{cases}$$

When combining this with the saturation-head based Richards equation (1.4), several authors have confirmed that global mass balance is perfect for the system being considered ([12], [43], [44], [58]); indeed, this global mass conservation comes from the saturation-head based Richards equation being in conservation form, and the mass conservation inherent in the solution satisfying the discrete weak form, where for any $v_h \in V_h$, the solution at each fixed time t^n, Ψ_h^n satisfies

$$(\theta^n, v_h) + \tau(K^n\nabla(\Psi_h^n + z), \nabla v_h) = \tau(f^n, v_h) + (\theta^n, v_h). \tag{4.2}$$

There are, however, potential problems inherent in using this approach. As pointed out ([12], [27], [44], [14], [11]), in problems involving steep saturation fronts or any other problems where the water content can deviate significantly, the discrete analog of the maximum principle for the elliptic problem being solved on each timestep may be violated depending on discretization parameters (specifically, based on the ratio of the time step size τ and the characteristic mesh size h).

This is a problem that is inherent in the discretization of elliptic problems of the form (4.2), and is a well-documented phenomenon in the literature. The authors of [59] illuminate some

of the difficulties regarding this issue for $P1$-Lagrange discretizations of the Poisson problem. The discrete maximum principle for the diffusion problem

$$(D(x)\nabla u_h, \nabla v_h) = (f, v_h), \quad v_h \in V_h,$$
$$u_h - \tilde{u}_h \in V_h,$$

with homogeneous data $f = 0$ and positive definite diffusion tensor $D(x)$ can be stated as

$$||u_h||_{\infty,\mathcal{T}_h} \leq ||\tilde{u}_h||_{\infty,\partial\mathcal{T}_h},$$

With $||\cdot||_\infty$ being the $L^\infty(\mathcal{T}_h)$ norm. In particular, for the Poisson problem ($D = I$), Scott et al. [59] showed that if certain geometry constraints on the mesh are not satisfied, then the discrete Green's operator $G_y^h(x)$ that solves the problem above with $f = \delta_y$, with δ_y being the standard Kronecker delta distribution centered at y, attains negative values; this in itself implies that there exists values $u_h \geq \tilde{u}_h$ inside the domain, violating the discrete maximum principle, hence allowing for non-physical solutions.

The geometry constraints in question involve the interior angles of elements of the the mesh in question; in two dimensions, if the sum of angles opposite to an edge is greater than π, then the discrete max principle will be violated, leading to nonphysical solutions. There are many conditions that one can use to enforce the monotonicity of elliptic discretizations. One of the simplest is to have a sequence of triangulations $\{\mathcal{T}_h\}_{h\to 0}$ be *shape-regular*, which implies that there exists a constant $C > 0$ such that

$$\max_{T\in\mathcal{T}_h} \frac{\operatorname{diam}(T)^d}{|T|} \leq C,$$

uniformly in h. In two dimensions, this is equivalent to the existence of a uniform lower bound with respect to h of the minimal angle of any element in the mesh.

An even weaker condition that can still guarantee monotonicity in two dimensions is the notion of a *Delaunay triangulation*, which can be interpreted as the dual triangulation induced by the Voronoi diagrams of a set of finite points in a bounded domain $\Omega \subset \mathbb{R}^d$. In two dimensions, one can show that the (unique) Delaunay triangulation of a mesh maximizes the smallest angle in any triangle of the mesh, and also minimizes the maximal diameter of any triangle in the mesh. One can also show that $P1$-Lagrange interpolations of any function $f \in H_0^1$ obtain a minimal H_0^1 semi-norm on Delaunay triangulations; in this sense, these triangulations can be seen as an ideal choice to work with $P1$-conforming finite elements.

Further, given a finite set of nodes in two or three dimensions, constructing their Voronoi diagram is relatively inexpensive at $\mathcal{O}(N \log N)$ operations. Refining the mesh is also relatively simple, as one can use an alternate definition of a triangulation being Delaunay, namely, that the Delaunay triangulation of a set of nodes is the unique triangulation in which the circumsphere of each triangle contains no other point in the mesh, to choose new points, or to switch edges as necessary. Such procedure can be done in $\mathcal{O}(n \log n)$ operations as well, implying that Delaunay triangulations in 2D are very amenable to unstructured meshes and automatic mesh refinement procedures.

However, it should be noted that in three dimensions, the Delaunay triangulation no longer guarantees a maximal minimum angle between faces, and in practice, going through this Delaunay triangulation and then using a circumsphere criterion to edge swap also cannot guarantee convergence to the actual mesh that maximizes the minimum angle, as the order in which you proceed through the triangulation affects this process. What is true however is that if a 3D triangulation has the property that the center of the circumsphere for a given simplex in the triangulation lies inside the element, then that triangulation is automatically Delaunay, and in more recent times, a randomized edge swapping method has been introduced that guarantees convergence to the Delaunay triangulation with high probability. We refer to [60] for more details on Delaunay triangulations.

Finally, given a positive definite diffusion tensor $D(x)$, one can refine the circumsphere property that a Delaunay triangulation must satisfy as one in which no node of the mesh lies inside the circumsphere of a given element of the triangulation, with metric induced by the diffusion coefficient D, i.e, using the norm

$$||x||_D^2 := x^T D x.$$

In practice, this metric transforms the circumspheres into ellipsoids, which imply the generation of mesh refinements that naturally adapt to any anisotropies that might be present in the diffusion coefficient, as well as any sharp variations in D, for instance near discontinuities.

However, even with a grid that guarantees that a discrete maximum principle holds for the elliptic part, non-physical oscillations may still exist for the discretization of the parabolic problem (4.2). The reason for this can be best illustrated by considering the discretization of the problem in one dimension for a uniform grid, as was done in [12].

Given a one dimensional domain split into equal length intervals of length $h = L/NV$, $\Omega = [0, L] = \cup_{i=0}^{NV-2}[x_i, x_{i+1}]$, with $x_0 = 0$, $x_{NV} = L$, the $P1$-Lagrange discretization of (4.2) forms a linear system of equations with unknowns $P = [P_0, P_1, \ldots, P_{NV-1}]$, where P_i is the

value of Ψ_ϵ at node x_i.

For a node i in the interior, when using the modified Picard, L-scheme, or Picard linearizations, the linear system to be solved is of the form

$$\begin{aligned}
&\frac{h}{6}J\theta_{i-1}P_{i-1} + \frac{h}{3}J\theta_i P_i + \frac{h}{6}J\theta_{i+1}P_{i+1} + \frac{\tau}{h}\left[-K_{i-1}P_{i-1} + (K_{i-1}+K_i)P_i - K_{i+1}P_{i+1}\right] \\
&= P_{i-1}\left(\frac{h}{6}J\theta_{i-1} - \frac{\tau}{h}K_{i-1}\right) + P_i\left(\frac{2h}{3}J\theta_i + \frac{\tau}{h}(K_{i-1}+K_i)\right) + P_{i+1}\left(\frac{h}{6}J\theta_i - \frac{\tau}{h}K_i\right) \\
&= F_i,
\end{aligned}$$

where the coefficients K and $J\theta$ are taken to be piecewise constant on the intervals $I_k = [x_k, x_{k+1}]$, $k = i-1$, $i.$, and $J\theta$ is the approximation of θ^n used in each symmetric linearization.

Note that depending on the values of physical parameters $J\theta$ and K, and the discretization parameters h and τ, the coefficients of the off-diagonal terms corresponding to P_{i-1} and P_{i+1} can be made positive. This would result in a coefficient matrix that can (and in practice, typically does) produce solutions that violate the discrete maximum principle. This violation often manifests itself in the form of non-physical oscillations, particularly around points where the physical parameters vary significantly, i.e, near infiltration fronts into dry media. In particular, given fixed mesh size and timestep, and that the sign of both θ and its derivative $\partial\theta$ are positive, a condition for these off-diagonal terms to be negative for all i is

$$h < \left(6\tau\frac{\min_i K_i}{\max_i J\theta_i}\right)^{1/2}.$$

This dependence of mesh size to the timestep size, and the values of K and $J\theta$ imply potential violations of the discrete max principle when there are large changes in the Jacobian of the water content or when K is small (i.e, dry media), which is consistent with findings in the literature [44], [14], [43].

There have been many suggested remedies for this issue. The most popular suggestion is also the simplest; namely, [12] suggest that instead of using the consistent (distributed) mass discretization for the $J\theta$ term, one can use a lumped mass discretization. This discretization can be understood as a lower order quadrature of the mass bilinear form, in which one "lumps" all of the distributed mass onto the central weight; i.e, set the coefficients in the off-diagonal mass matrix as 0, and set the diagonal entry as the sum of the lumped masses. This entirely removes positive contributions to the off-diagonal entries of the coefficient matrix, and hence eliminates the potential for non-physical oscillations to occur. As such, we also use the

lumped mass discretization for our $P1$-Lagrange formulation.

Another potential source of non-physical oscillations was described by Forsyth et al. [27]. In this work, the authors show that the choice of K used in constructing the stiffness matrix is also influential in preventing non-physical oscillations from occuring at "homogeneous" interior nodes (i.e, nodes where there are no sharp discontinuities in material properties). In multiple dimensions, Forsyth et al. were able to show that for each nonlinear iteration, the weights K_i should be chosen via an upwind algorithm. If one of the symmetric linearizations in two dimensions for a uniform shape regular mesh are used, then on the jth nonlinear iteration, the following system of linear equations must be solved for P_i:

$$P_i \left(\int_\Omega J\theta_i^j \varphi_i^2 \, \mathrm{d}x \right) + \sum_{k \in \Omega_i} P_k K_{i,k}^j \gamma_{i,k} = R_i^j,$$

where the weights

$$\gamma_{i,k} = \int_\Omega \nabla\varphi_k \cdot \nabla\varphi_i \, \mathrm{d}x,$$

and Ω_i is the set of vertices x_k with $\mathrm{Supp}(\varphi_k) \cap \mathrm{Supp}(\varphi_i) \neq \{0\}$.

In this context, the upstream weighting suggested in [27] is of the form

$$K_{i,k}^j = \begin{cases} K_k^j = K(\Psi_h^j(x_k)), & \gamma_{i,k}(\Psi_k^j - \Psi_i^j) > 0, \\ K_i^j = K(\Psi_h^j(x_i)), & \gamma_{i,k}(\Psi_k^j - \Psi_i^j) \leq 0. \end{cases} \tag{4.3}$$

The authors proceed to show that discretizations using this upwinding approach are unconditionally monotone, implying there will be no non-physical oscillations, while using discretizations with average weighting and centroid weighting (where one approximates K locally using the centroid of the simplices, and assembles the matrix using these terms) of K can lead to non-physical oscillations, particularly for heterogeneous media where K may become discontinuous. However, this choice can slow down the simulation, as this choice needs to be made before every linear iteration.

A more attractive choice as suggested by Zaidel and Russo [61], is to use the arithmetic mean saturation (KAMS):

$$K_{i,k} = K\left(\frac{\theta_i^j + \theta_k^j}{2} \right). \tag{4.4}$$

C.T. Miller et al. [45] performed numerical solutions using $P1$-Lagrange finite elements with high order time integration on the head-based RE (1.2) for various infiltration problems, linearizing the problem with modified Picard, and interpolating K with either KAMS, an

integration based interpolation, or the central weighting approach, which is just an arithmetic average of K at the two points. What they found was that the KAMS K was the most robust with respect to choice of material, in that the error $||\frac{\Psi_* - \Psi_h}{\Psi_*}||_{\ell^2}$ for some solution Ψ_* computed on a dense grid was close to the smallest for all examples, and also led to convergence of modified Picard for the most simulations.

For our work, we wanted to use a discretization of (4.2) that could yield a monotone discretization for all linearizations, including for the full Newton linearization. This linearization for $P1$-Lagrange finite elements yields the following sequence of nonsymmetric linear problems: Find $\Psi_h \in V_h$ such that, for all $v_h \in V_h$,

$$(\partial\theta^j \Psi_{h,\epsilon}, v_h) + \tau(\partial K^j \nabla(\Psi_h^j + z)\Psi_{h,\epsilon} + K^j \nabla\Psi_{h\epsilon}, \nabla v_h) = R^j. \tag{4.5}$$

This can be interpreted as a time-transient convection-diffusion problem, with the conductivity Jacobian term playing the role of the convection. As such, we decided to use an alternate discretization of the problem that has been shown to produce monotone discretizations of convection-diffusion problems with the milder restriction of the triangulation being Delaunay with respect to the diffusion coefficient, K^j. This scheme is called the Edge Averaged Finite Element scheme, or EAFE for short.

This scheme, formally introduced in [19], was partially based on the work of Markowich and Zlamal [62] and Brezzi, Marini and Pietra [63], who each introduced discretizations of an electron drift model of the form

$$-\operatorname{div}(\nabla u + \nabla\psi) = f, \quad x \in \Omega.$$

The authors were able to generalize these works, and proposed a method for convection-diffusion problems of the form,

$$\begin{cases} \mathcal{L}u \equiv -\operatorname{div}(\alpha(x)\nabla u + \beta(x)u) = f(x) & x \in \Omega, \\ u = 0 & x \in \partial\Omega, \end{cases} \tag{4.6}$$

with assumptions $\alpha \in C^0(\bar{\Omega})$ with $0 < \alpha_{min} \leq \alpha(x) \leq \alpha_{max}$ for every $x \in \Omega$, $\beta \in \left(C^0(\bar{\Omega})\right)^2$, and $f \in L^2(\Omega)$. This problem admits the following weak formulation:

Find $u \in H_0^1(\Omega)$ such that

$$a(u, v) = f(v), \qquad \text{for every } v \in H_0^1(\Omega), \tag{4.7}$$

where

$$a(u,v) = \int_\Omega (\alpha(x)\nabla u + \beta(x)u)\cdot\nabla v dx, \quad f(v) = \int_\Omega f(x)v dx. \tag{4.8}$$

This problem is uniquely solvable, and has the monotonicity property, i.e,

$$\text{If } (\mathcal{L}u)(x) \geq 0 \text{ for all } x \in \Omega \text{ then } u(x) \geq 0 \text{ for all } x \in \Omega. \tag{4.9}$$

In this paper, they introduce their scheme, which they prove to have the discrete equivalent to the monotonicity property, i.e, if one is using $P1$-Lagrange finite elements to approximate (4.6), and $\mathcal{L}_h$ is the corresponding discretization for $\mathcal{L}$, then

$$(\mathcal{L}_h^{-1} f_h)(x) \geq 0 \text{ for all } x \in \Omega, \text{ if } f_h^{(i)} = f(\varphi_i) \geq 0 \text{ for all } i = 1, \ldots, NV. \tag{4.10}$$

The core idea is to rewrite the standard $P1$-Lagrange bilinear form from a sum of values of the interpolant $u(v_i)$ with various weights, into a product of differences along edges in the mesh.

This can best be illustrated on the simplest case, that of Poisson's equation ($\alpha = 1, \beta = 0$).

Then, for u_h, $v_h \in V_h$, we have

$$\int_T \nabla u_h \cdot \nabla v_h dx = \sum_{i,j} a_{ij}^T u_h(x_i) v_h(x_j),$$

With the weights

$$a_{ij}^T = \int_T \nabla\varphi_j \cdot \nabla\varphi_i \,\mathrm{d}x.$$

As $a_{ii}^T = -\sum_{j\neq i} a_{ij}^T$, we can rearrange the sum to obtain the following simple, but important identity

$$\int_T \nabla u_h \cdot \nabla v_h dx = -\sum_{i<j} a_{ij}^T (u_h(x_i) - u_h(x_j))(v_h(x_i) - v_h(x_j)),\ u_h, v_h \in V_h. \tag{4.11}$$

Using the relation (4.11) and defining $\delta_E(f) = f(x_i) - f(x_j)$ for an edge $E = (x_i, x_j)$ with direction $\tau_E = \delta_E(x) = x_i - x_j$, we can rewrite the bilinear form in the following way:

$$\int_\Omega \nabla u_h \cdot \nabla v_h dx = \sum_{T\in\mathcal{T}_h} \sum_{E\subset T} \omega_E^T \delta_E u_h \delta_E v_h. \tag{4.12}$$

where $\omega_E^T = -a_{ij}^T$ with E connecting the vertices x_i and x_j.

For the weights ω_E^T the following simple identity holds

$$\omega_E^T = \frac{1}{d(d-1)} |\kappa_E^T| \cot \zeta_E^T, \tag{4.13}$$

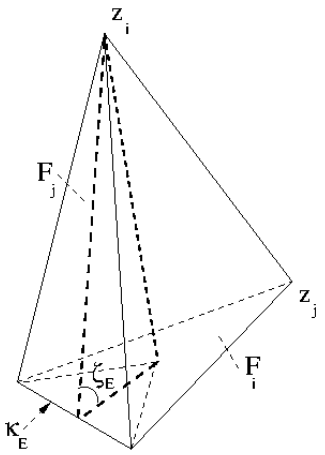

Figure 4.1. 3D element with vertices z_i, z_j.

where ζ_E^T is the the angle between the faces not containing edge E (see Fig. 4.1.1), and their intersection forms κ_E^T (the $n-2$ dimensional simplex opposite to the edge E), where for our purposes we take the volume of a vertex (0-dimensional simplex), 1.

Using this edge-based formulation of (4.12) gives us a relatively simple criterion for the triangulation, directly in terms of the weights in the bilinear form. Plugging in $u_h = \varphi_i$ and $v_h = \varphi_j$ into (4.12) and summing up the local contributions, it's a simple matter to check that

Lemma 4.1.1 ([19], lemma 2.1)**.** *The stiffness matrix for the Poisson equation is an M-matrix if and only if, for any fixed edge E the following inequality holds:*

$$\omega_E \equiv \frac{1}{n(n-1)} \sum_{T \supset E} |\kappa_E^T| \cot \zeta_E^T \geq 0, \tag{4.14}$$

where $\sum_{T \supset E}$ means summation that takes all simplexes T containing E.

This condition takes into account our discussion of earlier, namely that in two dimensions, a sufficient condition for discrete monotonicity requires the sum of the two angles opposite the same edge be less than or equal to π; indeed, Delaunay triangulations (with some possible exceptions at the boundary) satisfy this criterion, which imply that Delaunay triangulations can guarantee discrete monotonicity for the Poisson problem.

If one were to construct the same bilinear form for (4.8) and rewrite it as a sum of edge differences, one would get the following bilinear form:

$$\int_T J(u_h) \cdot \nabla v_h dx = \sum_E \omega_E^T \delta_E(J(u_h)) \delta_E v_h. \tag{4.15}$$

To approximate the difference in flux $J(u) \equiv \alpha \nabla u_h + \beta u_h$, the authors first introduce a function ψ_E defined locally on E whose tangential derivative along E is given by

$$\frac{\partial \psi_E}{\partial \tau_E} = \alpha^{-1}\beta \cdot \frac{\tau_E}{|\tau_E|}. \tag{4.16}$$

Using this auxiliary function, one can show the following, by solving the given edge-based ODE using integrating factors,

Lemma 4.1.2 ([19], lemma 3.1)**.** *Let $u \in H_0^1(\Omega) \cap C^0(\bar{\Omega})$. Then the following identity holds:*

$$\delta_E(e^{\psi_E} u) = \frac{1}{|\tau_E|} \int_E \alpha^{-1} e^{\psi_E} (J(u) \cdot \tau_E)\, ds, \tag{4.17}$$

where $J(u) = \alpha \nabla u + \beta u$.

If one approximates the flux $J(u)$ on the edge E with a constant J_E, then one can rewrite the edge difference $\delta_E J(u) = J_E \cdot \tau_E$ using (4.17) by replacing it with the edge difference of $e^{\psi_E} u$, by pulling out the constant $J_E \cdot \tau_E$ from the integral on the right and dividing both sides of the equation above by the other terms:

$$J_E \cdot \tau_E = \tilde{\alpha}_E(\beta) \delta_E(e^{\psi_E} u),$$

where $\tilde{\alpha}_E(\beta)$ is the harmonic average of $\alpha e^{-\psi_E}$,

$$\tilde{\alpha}_E(\beta) = \frac{|\tau_E|}{\int_E \alpha^{-1} e^{\psi_E} ds}. \tag{4.18}$$

We can then use this to simplify (4.15) further:

$$\sum_{E \subset T} \omega_E^T \delta_E(J(u_h)) \delta_E v_h \approx \sum_E \omega_E^T \delta_E (J_E \cdot \tau_E) \delta_E v_h = \sum_E \omega_E^T \delta_E \tilde{\alpha}_E(\beta) \delta_E(e^{\psi_E} u) \delta_E v_h. \tag{4.19}$$

Summing over all elements $T \in \mathcal{T}_h$, one gets an edge-based bilinear form with exponential

weighting,

$$a_h(u_h, v_h) = \sum_{T \in \mathcal{T}_h} \left\{ \sum_{E \subset T} \omega_E^T \tilde{\alpha}_E(\beta) \delta_E(e^{\psi_E} u_h) \delta_E v_h \right\}. \tag{4.20}$$

Given the continuity of $J(u) \cdot \tau_E$ across edges of a mesh (which is guaranteed by the interpolation using $P1$-Lagrange finite elements, as well as the continuity of the coefficients α and β), one can further simplify (4.20) by reordering the sum over edges in the mesh:

$$a_h(u_h, v_h) = \sum_{E \in \mathcal{T}_h} \omega_E \tilde{\alpha}_E(\beta) \delta_E(e^{\psi_E} u_h) \delta_E v_h, \tag{4.21}$$

with ω_E defined in (4.14).

Showing that the stiffness matrix generated by (4.21) is an M-matrix is straightforward:

Lemma 4.1.3 ([19], lemma 3.2)**.** *The stiffness matrix corresponding to the bilinear form (4.20) is an M-matrix for any continuous functions $\alpha > 0$ and β if and only if the stiffness matrix for Poisson equation is an M-matrix, namely if and only if the condition (4.14) holds.*

Proof. Given $j \in \{1, \ldots, NV\}$, consider the corresponding node x_j. Obviously, if x_i is a neighbor of x_j,

$$A_{ij} = \sum_{E \ni x_j} \omega_E \tilde{\alpha}_E(\beta) \delta_E(e^{\psi_E} \varphi_j) \delta_E \varphi_i = -\omega_E \tilde{\alpha}_E(\beta)(e^{\psi_{j,E}}) \leq 0. \tag{4.22}$$

Here $E \ni x_j$ means all the edges having x_j as an endpoint and $\psi_{j,E} = \psi_E(x_j)$.

Now, if x_j has no neighboring node on the boundary, then the $j-$th column sum of A is zero:

$$\sum_i A_{ij} = \sum_{E \ni x_j} \omega_E \tilde{\alpha}_E(\beta) \delta_E(e^{\psi_E} \varphi_j) \delta_E \sum_i \varphi_i = \sum_{E \ni x_j} \omega_E \tilde{\alpha}_E(\beta) \delta_E(e^{\psi_E} \varphi_j) \delta_E 1 = 0,$$

which means that $A_{jj} = \sum_{i \neq j} |A_{ij}|$. And if x_j has a neighboring node on the boundary, it is easy to see $\sum_i A_{ij} > 0$, or $A_{jj} > \sum_{i \neq j} |A_{ij}|$. This completes the proof. □

Remark 4.1.1. Note that the key to the M-matrix properties being satisfied lies in the replacement of the flux term in each sum by the exponential weighting term, e^{ψ_E}, which fixes the sign of the edge difference $\delta_E(e^{\psi_E} \varphi_j)$ to be positive. This term essentially enforces an automatic upwinding of the scheme that guarantees that the M-matrix property holds in the case that the edge weights ω_E^T are positive, and hence upholds the discrete monotonicity of the scheme provided the triangulation is Delaunay.

In order to simplify (4.20) for implementation, the authors are able to show a more computable form, where for β being constant,

$$A_{ij} = \sum_{E=(x_i,x_j)} \omega_E \tilde{\alpha}_E(0) \left[B(\frac{-\beta \cdot \tau_E}{\tilde{\alpha}_E(0)}) u(x_i) - B(\frac{\beta \cdot \tau_E}{\tilde{\alpha}_E(0)}) u(x_j) \right] = \sum_{T \ni x_i} \int_T f \varphi_i \, \mathrm{d}x, \tag{4.23}$$

with $\tau_E = x_i - x_j$, the summation is over all $x_j \neq x_i$, such that (x_i, x_j) is an edge, and B is the Bernoulli function,

$$B(s) = \begin{cases} \frac{s}{e^s - 1} & s \neq 0 \\ 1 & s = 0. \end{cases} \tag{4.24}$$

Xu and Zikatanov [19] also show that the discrete monotonicity holds, even for a more general problem with piecewise continuous β and α, and a mass term γu ($\gamma(x) \geq 0$) added to the bilinear form defined in (4.6). The corresponding bilinear form with this added term is written as

$$a_h(u_h, v_h) = \sum_{T \in \mathcal{T}_h} \left\{ \sum_{E \subset T} \omega_E^T \tilde{\alpha}_E(\beta) \delta_E(e^{\psi_E} u_h) \delta_E v_h + \gamma_T(u_h v_h) \right\}, \tag{4.25}$$

where mass lumping quadrature is used for the mass term,

$$\gamma_T(u_h v_h) = \frac{|T|}{n+1} \sum_{i=1}^{n+1} \gamma(x_i) u_h(x_i) v_h(x_i).$$

However, the discontinuity has to be aligned with the mesh, and the angles θ_E on the discontinuity line have to satisfy the stricter condition on the angles, $0 < \theta_E^T \leq \pi/2$, for all $T \supset E$. In this context, if one has a non-obtuse triangulation (where the property above regarding sums of angles opposite an edge holds for all edges in the mesh), then the monotonicity follows.

Finally, some mention has to be made on the way that β's behavior affects the scheme. In general, convection-dominated problems can be difficult to solve, which is a point that will be expanded upon in the last section of this chapter. In this paper, the scheme reflects this, in that the stability of the scheme depends directly on $||\beta||_\infty$. This complication is in some sense lessened by the asymptotic tendency of the Bernoulli function B to tend towards 0 as $\beta \to \infty$. However, in the case of negative $\beta \cdot \tau$, B approaches $-\infty$; in this sense, the sign of $\beta \cdot \tau_E$ can cause significant changes in B values, leading to very strong upwinding, which maintains the monotonicity of the scheme, but also can significantly impact the conditioning of the system.

Finally, it needs to be mentioned here that, as with most other monotone schemes, the convergence is only first order; i.e, they prove the following result for a more general problem including mass term γu and loosening the continuity constraints on α and β:

Theorem 6 ([19], Theorem 6.3)**.** *Let u be the solution of the problem (4.6), with nodal interpolant $u_I = \sum_{x_x \in \mathcal{T}_h} \hat{u}_i \varphi_i$. Assume that for all $T \in \mathcal{T}_h$ $\alpha \in W^{1,\infty}(T)$, $\beta \in [W^{1,\infty}(T)]^n$, $J(u) \equiv \alpha(x)\nabla u + \beta(x)u \in (W^{1,p}(T))^n$, and $\gamma(x) \in C(\bar{T})$ and $\gamma u \in W^{1,r}(T)$. Then the following estimate holds:*

$$||u_I - u_h||_{1,\Omega} \leq Ch \left\{ \sum_{T \in \mathcal{T}_h} |J(u)|^2_{1,p,T} + \sum_{T \in \mathcal{T}_h} |\gamma u|^2_{1,r,T} \right\}^{\frac{1}{2}} \tag{4.26}$$

for sufficiently small h.

One thing to note is that this "small enough h" condition is only required for the more general Delaunay triangulation, and the minimal mesh size in this context depends on an upper bound for the weak derivatives of α and β; in the case of non-obtuse triangulations however, the monotonicity of the scheme holds independent of mesh size, due to the property 4.14 holding, independent of β and α.

4.1.1.1 EAFE discretization of Richards' Equation

In the context of Richards' equation, this method can be seen as a generalized version of standard upwinding schemes such as those suggested in [27]; indeed, the authors of [19] show that the EAFE discretization limits to a variant of the upwinding scheme suggested by Forsyth in the case when $K \to 0$, or large variations in K (i.e, $|\beta| \to \infty$). To the authors' knowledge, EAFE or any variant of it has not been previously considered for finite element approximations of Richards' equation.

The edge-based bilinear form of any of the linearizations described before for linearization step κ is

$$a_h(u_h, v_h) = \tau \sum_{T \in \mathcal{T}_h} \left\{ \sum_{E \subset T} \omega_E^T \tilde{\alpha}_E(\tilde{K}^\kappa) \delta_E(e^{\psi_E} u_h) \delta_E v_h + J\theta_T^\kappa(u_h v_h) \right\}, \tag{4.27}$$

with the harmonic average of $Ke^{-\psi_E}$ on edge $E = (i, j)$ defined as

$$\tilde{\alpha}_E(\tilde{K}^\kappa) = \frac{|\tau_E|}{\int_E K^{-1,\kappa} e^{\psi_E} ds}, \tag{4.28}$$

and a ψ_E whose tangential derivative is defined using ∂K:

$$\frac{\partial \psi_E}{\partial \tau_E} = \tilde{K}^\kappa \cdot \frac{\tau_E}{|\tau_E|}, \tag{4.29}$$

with $\tilde{K}^\kappa \equiv [\partial K(\nabla \Psi + e_z)]^\kappa$.

In the case of symmetric linearizations, $\partial K = 0$; then the function ψ_E defined on the edge is a constant, so $\tilde{\alpha}_E$ simplifies to the harmonic average of of K over the edge multiplied by a constant:

$$\tilde{\alpha}(0) = \mathcal{H}_E(K^\kappa) = e^{-\psi_E} \frac{|\tau_E|}{\int_E K^{-1,\kappa}(s)\,\mathrm{d}s}. \tag{4.30}$$

Plugging in $\mathcal{H}_E(K^\kappa)$ into (4.27) cancels the exponential term, so that the edge-based bilinear form simplifies to

$$a_h(u_h, v_h) = \sum_{T \in \mathcal{T}_h} \left\{ \sum_{E \subset T} \tau \omega_E^T \mathcal{H}_E(K^\kappa) \delta_E u_h \delta_E v_h + J\theta_T^\kappa(u_h v_h) \right\}, \tag{4.31}$$

where we can now without loss of generality omit the exponential term in $\mathcal{H}_E(K^\kappa)$ so that it is just the harmonic average of K^κ over the edge. To compute the quadrature along the edge, note that, if one assumes $K = K_E$ constant on each edge E, then $\mathcal{H}(K^\kappa) = K_E^\kappa$. One reasonable choice of approximating this quadrature in the case of constant K_E would then be using the midpoint rule along the edge, $K\left(\frac{\theta(P_i)+\theta(P_j)}{2}\right)$, which exactly corresponds to the choice Zaidel and Russo suggest for the values at the edge [61], which C.T. Miller et al. [45] showed had good robustness in the tests they worked with, and is cheap to compute. As such, in our local stiffness matrix assembly, we evaluate

$$\tilde{\alpha}(0) = \mathcal{H}_E(K^\kappa) \approx K_E^\kappa = K\left(\frac{\theta(P_i^\kappa) + \theta(P_j^\kappa)}{2}\right),$$

i.e, the KAMS approach on each edge in our EAFE scheme.

In the case of $\partial K \neq 0$, one can estimate the term $\tilde{\alpha}_E(\tilde{K}^\kappa) e^{\psi_E(\tilde{K}^\kappa)}$ using the same technique to derive the analog of (4.23), which (fixing $\tilde{K}^\kappa$ constant on each edge) is

$$A_{ij} = \tau \sum_{E=(x_i,x_j)} \omega_E K_E^\kappa \left[B\left(-\left[K^{-1}\tilde{K}\right]_E^\kappa \cdot \tau_E \right) u(x_i) - B\left(\left[K^{-1}\tilde{K}\right]_E^\kappa \cdot \tau_E \right) u(x_j) \right]. \tag{4.32}$$

In scenarios when $\left[K^{-1}\tilde{K}\right]_E \cdot \tau_E >> K_E$,

$$K_E B\left(\left[K^{-1}\tilde{K}\right]_E \cdot \tau_E\right) \to 0,$$

and

$$K_E B\left(-\left[K^{-1}\tilde{K}\right]_E \cdot \tau_E\right) \to \tilde{K} \cdot \tau_E.$$

Let us consider the local contributions of an element T containing edge $E = (x_i, x_j)$ to (4.32) in this convection-dominated case. To compute these local contributions, one must compute the local stiffness matrix M_T. Plugging in $P_T = \sum_{x_i \in T} P_i \varphi_i$ as u_h and φ_j as v_h in (4.27), one can construct a local stiffness matrix M_T whose off-diagonal entries corresponding to edge $E = (x_i, x_j)$ are

$$M_T(i,j) = B\left(\left[K^{-1}\tilde{K}\right]_E^{\kappa} \cdot \tau_E\right) \int_T \nabla\varphi_j \cdot K_E \nabla\varphi_i \,\mathrm{d}x, \tag{4.33}$$

with tangential direction given as $\tau_E = x_i - x_j$.

To compute $\nabla\Psi^{\kappa}$ as required in $\tilde{K}_E^{\kappa}$, we must use P_T^{κ}; i.e, $\nabla\Psi_T^{\kappa} = \sum_{x_i \in T} P_i^{\kappa} \nabla\varphi_i$. Then, as we dot the gradient with τ_E, the relevant change in P^{κ} is in the direction of the edge, i.e,

$$\nabla\Psi_E^{\kappa} = \frac{P_i^{\kappa} - P_j^{\kappa}}{|x_i - x_j|^2}\tau_E.$$

We can thus give a more explicit form of (4.33):

$$M_T(i,j) = B\left(\left[K^{-1}\partial K\right]_E^{\kappa}\left(\frac{P_i^{\kappa} - P_j^{\kappa}}{|x_i - x_j|^2}\tau_E + e_z\right) \cdot \tau_E\right) \int_T \nabla\varphi_j \cdot K_E \nabla\varphi_i \,\mathrm{d}x. \tag{4.34}$$

As the evaluations of K^{-1} and ∂K are the same for the same edge, the mirror term from x_i to x_j is

$$M_T(j,i) = B\left(-\left[K^{-1}\partial K\right]_E^{\kappa}\left(\frac{P_i^{\kappa} - P_j^{\kappa}}{|x_i - x_j|^2}\tau_E + e_z\right) \cdot \tau_E\right) \int_T \nabla\varphi_i \cdot K_E \nabla\varphi_j \,\mathrm{d}x. \tag{4.35}$$

As K and ∂K are always nonnegative, then the Bernoulli function coefficients weight the contributions to the global stiffness matrix A with smaller weights in the direction of increasing P_T; e.g, if $P_i > P_j$, then the Bernoulli weight in (4.35) which contributes to the global matrix entry A_{ji} is larger in magnitude (and by construction, same sign) than that of (4.34) for the local contribution to global matrix entry A_{ij}. In the limit of small K or large ∂K, these

weights smoothly go to 0 for the A_{ij} contribution, or to $\partial K_E^\kappa \nabla(\Psi_E^\kappa + z) \int_T \nabla\varphi_i \cdot \nabla\varphi_j \,\mathrm{d}x$ for the A_{ji} contribution. This upwinding scheme is similar to the manual upwinding process (4.3) as suggested in [27], but with distinct advantage of being automatically given when assembling locally.

4.1.2 Mixed finite element method

There are several applications where local conservation of fluid flux is a highly desirable or necessary trait for numerical simulations to have. In the context of Richards' equation, one of these applications involves modeling the transport of pollutants by groundwater flows; in this scenario, Richards' equation would be used to model the infiltration of water into the porous material, and the result would feed into the contaminant transport model, predicting the rate of propagation and distribution of the contaminant in the soil; such models are relevant for many environmental simulations that involve potentially hazardous chemicals infiltrating local water supplies, such as fracking surfactants, nuclear runoff, and tracking the dissemination of pesticides used in agricultural processes. In this context, the use of the standard $P1$-Lagrange elements may lead to inherently inaccurate contaminant fate predictions.

To see this, consider the elliptic term in (4.1), after discretizing with $P1$-Lagrange finite elements. The argument of this term corresponds to the discrete Darcy fluid flux, which we label as

$$q_h = K\nabla(\Psi_h + z). \tag{4.36}$$

Note that with the $P1$-Lagrange basis, q_h is piecewise constant on each element T, with potential jumps across element faces (indeed, the shape functions always have discontinuities in the normal component of their gradients). As such, $\operatorname{div} q_h$ can only be defined weakly for the entire domain:

$$\int_\Omega \varphi_i \operatorname{div} q_h \,\mathrm{d}x = \sum_{T\in\mathcal{T}_h} \int_{\partial T} \varphi_i (q_T \cdot n) \,\mathrm{d}x - \int_T q_T \cdot \nabla\varphi_i \,\mathrm{d}x. \tag{4.37}$$

Expanding the first sum of flux integrals over the boundary of each face $f \in T$

$$\begin{aligned} \sum_{T\in\mathcal{T}_h} \int_{\partial T} \varphi_i (q_T \cdot n) \,\mathrm{d}x &= \sum_{T\in\mathcal{T}_h} \sum_{f\in T} \int_f \varphi_i (q_T \cdot n) \,\mathrm{d}x \\ &= \sum_{f\in\mathcal{T}_h} \int_f [[q_h \cdot n]]_f \varphi_i \,\mathrm{d}x, \end{aligned} \tag{4.38}$$

where $[[q_h \cdot n]]_f$ is the difference of the values of $q_h \cdot n$ from the two elements sharing face f. As q_Ts are sums of the gradients of the shape functions φ_i, and the normal components of these gradients are discontinuous across faces, so too are the q_Ts, meaning that the weak derivative as defined above cannot be square-integrable, which implies that div cannot be defined weakly.

This implies that solutions to the weak formulation (4.2) do not have to conserve mass locally, as this lack of continuity of the normal component of q_h across faces implies that for certain functions $\Psi_h \in V_h$, the sum of the fluxes across a face need not be equal to 0. In fact, since the tangential components of the gradients of the shape functions are continuous, it turns out that the only function in V_h for which $[[q_h \cdot n]]_f = 0$ everywhere locally and globally is the constant function $\Psi_h = C$.

One way to overcome this problem is to properly exploit the local subdomains where the mass conservation property of the finite element scheme is inherently satisfied. For example, on two-dimensional Delaunay triangulations these subdomains can be defined as the Voronoi (or Thiessen) polygons [64]. A local postprocessing procedure can then be used to yield accurate and mass conserving velocity fields. However, this approach has the limitation of not being directly applicable to three dimensions. Conversely, Integrated Finite Difference (IFD) or Finite Volume (FV) schemes are inherently mass conservative, as the subdomain of definition of the mass conservation principle coincides with the mesh element. However, the applicability of IFD or FV to general domains and boundary conditions is often cumbersome, and analysis of stability and error estimates are not straightforward to derive.

To this end, we seek to use a finite element formulation that can guarantee local mass conservation. There are various formulations that allow for this; for instance, one method to work with the weak divergence definition (4.37) consistent is to track the jump in fluxes (4.38) and enforce their continuity as a Lagrange multiplier constraint; this results in the discontinuous Galerkin (DG) formulation, which has been used in modeling contaminant transport problems [65].

However, as the Darcy velocity itself is a variable of interest in our context, we elect to use mixed finite elements to discretize (4.1). This finite element approximation swaps the minimization problem (4.2) posed in the previous section with a saddle point problem, find $(\Psi, q) \in S \times Q$ such that, for each $(v, r) \in S \times Q$,

$$\begin{aligned} \tau(q^n, r) - \tau(K(\Psi^n)\nabla(\Psi^n + z), r) &= \tau(g, r)_{\partial\Omega}, \\ -(\theta^n, v) + \tau(\operatorname{div} q^n, v) &= -\tau(f, v) - (\theta^{n-1}, v). \end{aligned} \tag{4.39}$$

The spaces S and Q need to be defined appropriately so that the resulting saddle point problem is well-posed.

Omitting the timestep index, introducing the linearization index, and computing the residuals for each equation as was done in the previous section, we get the following linearizations for (4.39):

(Mixed Picard)

$$\begin{aligned} &\tau(K^{-1,j}q_\epsilon, r) - \tau(\nabla\Psi_\epsilon, r) = \tau(G^j, r), \\ &\tau(\operatorname{div} q_\epsilon, v) - (\theta^j\Psi_\epsilon, v) = (F^j, v), \end{aligned} \tag{4.40}$$

(Full Newton)

$$\begin{aligned} &(K^{-1,j}q_\epsilon, r) - \tau(\nabla\Psi_\epsilon, r) - \tau\left(\left[K^{-1}\partial K K^{-1}\right]^j q^j\Psi_\epsilon, r\right) = \tau(G^j, r) \\ &\tau(\operatorname{div} q_\epsilon, v) - (\partial\theta^j\Psi_\epsilon, v) = (F^j, v), \end{aligned} \tag{4.41}$$

(Modified Picard)

$$\begin{aligned} &\tau(K^{-1,j}q_\epsilon, r) - \tau(\nabla\Psi_\epsilon, r) = \tau(G^j, r) \\ &\tau(\operatorname{div} q_\epsilon, v) - (\partial\theta^j\Psi_\epsilon, v) = (F^j, v), \end{aligned} \tag{4.42}$$

(L-scheme)

$$\begin{aligned} &\tau(K^{-1,j}q_\epsilon, r) - \tau(\nabla\Psi_\epsilon, r) = \tau(G^j, r) \\ &\tau(\operatorname{div} q_\epsilon, v) - (L\Psi_\epsilon, v) = (F^j, v). \end{aligned} \tag{4.43}$$

Denoting the linear and bilinear forms with $v \in S$ and $r \in Q$,

$$\begin{aligned} (J^j\theta\Psi_\epsilon, v) &\to d(\Psi_\epsilon, v), \\ \tau(K^{-1,j}q_\epsilon, r) &\to a(q_\epsilon, r), \\ \tau(\operatorname{div} q_\epsilon, v) &\to b_{\operatorname{div}}(q^j, v), \\ \tau\left(\left[K^{-1}\partial K K^{-1}\right]^j q^j\Psi_\epsilon, r\right) &\to b_{K'}(\Psi_\epsilon, r), \\ (F^j, v) &\to F^j(v), \\ \tau(G^j, r) &\to G^j(r), \end{aligned}$$

where $J^j\theta = \partial\theta^j, \theta^j$ (depending on linearization). We can reformulate each of the linearizations of (4.39) as a symmetric saddle point problem for $(q_\epsilon, \Psi_\epsilon) \in Q \times S$,

$$\begin{aligned} a(q_\epsilon, r) + b^*_{\text{div}}(\Psi_\epsilon, r) &= G^j(r), \quad r \in Q, \\ b_{\text{div}}(q_\epsilon, v) - d(\Psi_\epsilon, v) &= F^j(v), \quad v \in S, \end{aligned} \tag{4.44}$$

or in the case of the Newton linearization, the non-symmetric saddle point problem

$$\begin{aligned} a(q_\epsilon, r) + b^*_{\text{div}}(\Psi_\epsilon, r) - b_{K'}(\Psi_\epsilon, r) &= G^j(r), \quad r \in Q, \\ b_{\text{div}}(q_\epsilon, v) - d(\Psi_\epsilon, v) &= F^j(v), \quad v \in S, \end{aligned} \tag{4.45}$$

with b^*_{div} denoting the adjoint operator of b_{div}. To motivate the definition of S and Q we use, re-consider the weak div definition, (4.37). In order for the integrals involved to be well-defined, one sufficient set of conditions are that the divergence of flux q_h be square integrable, and the test function φ to be square integrable; in other words, it will be sufficient to have

$$q_h \in H_{\text{div}}(\Omega) = \{v_h \, : v_h \in L^2(\Omega)^d, \text{div}(v_h) \in L^2(\Omega)\}$$

and

$$\varphi \in L^2(\Omega).$$

Further, to ensure the local consistency of the right hand side of (4.37), it suffices to choose a subspace $Q_h \subset H_{\text{div}}(\Omega)$ whose basis functions $\{\phi_f\}_{f \in \mathcal{T}_h}$ have continuous normal component across each of the faces in the triangulation, and subspace $S_h \subset L^2(\Omega)$ whose basis functions $\{\chi_T\}_{T \in \mathcal{T}_h}$ are piecewise constant on each of the elements.

Choosing then the degrees of freedom for the discrete flux q_h as the value of the flux across a face $f \in \mathcal{T}_h$,

$$\sigma_{f_i}(q_h) = \int_f q_h \cdot n_f \, \mathrm{d}x,$$

and the degrees of freedom for the discrete pressure Ψ_h as the value at the barycenter, x^*_T:

$$\sigma_{T_i}(\Psi_h) = \Psi_h(x^*_{T_i}),$$

the lowest order induced shape functions for the discrete flux are the zeroth order Raviart-

Thomas finite elements, whose definition restricted to a simplex $T \in \mathcal{T}_h$ is

$$\phi_{f,T} = s_{f,T} \frac{x - x_f}{d|T|}. \tag{4.46}$$

Here x_f is the sole vertex in T that is not in the face f, and the constant $s_{f,T}$ serves the purpose of enforcing the continuity of the normal component of the flux across the face f. It is defined as follows:

$$s_{f,T} = \begin{cases} (n_{f,T} \cdot n_f) = \pm 1, & f \in \partial T \\ 0, & f \notin \partial T. \end{cases}$$

In what follows, for $f \in \partial T$, $n_{f,T}$ is the normal vector to f pointing outward with respect to T, and n_f is the globally oriented normal vector to f whose direction is chosen independently of T and is fixed for every f. In practice, the typical way to choose this globally oriented normal vector involves numbering all d-dimensional simplices as $1, \ldots, |\mathcal{T}_h| = n_T$ and for $T \in \mathcal{T}_h$; $i(T)$ is the number of T in this sequence. Then,

$$n_f = \text{sign}(i(T) - I(T'))n_{f,T}, \quad \text{iff} \quad f = T \cap T'.$$

To show the unisolvence of this finite element for the discrete fluxes, it suffices to show:

Lemma 4.1.4. *For two faces $f, f' \in \mathcal{T}_h$,*

$$\int_{f'} \phi_f \cdot n_{f'} = \begin{cases} 1, & f = f', \\ 0, & \textit{otherwise.} \end{cases} \tag{4.47}$$

Proof. We first note that for $f' \neq f$ and $x \in f'$ we have $(x - x_f) \cdot n_{f'} = 0$ because $x_f \in f'$. We then only need to to show the identity (4.47) for $f = f'$. We recall the following identities

$$n_f = -\frac{1}{|\nabla \lambda_f|} \nabla \lambda_f, \quad |\nabla \lambda_f| = \frac{|F|}{d|T|},$$

where λ_f is the barycentric coordinate function satisfying

$$\lambda_f(x_{f'}) = \begin{cases} 1, & f = f', \\ 0, & \text{otherwise.} \end{cases}$$

Since λ_f is linear, the first order Taylor expansion of λ_f centered at x_f yields for $x \in f$,

$$0 = \lambda_f(x) = \lambda_f(x_f) + \nabla\lambda_f \cdot (x - x_f) = 1 + \nabla\lambda_f \cdot (x - x_f).$$

This shows that on f the quantity $(\phi_f \cdot n_f)$ is a constant. Therefore, for $x \in f$ we have

$$\begin{aligned} \phi_f \cdot n_f &= -\frac{1}{d|T|\,|\nabla\lambda_f|}(x - x_f) \cdot \nabla\lambda_f \\ &= \frac{1}{|F|}, \end{aligned}$$

which clearly implies (4.47). □

For the piecewise constant pressure, the induced shape functions are the characteristic functions

$$\chi_{T_i}(x) = \begin{cases} 1, & x \in T_i, \\ 0 & \text{otherwise.} \end{cases} \tag{4.48}$$

The unisolvence of these finite elements with their dual basis follow from the defintion, since the map

$$p \ni S_h \to \Big(p(x^*_{T_1}), p(x^*_{T_2}), \ldots, p(x^*_{T_{n_T}})\Big) \in \mathbb{R}^{n_T}$$

is clearly bijective.

Now we can define the finite dimensional formulation of the linearizations of (4.39). On the domain $\mathcal{T}_h$, we seek the solution $(\Psi^n_h, q^n_h) \in S_h \times Q_h$ such that for every $(v_h, r_h) \in S_h \times Q_h$,

$$\begin{aligned} a_h(q_{h,\epsilon}, r_h) + b^*_{h,\mathrm{div}}(\Psi_{h,\epsilon}, r_h) &= G^j(r_h), \quad r_h \in Q_h, \\ b_{h,\mathrm{div}}(q_{h,\epsilon}, v_h) - d_h(\Psi_{h,\epsilon}, v_h) &= F^j(v_h), \quad v_h \in S_h, \end{aligned}$$

Or in the case of the Newton linearization,

$$\begin{aligned} a_h(q_{h,\epsilon}, r_h) + b^*_{h,\mathrm{div}}(\Psi_\epsilon, r_h) - b_{h,K'}(\Psi_{h,\epsilon}, r_h) &= G^j(r_h), \quad r_h \in Q_h, \\ b_{h,\mathrm{div}}(q_{h,\epsilon}, v_h) - d_h(\Psi_{h,\epsilon}, v_h) &= F^j(v_h), \quad v_h \in S_h, \end{aligned}$$

where the functionals defined in the above are the finite dimensional equivalents of the continuous linear forms defined before.

One can guarantee the well-posedness of finite element discretizations of a continuous variational problem by using stability estimates, which bound the energy of the solution to

the variational problem on finite dimensional subspaces by that of the source and boundary data, with a constant independent of the size of the subspaces. This implies that for zero source data, the only solution can be 0, and one can use these to form error estimates that can show the rate (in h and τ) at which the discrete solutions approach the continuous one.

The stability of mixed finite element discretizations of time discretized parabolic problems of the form

$$\begin{aligned} &Lu^n - \tau \operatorname{div}(K(x)\nabla u^n + \beta(x)u^n) = \tau f^n + Lu^{n-1}, && x \in \Omega, \\ &u(x,0) = u_0(x), && \Omega \times \{t = 0\}, \\ &u^n|_{\Gamma_D} = u_D(x,t), \\ &\partial_\nu u^n|_{\Gamma_N} = g(x,t) \end{aligned}$$

is well established for uniformly elliptic K and $L > 0$ (see, for instance, [66], chapter 7 for a detailed discussion).

As each linear problem of form (4.44), (4.45) can be written as a problem of this type, the well-posedness on each linear iteration for the mixed finite element discretization follows. The error estimates are of the form

$$||q^n - q_h^n||_{L^2(\Omega)} \leq C(h+\tau)||q^n||_{L^2(\Omega)},$$

and for the pressure,

$$||\Psi^n - \Psi_h^n||_{L^2(\Omega)} \leq C(h+\tau)||\Psi^n||_{L^2(\Omega)},$$

where (q, Ψ) are the continuous solutions of(4.39), with appropriate boundary conditions.

Establishing the stability and error estimates for the nonlinear problem is not straightforward; essentially, the well-posedness of the problems is dependent on the Jacobian of the system, which can become singular for mixed saturated-unsaturated problems. One common technique to circumvent this issue is to use the Kirchoff transformation as defined in chapter 1. However, one problem that persists is the low regularity of the solutions; particularly that in time, [9] manage to show that (for the full Dirichlet problem, for simplicity) u and $b(u)$ are in the spaces

$$\begin{aligned} b(u) &\in L^\infty(0,T;L^1(\Omega)), \\ \partial_t b(u) &\in L^2(0,T;H^{-1}(\Omega)), \\ u &\in L^2(0,T;H_0^1(\Omega)), \end{aligned} \tag{4.49}$$

$$K(b(u))e_z \in L^2(\Omega \times (0,T)).$$

This implies that the Kirchoff flux,

$$q_b = -(\nabla u + k(b(u))e_z) \in L^2(0,T;(L^2(\Omega))^d).$$

However, due to the degeneracy of b near full saturation, the weak problem requires test functions that are in $H^1(\Omega)$, which are more restrictive than the test functions used for the weak formulation. However, after redefinition on a set of measure zero, $b(u) \in C(0,T;H^{-1}(\Omega))$; thus one can integrate in time as is done in [5], to get the continuous in time relation

$$b(u(t)) + \operatorname{div} \int_0^t q_b(s)\,\mathrm{d}s = b(u_0), \quad t \in (0,T]$$

defined in the H^{-1} sense. If one further assumes that b is Hölder continuous of exponential order $0 < \alpha < 1$, then the regularity of u implies that

$$b(u) \in L^2(0,T;L^{2/\alpha}(\Omega)) \subset L^2(0,T;L^2(\Omega)).$$

Thus, if we assume that initial data $u_0 \in L^2(\Omega)$, the Hölder continuity of b gives us that

$$\int_0^t q\,\mathrm{d}s \in H^1(0,T;(L^2(\Omega))^d) \cap L^2(0,T;H_{\operatorname{div}}(\Omega)),$$

giving us that $\operatorname{div} \int_0^t q\,\mathrm{d}s \in L^2(\Omega)$ for a.e t. Then, the authors of [20], one consider the time integrated form of (4.39), and then after discretizing in time using Backward Euler, prove stability in the form

$$\begin{aligned}
&\tau \sum_{n=1}^N ||\Psi^n||_1^2 + \tau \sum_{n=1}^N ||q^n||^2 \le C,\\
&\sum_{n=1}^N (b(p^n) - b(p^{n-1}), p^n - p^{n-1}) + \tau \max_{n=1,\dots,N} ||q^n||^2 + \tau \sum_{n=1}^N ||q^n - q^{n-1}||^2 \le C\tau, \qquad (4.50)\\
&\tau \sum_{n=1}^N ||\operatorname{div} q^n||^2 \le C\tau^{-\frac{2(1-\alpha)}{\alpha}},
\end{aligned}$$

and error estimates between the semidiscrete and fully discrete solutions of the form

$$\sum_{n=1}^N \int_{t_{n-1}}^{t_n} ||b(\Psi(t)) - b(\Psi_h^n)||_{L^{1+\frac{1}{\alpha}}(\Omega)}^{1+\frac{1}{\alpha}}\,\mathrm{d}t + \left|\left|\sum_{n=1}^N \int_{t_{n-1}}^{t_n} (\Psi(t) - \Psi_h(t))\,\mathrm{d}t\right|\right|^2 \qquad (4.51)$$

$$+ \left|\left|\sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} (q(t) - q_h^n)\, \mathrm{d}t\right|\right|^2 \leq C\left(\tau + h^2 \tau^{-\frac{2(1-\alpha)}{1+\alpha}}\right).$$

These estimates, when combined with the proof in [16] that solutions to the time integrated variational problem are equivalent to solutions of the variational problem without time integration, provide stability for both the $P1$-Lagrange formulation and the mixed RT0-P0 formulation, for small enough h, dependent on the Hölder exponent α of b.

4.2 Krylov subspace iterative solvers

Now that we have linearized and discretized our problem, we must discuss the particulars of constructing and solving the linear systems that are generated.

As a precursor, one issue of particular relevance is the size N of the linear systems, which increase exponentially as $h \to 0$. For instance, the system (4.2) discretized with $P1$-Lagrange elements in d spatial dimensions has the number of DOF (unknowns) equal to the number of vertices in the mesh, which scale as $\mathcal{O}(h^{-d})$. Even the best direct solvers scale unfavorably with the size of the system, which can become very large (for instance, on a unit cube split into equal size subcubes of size $h = 1/128$, the number of unknowns to solve exceeds 2 million for the three dimensional problem). In these scenarios, even $\mathcal{O}(N^2 \log(N))$ operations to solve the given linear system could be prohibitively expensive.

To this end, solving linear systems with iterative methods that take an initial guess and repeat some low cost approximate solve until some tolerance is reached is a much more scalable alternative, as the simplest methods involve $\mathcal{O}(N)$ calculations, implying much faster linear solve times. The caveat of working with iterative methods is that the amount of iterations required for an iterative solver to converge within some given tolerance might be rather large (so large in fact that in some cases, the use of a direct solver would have solved the problem faster). One indicator that can predict how effectively an iterative solver can converge to a solution is what is known as the *condition number* of the system. For an invertible linear operator A, the condition number of A can be defined as

$$\kappa(A) = ||A^{-1}||||A||,$$

where for A is a linear transformation from a finite dimensional Hilbert space V to itself with

norm $||\cdot||_V$, which is the standard operator norm $||A||$ can be defined as

$$||A|| \equiv \sup_{v \in V} \frac{||Av||_V}{||v||_V}.$$

In the case $V = \mathbb{R}^n$, if we consider the matrix representation A, the operator norm becomes

$$||A|| = \sup_{v \in \mathbb{R}^n} \left(\frac{\langle Av, Av \rangle}{\langle v, v \rangle} \right)^{1/2},$$

with $\langle x, x \rangle$ being the standard Euclidean inner product. Then

$$\begin{aligned} \kappa(A) &= \sup_{v = Aw \in \mathbb{R}^n} \left(\frac{\langle w, w \rangle}{\langle Aw, Aw \rangle} \right)^{1/2} \sup_{v \in \mathbb{R}^n} \left(\frac{\langle Av, Av \rangle}{\langle v, v \rangle} \right)^{1/2} \\ &= \frac{\sigma_{\max}}{\sigma_{\min}} \geq 1, \end{aligned}$$

with $\sigma_{\max\,/\,\min}$ denoting the maximum/minimum singular value of A.

When solving $Ax = b$ with an iterative method, the conditioning of the system is very important. All iterative solvers can be shown to form contractive sequences of iterates $\{x^i\} \to x$, where the convergence has rate

$$||x - x_i|| \leq \left[1 - f\left(\kappa(A)\right)\right]^i ||x - x_0||, \tag{4.52}$$

where $0 < f(s) < 1$ is a decreasing function, so that the number of iterations required to converge increases exponentially as κ increases. For instance, for the Gauss-Seidel iteration, $f(s) = s^{-1}$.

Discretizing either of the linearizations (4.31), (4.32) with $P1$-Lagrange finite elements over a given triangulation $\mathcal{T}_h$ gives a linear system of the form

$$(D_\theta + \tau A_E) P_\epsilon = r^\kappa, \tag{4.53}$$

to solve on each linearization step κ, with the vector $P_\epsilon = [\Psi_\epsilon^1, \ldots, \Psi_\epsilon^{NV}]^T$ representing the DOF of the $P1$-Lagrange interpolant of the correction,

$$\sum_{T \in \mathcal{T}_h} \sum_{E \subset T} \omega_E^T J \theta_T^\kappa (\varphi_j \varphi_i) \to D_\theta^{ij},$$

$$\sum_{T\in\mathcal{T}_h}\sum_{E\subset T}\omega_E^T\tilde{\alpha}_E(\tilde{K}^\kappa)\delta_E(e^{\psi_E}\varphi_j)\delta_E\varphi_i \to A_E^{ij},$$

with $\tilde{\alpha}_E$, ψ_E, $\tilde{K}^\kappa$ as defined in section 4.1.1.1, and the weights ω_E^T as defined in (4.13).

Provided $J\theta^\kappa(x_j) > 0$ for all j, the matrix D_θ is conditioned well, as the condition number is just the ratio between the highest and lowest $J\theta^\kappa$ values. Note that for the L-scheme, even in saturated-unsaturated problems, $J\theta \equiv L$, and so this matrix is perfectly conditioned.

In the cases when $\tilde{K}^\kappa = 0$, the stiffness matrix A_E can be simplified to (4.31), which we reproduce here:

$$\sum_{T\in\mathcal{T}_h}\left\{\sum_{E\subset T}\omega_E^T K_E\delta_E(u_h)\delta_E v_h + J\theta_T^\kappa(u_h v_h)\right\}.$$

In the symmetric case, the conditioning of this system is of the same order in h as the conditioning of the regular Laplace discretization, which is known to scale like $\mathcal{O}(\tau h^{-2})$, with a constant related to the ratio of the max and min values of K_E. As such, the contraction coefficient (4.52) goes to 1 as $h \to 0$, implying that iterative techniques would require many iterations to converge.

In the case of mixed finite elements, one gets the following saddle point system to solve, after discretizing using RT0 elements for the Darcy flux, $q = \sum_{f\in\mathcal{T}_h} q_f^{RT}\phi_f$, and $P0$ elements for the pressure head, $P^{RT} = \sum_{T_i\in\mathcal{T}_h} P_T^{RT}\chi_{T_i}$:

$$\begin{bmatrix} A_{qq} & B_{\text{div}}^T - B_{K'} \\ B_{\text{div}} & -D_\theta \end{bmatrix}\begin{pmatrix} q_\epsilon^{RT} \\ P_\epsilon^{RT} \end{pmatrix} = \begin{pmatrix} \tilde{f} \\ \tilde{g} \end{pmatrix}, \tag{4.54}$$

with

$$\tau\sum_{T\ni f'}\sum_{f,f'\in T}\int_T \phi_f K_T^{-1,\kappa}\cdot\phi_{f'}\,\mathrm{d}x \to A_{qq}^{f',f},$$

$$\tau\sum_{f\in T_i}\int_f \phi_f\cdot n_f\,\mathrm{d}x \to B_{\text{div}}^{T_i,f},$$

$$\tau\int_{T_i}\left[K^{-1}\partial K K^{-1}\right]^\kappa q^\kappa\cdot\phi_f\,\mathrm{d}x \to B_{K'}^{f,T_i},$$

$$\int_{T_i} J^\kappa\theta\,\mathrm{d}x \to D_\theta^{T_i,T_i}.$$

Here the total number of degrees of freedom is equal to the number of unique faces $f \in \mathcal{T}_h \approx \frac{d}{2}d!h^{-d}$, plus the number of elements $T \in \mathcal{T}_h \approx d!h^{-d}$. As this linear system is spectrally equivalent to (4.53), the conditioning of this system scales with the same order in h.

For the nonsymmetric linearization, the first order convection changes the scaling of the condition number to a first order $\kappa \sim \mathcal{O}(\tau h^{-1})$. This, coupled with the fact that nonsymmetric problems are harder to precondition than nonsymmetric ones, implies that for problems where the gradients of K become large, the number of iterations will increase immensely, particularly for problems where the convection-dominated areas can change per linearization step.

In this sense, solving (4.53) and (4.54) as $h \to 0$ introduces challenges: every time one halves the characteristic mesh size, the problem size N increases exponentially, and the condition number of the system grows by a factor of roughly 4 for the symmetric case, and by 2 for the nonsymmetric case. In this sense, scalable solvers coupled with preconditioners are necessary to efficiently solve the system. We will discuss preconditioners in chapter 5; for now, let us first define the iterative solvers that we use, and mention some of the properties of these solvers.

Solving a linear system $Ax = b$ can be seen as finding critical points of the quadratic functional,

$$\varphi(x) = \frac{1}{2}(Ax, x) - (b, x).$$

Projection methods do this by extracting an approximate solution x_m from an affine subspace $x_0 + \mathcal{K}_m$ of dimension m by imposing the Petrov-Galerkin condition,

$$b - Ax_m \perp \mathcal{L}_m,$$

where $\mathcal{L}_m$ is another subspace of dimension m and x_0 is an arbitrary initial guess to the solution. A Krylov subspace method is a method for which the subspace $\mathcal{K}_m$ is the Krylov subspace

$$\mathcal{K}_m(A, r_0) = \text{span}\{r_0, Ar_0, A^2 r_0, \ldots, A^{m-1} r_0\},$$

where $r_0 = b - Ax_0$. The different versions of Krylov subspace methods arise from different choices of the subspace $\mathcal{L}_m$ and from the ways in which the system is preconditioned, a topic that will be covered in the next chapter. Two broad choices for $\mathcal{L}_m$ give rise to the best-known techniques. The first is simply $\mathcal{L}_m = \mathcal{K}_m$ and the minimum-residual variation $\mathcal{L}_m = A\mathcal{K}_m$, which minimize the residual $||r_i||_{A^{-1}} = ||b - Ax_i||_{A^{-1}}$ and $||r_i||_2$, respectively, for an SPD A. In the case that A is not symmetric, one can define $\mathcal{L}_m$ to be a Krylov subspace method associated with A^T, namely, $\mathcal{L}_m = \mathcal{K}_m(A^T, r_0)$. This leads to methods like BiConjugate Gradient.

4.2.1 Conjugate Gradient method

In the case that $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite (SPD), the quadratic linear form defined above has a unique minimum. One of the most powerful and popular iterative solvers can be seen as a method of finding this minimum,

$$x_* = \arg\min_{x\in\mathbb{R}^n} \varphi(x), \quad \varphi(x) = \frac{1}{2}(Ax, x) - (b, x), \tag{4.55}$$

In this scenario, the linear form is convex, and it is clear that the unique minimizer is the solution to the linear problem $Ax = b$.

A standard method of finding the minimizer x_* is to use some line search method, all of which can be written in the form

1. Find $\alpha_j = \arg\min \varphi(x_j + \alpha p_j)$,
2. Set $x_{j+1} = x_j + \alpha_j p_j$.

The key to the successs of a line search method is in choosing a good set of search directions $\{p_j\}$. The CG method creates a sequence of search directions p_j that are "conjugate" to each other, i.e, orthogonal with respect to the inner product $||x||_A = (Ax, x)$ by using a Gram-Schmidt process known as the Arnoldi iteration; this A-orthogonality ensures that each succesive residual $r_k = b - Ax_k$ is minimized in the subspace spanned by all prior search directions, and also orthogonal to every other residual before it, which can be used to further simplify the algorithm used to generate the new search path and iterative solution.

The algorithm in simplest form is as follows, as described in [67]:

Algorithm 4.2.1 (Conjugate Gradient)**.** Result: Approximate solution $\widetilde{x}$ to $Ax = b$. Let x_0 be given initial guess. Compute initial residual $r_0 := b - Ax_0$, $p_0 := r_0$. For $i = 0, \ldots$ until convergence,

1. Compute correction coefficient $\alpha_i := (r_i, r_i)/(Ap_j, p_j)$
2. Compute corrected guess, $x_{j+1} := x_j + \alpha_j p_j$.
3. A-orthogonalize the residual, $r_{j+1} := r_j - \alpha_j A p_j$.
4. A-orthogonalize the conjugate direction,

$$p_{j+1} := r_{j+1} + \beta_j p_j,$$

where $\beta_j := (r_{j+1}, r_{j+1})/(r_j, r_j)$.

As can be seen, this method is optimal in the sense that no matrix inverses are ever computed, meaning that the running cost of this algorithm per iteration is on the order of a few matrix vector multiplications, each of which are $\mathcal{O}(N)$ for the sparse systems (4.53) and (4.54).

Note that, because A is positive definite, the iterative procedure will converge in at most n iterations; however, the hope is that convergence will be reached for $k << n$. The number of iterations to converge, however, can vary significantly depending on the conditioning of the system. It can be shown that the convergence rate of this algorithm is dependent on the condition number of A in the following sense:

$$\|x_* - x_l\|_A \leq \frac{2}{\left(\frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}\right)^l + \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^l} \|x_* - x_0\|_A \leq 2\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^l \|x_* - x_0\|_A. \tag{4.56}$$

One way to think of this is, the higher the condition number, the higher the spread of the eigenvalues, and hence the worse this scheme does, which indicates that the performance of PCG will suffer significantly for the problems we consider. Further, this method can only work for symmetric systems, as for nonsymmetric problems, the norm $||\cdot||_A$ cannot be defined. There are ways around this, i.e, by generating orthogonal bases in the Krylov subspaces

$$\mathcal{K}_m = \text{span}\{r_0, A^T r_0, \ldots, (A^T)^{m-1} r_0\},$$

one can define the Bi-Conjugate Gradient method. Combined with a stabilization procedure to prevent failure of the algorithm in certain edge cases, this becomes the well-known BiCG-STAB method.

4.2.2 Generalized Min Res method

Similar to CG, the Generalized Minimal Residual Method also uses an orthogonal basis of the Krylov subspaces $\mathcal{K}_m$ via the Arnoldi iteration to minimize some norm of the error $||r_m||$. However, in this case, the norm is $||\cdot||_{\ell^2}$, rather than the norm induced by A. In particular, on iteration m, if the resulting orthogonal basis of $\mathcal{K}_j$ is written as

$$Q_m = [p_1, p_2, \ldots, p_m],$$

then the matrix A can be shown to satisfy

$$AQ_m = Q_{m+1}\bar{H}_m,$$

where $\bar{H}_m$ is an $(m+1) \times m$ upper Hessenberg matrix generated by the Arnoldi iteration.

In this case, the residual can be computed in terms of the orthonormal Q's and $\bar{H}_n$:

$$||Ax_m - b||_{\ell^2} = ||\bar{H}_m y_m - Q_{m+1}^T b||_{\ell^2} = ||\bar{H}_m y_m - \beta e_1||_{\ell^2},$$

with e_1 being the first column of the $(m+1) \times (m+1)$ identity matrix, and $\beta = ||b - Ax_0||$, with x_0 being some initial guess. In this sense, the problem has been reduced to finding the solution y_n to the least-squares minimization problem,

$$\arg\min_{y,} ||\bar{H}_m y_m - \beta e_1||_{\ell^2}. \tag{4.57}$$

Altogether, the algorithm in matrix form is

Algorithm 4.2.2 (GMRES)**.** Let x_0 be given initial guess.

Compute $\beta = ||b - Ax_0||_{\ell^2}$

While $||b - Ax_m||_{\ell^2} \neq 0$ **do**

Compute p_m with the Arnoldi method;

Solve the least-squares problem (4.57);

$x_m = Q_m y_m$;

endWhile.

The main motivation and useful feature of this solver is its robustness; indeed, it was proven that this method is guaranteed to converge in at most n iterations, even if the problem is non-symmetric, with a symmetric part that is indefinite (i.e, the real part of the eigenvalues of $M = \frac{A+A^T}{2}$ can be negative) [68]. However, in practice, it is not desirable to use the full GMRES, as this requires storing all of the previous p_m's, and computing the residual in (4.57) will also increase the number of multiplications, as $\frac{1}{2}m^2 n$. In this sense, the standard way to use GMRES is to keep only some maximal number (M) of previous p_m, and to restart the

algorithm with the initial guess being x_M. In this case, however, the guaranteed convergence is lost; in this case, the authors show that if the symmetric part of the matrix is not positive definite and one does not run the entire algorithm, then the residual can stagnate.

In particular, for matrices A that have positive definite symmetric part M, the authors get the following estimate on the error reduction:

$$||r_m||^2 \leq \left[1 - \frac{\lambda_{\min}^2(M)}{\lambda_{\max}(A^T A)}\right]^m ||r_0||. \tag{4.58}$$

For indefinite problems, the authors provide the following upper bound on the convergence rate:

Theorem 7 ([68], Theorem 5)**.** *Assume that A can be diagonalized as $A = XDX^{-1}$, and let*

$$\epsilon^m = \min_{p \in P_m,\, p(0)=1} \max_{\lambda_i \in \sigma} |p(\lambda_i)|.$$

Assume that there are ν eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_\nu$ of A with nonpositive real parts and let the other eigenvalues be enclosed in a ball of radius $R < C$ centered at a point $C > 0$. Then

$$||r_{m+1}|| \leq \kappa(X)\epsilon^m ||r_0||,$$

with an upper bound on ϵ^m,

$$\epsilon^m \leq \left[\frac{R}{C}\right]^{m-\nu} \max_{j=\nu+1,\ldots,n} \prod_{i=1}^{\nu} \frac{|\lambda_i - \lambda_j|}{|\lambda_i|} \leq \left[\frac{R}{C}\right]^{m-\nu} \left[\frac{D}{d}\right]^{\nu},$$

where

$$D = \max_{i=1,\ldots,\nu;j=\nu+1,\ldots,n} |\lambda_i - \lambda_j|, \quad d = \min_{i=1,\ldots,\nu} |\lambda_i|.$$

In this case, convergence of GMRES storing m previous iterates converges for any initial vector x_0 if

$$m\nu \log\left[\frac{DC}{dR}\kappa(X)^{1/\nu}\right] / \log\left[\frac{C}{R}\right].$$

In the case of (4.53), the problem is positive definite, so that any number of previous stored p_j's guarantees convergence, though the more that are stored, the better.

For the mixed discretization (4.54), the linear system, being of saddle point type, is indefinite; however, its Schur complement is well-known to be positive definite [69], which will be important for the preconditioners we use.

4.3 Multigrid algorithms

In this section, we describe an advanced iterative technique used to both solve and precondition linear systems that arrive from the discretization of elliptic PDE. It can be proven that the complexity of the algorithm is $\mathcal{O}(n)$, which implies that the work done per unknown is constant. The way that it manages this is that it relaxes errors by using standard relaxation (averaging on nearest neighbors) techniques on succesively coarser spaces, thus propagating information quickly through the mesh in a way that typical local relaxation techniques can't. Multigrid uses coarse grids to do divide-and-conquer in two related senses. First, it obtains an initial solution for an $n \times n$ grid by using an $(n/2) \times (n/2)$ grid as an approximation, which in turn takes an initial solution from an $(n/4) \times (n/4)$ grid, and so on, using simple interpolation and restriction techniques to transmit information between grids. Second, this smoothing and solving on coarser grids also divides the problem on the frequency domain, as the error on elliptic problems can be decomposed into sinusoidal eigenfunctions with different frequencies. In this sense, the problem on succesively coarser grids can efficiently divide and solve some portion of the error frequencies per grid level, by using cheap relaxation techniques.

To solve the system $Ax = b$, classical iterative solvers such as Jacobi, Gauss-Seidel, SOR work by taking some initial guess, decomposing the matrix A into some combination of its diagonal, lower triangular, or upper triangular parts, and iteratively inverting these simplified systems to get a correction that is then added to the initial guess. For instance, the simplest (Jacobi) iteration does the following:

Algorithm 4.3.1 (Jacobi method)**.** Let x_0 be given initial guess. Take $D = \text{diag}(A)$.

While $||r_k|| \geq tol$ **do**

Compute $r_k = b - Ax_k$.

Compute $e_k = D^{-1}r_k$;

$x_{k+1} = x_k + e_k$;

Set $k = k + 1$;

endWhile.

Such methods can be thought of as applications of operators that locally average values of the previous iterate to give the next iterate. It can be shown that the error becomes smooth after a few iterations of such a procedure, implying that the higher oscillatory modes of the error diminish very quickly. However, due to this local averaging property, the lower oscillatory modes of the error take considerably longer to diminish, resulting in slow convergence to the solution.

The core idea of multigrid is to take advantage of this property of rapid convergence of high frequency error modes by only performing a small number of one of these local averaging iterative methods to *relax* the error, so that the problem can be restricted to a coarser space, upon which the same relaxation can be performed, and restricted to an even coarser space. This is repeated until the problem is restricted on a coarse enough space where some direct solver can be used to solve for a coarse grid correction that can then be interpolated back up. The fixed number of iterations of the smoother on each level, combined with a cheap direct solve and cheap interpolation and restriction, can be shown to give a computational complexity of $\mathcal{O}(n)$ to the entire method, resulting in one of the fastest and computationally cheapest solvers for elliptic problems.

To define things more precisely, assume that we consider the problem on a grid with mesh size $h = 2^{-J}$. Let V_h be the corresponding finite element space. We consider the following sequence of subspaces of $V_J \equiv V_h$:

$$V_1 \subset V_2 \subset \ldots \subset V_{J-1} \subset V_J,$$

where V_k will be the finite element space corresponding to a grid with mesh size 2^{-k}. Let A_k denote the stiffness matrix corresponding to a discretization of the Poison equation on this grid, the operators R_k, P_k to be the restriction and interpolation operators between grids, and $n_k = \dim V_k$. Then the multigrid method is as follows:

1. Set initial guess $u^{(0)}$ and put $\ell = 0$, $L = 1$
2. **Do**
   (a) Compute $r_J = b - A_J u^{(\ell)}$

   #Loop down through the levels: \

   (b) **For** $k = J$ **to** 2 **step** -1 **do**

   #Pre-smoothing (with 0 initial guess):

(c) Set $e_k = R_k r_k$.

#Restrict the residual and keep it:

(d) Compute $r_{k-1} = (P^t_{k-1,k})(r_k - A_k e_k)$.

(e) **endFor**

Solve on the coarse grid, (for example by Conjugate Gradient method):

(f) $e_1 = A_1^{-1} r_1$.

#Loop up: /

(g) **For** $k = 2$ **to** J **step** 1 **do**

#Correct:

(h) Compute $e_k \leftarrow e_k + P_{k-1,k} e_{k-1}$.

Post smoothing:

(i) $e_k \leftarrow e_k + R^t_k(r_k - A_k)$.

(j) **endFor**

3. Update $u^{(\ell+1)} = u^{(\ell)} + e_J$; $\ell \leftarrow \ell + 1$.

4. **until** convergence

The above iteration is performed until some stopping criteria is satisfied. Typically, the restricion and interpolation operators can be taken as averaging operators. In the case considered above, $P_k : V_{k-1} \to V_k$ could work by averaging the degrees of freedom of nearest neighboring functions on the coarse space V_{k-1} to give the degree of freedom of a function on the finer space V_k, and the restriction operator R_k could take the dual action, which would take the value of a degree of freedom of a function on the coarse space V_{k-1} to be the average of its neighboring functions on the finer space V_k.

Given the explicit relationship of the coarse spaces and interpolation operators, this variant is known as Geometric multigrid, and is the first multigrid algorithm that was introduced. Such methods work well for sufficiently regular elliptic problems on geometrically regular domains, but for less regular domains, or problems where the magnitude of the entries of A in one row versus another vary greatly (i.e, in our case, where K and β can change by several orders of magnitude throughout Ω), one should consider forming coarse spaces not by the location of degrees of freedom in the domain, but by the relative weights and strength of connection in the entries of the matrix A (in other words, on *algebraic* properties of A); this

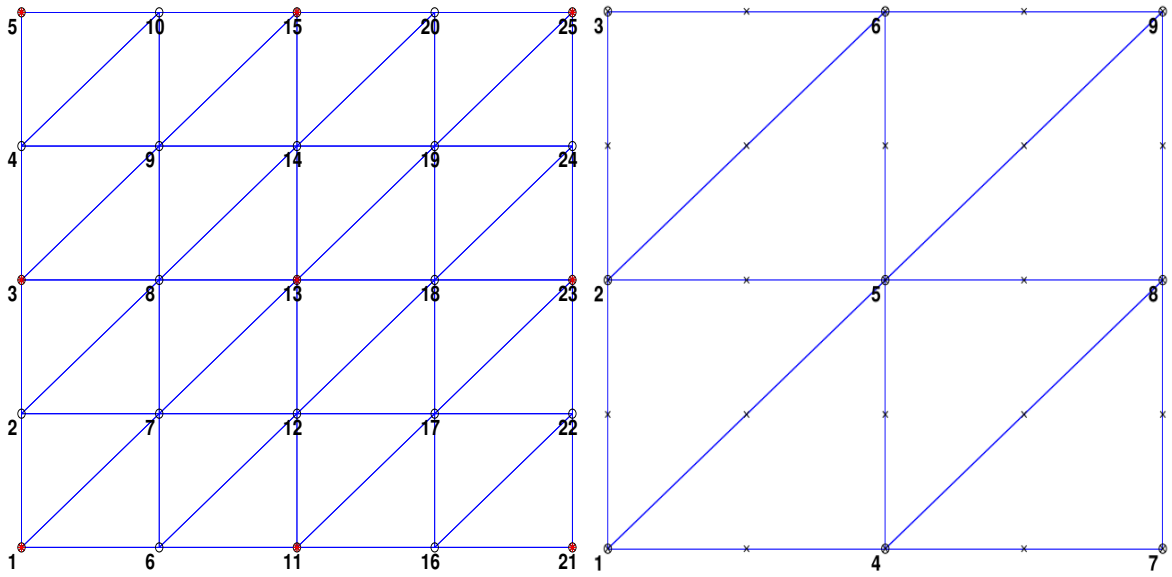

Figure 4.2. Fine and coarse grid

variant of multigrid is known as algebraic multigrid, which is what we use either to solve or to precondition the linear systems we work with.

4.3.1 Algebraic Multigrid method

The most important conceptual difference between geometric and algebraic multigrid is that geometric approaches employ fixed grid heirarchies and thus an efficient interplay between smoothing and coarse-grid correction has to be ensured by selecting appropriate smoothing processes. In contrast to this, AMG fixes the smoother to some simple relaxation scheme, like the damped Jacobi iteration, and enforces an efficient interplay with the coarse-grid correction by choosing the coarses levels and iterpolattion appropriately. This difference lends AMG a great deal of versatility in the types of problems it can solve, particularly problems in which a geometric grid heirarchy cannot be made a priori, or where the coefficients vary greatly, leading to various anisotropies that affect the convergence rate of geometric multigrid; for our problem in particular, K can vary by several orders of magnitude in different parts of the domain at different times, hence the motivation to use AMG, which trades consistent convergence for numerical work in developing the coarse spaces.

The overall procedure to solve the linear system with AMG is the same as the geometric algorithm; one uses a simple iterative method to presmooth the error on a fine grid, then restricts the problem to a coarser grid, and repeats until a coarsest grid is reached, where

the correction is solved for, and then interpolated back up and post smoothed, to give the solution on the finest grid. The key difference, and the bulk of the numerical work in using AMG, is in the construction of these coarse spaces, and their corresponding interpolation operators.

To define these notion of coarse and fine spaces, graph theory provides a natural language. An *undirected graph* (or simply a *graph*) $\mathcal{G}$ is a pair $(\mathcal{V}, \mathcal{E})$, where $\mathcal{V}$ is a finite set of points called *vertices* and $\mathcal{E}$ is a finite set of *edges*. As set of vertices we always consider subsets of $\{1, \ldots, n\}$ for some fixed n. An edge $e \in \mathcal{E}$ is an unordered pair $(j,\, k)$, where $j,\, k \in \mathcal{V}$. Similarly, a *directed graph* (or a *digraph*) $\mathcal{G}$, is a pair $(\mathcal{V},\, \mathcal{E})$, where $\mathcal{V}$ is the set of vertices as we just defined, but an edge $(j,\, k) \in \mathcal{E}$ is an ordered pair; that is, it indicates that there is a connection from j to k. We use the term *graph* to refer to both directed and undirected graphs.

If $(j,\, k)$ is an edge in an undirected graph $\mathcal{G} = (\mathcal{V},\, \mathcal{E})$, vertices j and k are said to be *adjacent*. If the graph is directed, vertex v is said to be *adjacent to* vertex w. A *path* from a vertex v to a vertex w is a sequence $(j_0,\, j_1,\, j_2, ...,\, j_l)$ of vertices where $j_0 = j,\, j_l = k$, and $(j_i,\, j_{i+1}) \in \mathcal{E}$ for all $i = 0,\, 1, ...,\, l - 1$. A vertex j is *connected* to a vertex k if there is a path from j to k. We shall adopt the convention, that every vertex is connected to itself. Two vertices, j and k, are said to be *strongly connected* if there is a path from j to k and a path from k to j. An undirected graph $\mathcal{G} = (\mathcal{V},\, \mathcal{E})$ is *connected* if every pair of vertices is connected by a path, otherwise it is said to be *disconnected*.

$$A = \begin{pmatrix} * & * & * & * & * & * \\ * & * & * & 0 & 0 & 0 \\ * & * & * & * & * & 0 \\ * & 0 & * & * & * & 0 \\ * & 0 & * & * & * & 0 \\ * & 0 & 0 & 0 & 0 & * \end{pmatrix}$$

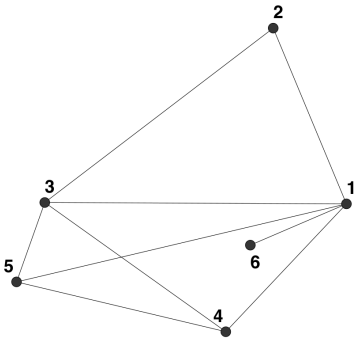

Figure 4.3. Sparsematrix (left) and the associated graph (right).

To relate graphs and sparse matrices, let us consider $A \in \mathbb{R}^{n \times n}$. The *adjacency graph* of

A is a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \{1, 2, ..., n\}$. The edges $\mathcal{E}$ are defined as $(j, k) \in \mathcal{E}$ if and only if $a_{jk} \neq 0$. We denote this graph by $\mathcal{G}(A)$. An example of a matrix with symmetric sparsity pattern and the corresponding graph is shown in Figure 4.3.

Let $\{\mathcal{V}_k\}_{k=1}^{n_c}$ be a non-overlapping splitting of the set of vertices

$$\mathcal{V} = \cup_{k=1}^{n_c} \mathcal{V}_k, \quad \mathcal{V}_j \cup \mathcal{V}_k = \emptyset, \quad \text{for} \quad j \neq k.$$

We then define

$$\mathcal{E}_k = \{(l, m) \in \mathcal{E} \;\Big|\; l \in \mathcal{V}_k \text{ and } m \in \mathcal{V}_k\} \tag{4.59}$$

to be the set of edges associated with $\mathcal{V}_k$.

4.3.2 Algebraic grid coarsening and interpolation

To describe a general two-level or multilevel multiplicative method, we denote $V = \mathbb{R}^n$, and also introduce a coarse space V_H, $V_H \subset V$, $n_H = \dim V_H$, $n_H < n$.

In order to transfer the correction up to the fine space, we need to define an operator $P : \mathbb{R}^{n_H} \mapsto \mathbb{R}^n$, such that $range(P) = V_H$. In this way, each column of P is formed by the coefficients in the expansion of a basis element from V_H via the finer basis in V:

$$\varphi_j^H = \sum_{i=1}^{n} p_{ij} \varphi_i^h, \quad j = 1, 2, \ldots, n_H. \tag{4.60}$$

Clearly P is full rank (because its columns are coefficients of a basis).

We would also like P to accurately interpolate smooth error back up to the fine space accurately, i.e,

$$[e_{\text{smooth}}]_i \approx \sum_{j=1}^{n_H} p_{ij} [e_{\text{smooth}}]_j, \quad i = 1, 2, \ldots, n. \tag{4.61}$$

A matrix that satisfies (4.60) and (4.61) is called a *prolongation* or *interpolation* matrix.

The construction of P usually contains two steps:

1. **Determine the sparsity of P:** The sparsity of P is determined by the sets of interpolatory vertices P_i. P_i is based on the connection between the F(ine)-vertex and neighboring C(oarse)-vertices. For reasons of efficiency and computational complexity, P_i should be a small subset of C vertices near i. This step is usually referred to *coarsening*.

2. **Compute p_{ij}:** Computing p_{ij} follows the principle that the smooth error should be

approximated well on the coarse level. It should be based on the characterization of the algebraic smooth error. One goal is to define p_{ij} such that (4.61) yields a reasonable approximation for any (algebraically) smooth error but does not require a large amount of computational work. This step is usually referred to as *construction of interpolation*.

The classical AMG method introduced by Brandt, McCormick and Ruge [70,71], is based on the observation that the algebraic smooth error varies less in the direction of relatively large (negative) off-diagonal coefficients of the matrix. This gives us an algebraic way to track the smooth error. In the classical formulation, the strength of connection between two degrees of freedom plays an important role in the construction of P; there are several viable metrics of strength of connection, all of which feature the generation of a "strength" adjacency graph that uses some measure of size of a non-zero off-diagonal entry a_{ij} in a row i relative to other nonzero entries in the same row to measure the strength of connection between i and j. This, combined with a ranking of new potential C-vertices generated via a "measure of importance" that initially prefers vertices that strongly connect to many vertices at the beginning of the coarsening, and then later prefers vertices that strongly connect to many fine vertices, produces consistent uniform coarsening.

Once this splitting of the unknowns into coarse and fine has been completed, constructing the interpolation operator then involves solving for the weights p_{ij} to minimize the error in (4.61). This is accomplished by observing that for smooth error, the error on the fine space should be orthogonal to the fine space in the Galerkin sense, i.e,

$$a_{ii}e_i + \sum_{j \in N_i} a_{ij}e_j \approx 0, \ i \in F, \tag{4.62}$$

where N_i is the total set of neighboring nodes connected to fine vertex i (not including i).

Using the relation (4.62), one can either use the nearest coarse neighbors to solve for the weights p_{ij} in (4.61) (direct coarsening), or we can include the nearest fine neighbors indirectly by using their respective coarse neighbors in the sum (indirect coarsening). Convergence estimates and the choice of optimal coarse space can be proven, but we will not go into such detail here. To read about these in more detail, a standard reference is [72].

4.3.3 Aggregation coarsening

In this section, we consider an alternate AMG algorithm based on *aggregation* coarsening. We focus on *unsmoothed* aggregation, as we use this method to solve and precondition certain

blocks of the saddle point system, or the primal system. The core idea is to first generate a collection of mutually disconnected aggregates $\mathfrak{a}$ of some subset of the adjacency graph of SPD operator A, and then define a very simple interpolation matrix P_0 ($p_{ij} = 1$ or 0) that can be used to relate average errors on each aggregate with errors on other aggregates. Applying this recursively with simple smoothers gives a numerically inexpensive algorithm for which the construction of the prolongation operator is easy to construct.

For the sake of exposition, we consider an SPD matrix $A \in \mathbb{R}^{n\times n}$ and the associated undirected adjacency graph $\mathcal{G} = \mathcal{V}, \mathcal{E}$ with vertices $|\mathcal{V}| = \{1, \ldots, n\}$ and set of edges $n_{\mathcal{E}} = |\mathcal{E}|$ and $n_{\mathcal{E}} \eqsim n$.

The term *aggregation* refers to a splitting of the set $\{1, \ldots, n\}$ in non-overlapping subsets. This can be done in many different and sophisticated ways, but since such algorithms are not our focus we introduce one of the simplest examples of such an algorithm:

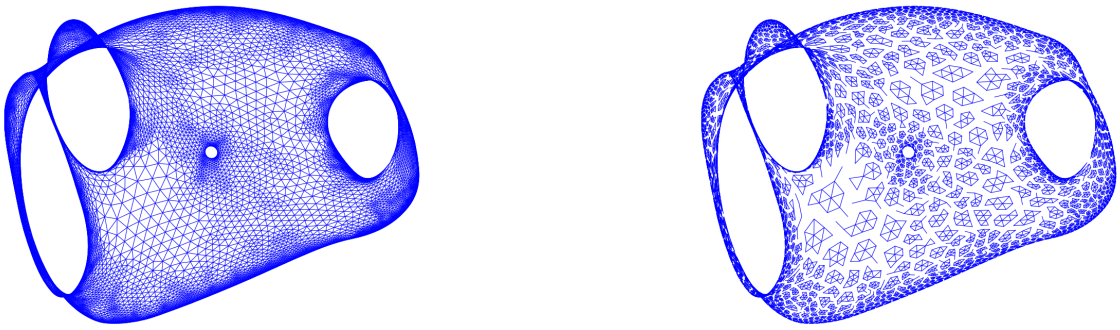

Figure 4.4. Graph of the matrix `barth5` from the University of Florida Sparse Matrix Collection [1] (left) and subgraphs formed by the greedy aggregation algorithm 4.3.2 (right). Courtesy of Ludmil. labelfig:example-sparse

Algorithm 4.3.2 (Greedy aggregation algorithm)**. Input:** Graph $\mathcal{G}$ with n vertices; **Output:** $\mathcal{V} = \cup_{k=1}^{n_c} \mathfrak{a}_k$, and $\mathfrak{a}_k \cap \mathfrak{a}_j = \emptyset$ when $k \neq j$.

1. Set $n_c = 0$ and for $k = 1 : n$ do:

   (a) If k and all its neighbors have not been visited, then: (a) we set $n_c = n_c + 1$; (b)

label with n_c the subgraph whose vertices are k and the neighbors of k; and (c) mark k and all its neighbors as visited.

(b) If at least one neighbor of k has been visited, we continue the loop over the vertices.

2. Since after this procedure there might be vertices which do not belong to any aggregate (but definitely have a neighboring aggregate), we add each such vertex to a neighboring aggregate and we pick the one which has minimal number of vertices in it.

3. The algorithm ends when all vertices are in a subset.

Such algorithm can be recursively applied and the resulting splitting is called an aggregation.

The splitting of the vertices in aggregates, naturally gives splitting of the adjacency graph of A. With each aggregate we associate a graph $\mathcal{G}_{\mathfrak{a}_k} = (\mathfrak{a}_k, \mathcal{E}_{\mathfrak{a}_k})$, where $\mathcal{E}_{\mathfrak{a}_k}$ is the subset of $\mathcal{E}$ of edges connecting vertices only from $\mathfrak{a}_k$. The following lemma is a corollary of Theorem 3.6 from [73], and is intuitive:

Lemma 4.3.1 (Corollary of [73]Thm 3.6 Kim, Xu, Zikatanov, 2003)**.** *If $\mathcal{G}$ is a connected graph, then the graphs corresponding to the aggregates $\{G_{\mathfrak{a}_k}\}_{k=1}^{n_c}$ obtained via Algorithm 4.3.2 are connected.*

The splitting of the adjacency graph is in bijective correspondence with the "piece-wise" constant vectors $\{\boldsymbol{\delta}_k\}_{k=1}^{n_c}$ defined as follows:

$$(\boldsymbol{\delta}_k)_j = \begin{cases} 1, & \text{if} \quad j \in \mathfrak{a}_k \\ 0, & \text{if} \quad j \notin \mathfrak{a}_k, \end{cases}, \quad j \in \mathcal{V}.$$

The matrix $P_0 = [\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_{n_c}]$ is called *tentative* (unsmoothed) prolongator. Clearly, range(P_0) has dimension n_c (by construction). In addition $[P_0^t P_0] \in \mathbb{R}^{n_c \times n_c}$ is a diagonal matrix and its k-th diagonal entry equals $|\mathfrak{a}_k|$.

We now consider the restriction of A on the space $V_H = \text{Range}(P_0)$

$$A_{H,0} = P_0^t A P_0,$$

If no cancellation is assumed in the triple product on the right side, the corresponding adjacency graph $\mathcal{G}_H(A_{H,0}) = (\mathcal{V}_H, \mathcal{E}_H)$ is formed by setting $V_H = \{1, \ldots, n_c\}$ and $(k, l) \in \mathcal{E}_H$ if and only if a vertex from $\mathfrak{a}_k$ is connected to a vertex in $\mathfrak{a}_l$ in $\mathcal{G}(A)$. A two-level and multilevel method utilizing P_0 is known as unsmoothed aggregation method.

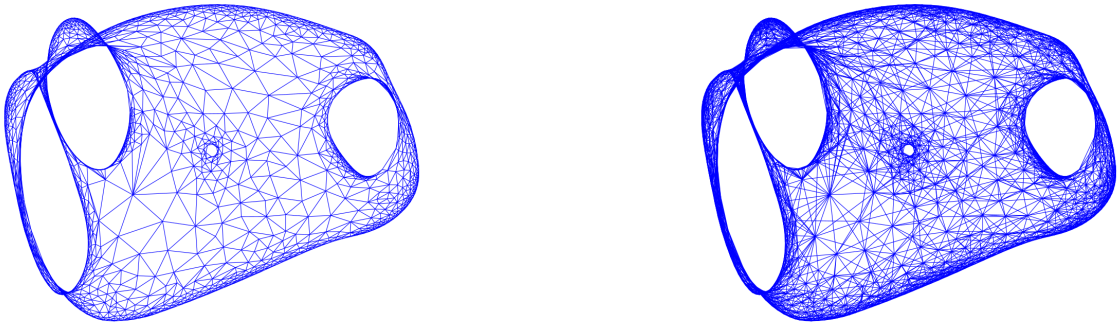

Figure 4.5. Graph of the coarse grid matrix corresponding to the unsmoothed aggregation (left) and the "denser" graph for the coarse grid matrix obtained by smoothed aggregation (right). Courtesy Ludmil.

An important property of this method is that it provides a sparse coarse matrix $A_{H,0}$, which neither smoothed aggregation nor classical AMG can guarantee; this is illustrated in figure 4.5, where the comparison of the coarse grid matrix after unsmoothed aggregation versus smoothed aggregation is plotted. However, it can be shown that due to the overly simplistic representation of error on the aggregates (constant on each aggregate), more iterations are typically required on the coarser levels, i.e, either W and F cycles must be used or the number of smoothing iterations per refinement level must be successively increased to guarantee convergence.

For the matrices that we use UA-AMG to precondition or solve, some (i.e, the mass matrix A_{qq}) are sufficiently well-conditioned that only a few applications are necessary to solve.

Chapter 5
Preconditioning Discretizations of the Linearized Richards equation

In this chapter, we focus on preconditioning the linear systems that result from discretizing any of the linearization schemes using the mixed finite element method as detailed in chapter 4. Section 5.1 explains the need for developing preconditioners for iterative solvers that are robust with respect to discretization parameters. In section 5.2, we first describe an iterative scheme to solve the indefinite saddle point systems that need to be solved on each iteration, which we call the Schur iteration. We prove that this iterative scheme can be used to precondition a GMRES solve of the saddle point system, uniform with respect to discretization parameters τ (timestep size), and h (characteristic mesh size), given a spectrally equivalent, sparse approximate Schur complement $\widetilde{S}_R$ that must be solved on each Schur iteration, and tuning parameter ω_R whose values are uniformly bounded due to the spectral equivalence of the approximate Schur complement and full Schur complement. To precondition the approximate Schur complement $\widetilde{S}_R$, we also introduce an auxiliary space preconditioner in section 5.3 that uses the Edge Average Finite Element scheme as described in the previous chapter as auxiliary space, an interpolation operator Π_{VS} that computes local averages to transfer between element-based degrees of freedom and nodal degrees of freedom, and a Jacobi smoother. By considering a representation of the bilinear form for $\widetilde{S}_R$ that considers differences across faces in the mesh as compared to the EAFE representation of the auxiliary problem A_E as differences of function values along edges of the mesh, we are able to prove the uniformity of the proposed auxiliary space preconditioner with respect to time and step size in arbitrary dimensions.

Finally, section 5.4 details some numerical experiments that verify the uniformity of the preconditioner for the mixed finite element-modified Picard discretization, first for an

exponential model of θ_N and K wherein the modified Picard iteration is used to linearize the Richards equation, and then for the VGM model described in chapter 1, for various materials. Finally, for mild K gradients, we show numerical tests that show both preconditioners also work when using Newton-Raphson, i.e, when the saddle point system and auxiliary linear systems are nonsymmetric. We finish with making some mention of potential future avenues to refine these preconditioners further, particularly for the case of steeper K gradients with Newton-Raphson.

For the sake of the reader, we reproduce the linear systems (4.53) and (4.54), introducing a $(-)$ to both sides of the $P1$-Lagrange system for consistency between the discretizations:

$$-(D_\theta + \tau A_E)P_\epsilon = -R^\kappa, \tag{5.1}$$

$$\begin{bmatrix} A_{qq} & B_{\text{div}}^T - B_{K'} \\ B_{\text{div}} & -D_\theta \end{bmatrix} \begin{pmatrix} q_\epsilon \\ \Psi_\epsilon \end{pmatrix} = \begin{pmatrix} \tilde{f} \\ \tilde{g} \end{pmatrix}. \tag{5.2}$$

5.1 Preconditioning: a primer

As was mentioned in the previous chapter, the poor conditioning of these systems as the mesh size $h \to 0$ require the use of preconditioners to solve these systems adequately. When solving a system $Ax = b$ with an iterative method, a preconditioner B on A acts to improve the condition of the linear system A, thus significantly improving efficiency of the iterative solver. More precisely, in the majority of cases, a *preconditioner* is an operator B that can act on A in one of two ways:

1. *Left* preconditioners are operators $z_i \leftarrow B^{-1}(Ax_i - b)$ that approximate A in some sense, but are much simpler to solve. The output z_i is then fed into the iterative algorithm as the transformed iterate $\hat{x}_i$; this is repeated until convergence.

2. *Right* preconditioners are operators that perform two solves, the first being a solve of $y_i \leftarrow AB^{-1}y_i = b$, and then $\hat{x}_i \leftarrow Bx_i = y_i$. The transformed iterate $\hat{x}_i$ is then fed into the iterative algorithm.

In general, left preconditioners are simpler to implement, and are the most commonly used. Roughly speaking, good preconditioners are operators that are "similar" to the original

problem, and are easier to solve. The goal of a left preconditioner, for instance, is that

$$1 \leq \kappa(BA) << \kappa(A),$$

where again the condition number of a system $\kappa(M) = ||M^{-1}||_X ||M||_X$. In the case of SPD matrices, this can be interpreted by saying that B is spectrally equivalent to $A : V_h \to V_h'$, i.e there exist constants μ_0, $\mu_1 > 0$ such that

$$\mu_0(Bx, x) \leq (Ax, x) \leq \mu_1(Bx, x), \quad x \in V_h. \tag{5.3}$$

The implication here is that the spectrum of the two operators should be close to each other, with $Br = z$ requiring less computational effort to solve. Some basic examples of preconditioners include other iterative solvers that take B to be some part of A, like damped Jacobi (diagonal of A) or Gauss-Seidel (Lower and Upper triangular parts of A). Multigrid is more often used to precondition linear systems generated by discretizing elliptic PDE than to solve them, to great effect in many cases.

5.2 Auxiliary space preconditioning

In chapter 3, we described two finite element discretizations for the same problem, with their various discretizations. The standard $P1$-Lagrange finite element discretization with EAFE discretization provides a discretization in finite dimensional subspaces of $H_0^1(\mathcal{T}_h)$, whose degrees of freedom are nodal. While this method is shown to be monotone and satisfy a discrete maximum principle through automatic upwinding, local mass conservation is lost. One way to retain local mass conservation is by using mixed finite elements, in which continuity of the fluid flux through faces is enforced and monitored via the addition of more degrees of freedom that track the flux; the price that is paid in doing this, however, is in a severe increase in the number of DOF of the system, and the generation of a potentially indefinite linear system. However, heuristically speaking, the solution of the two systems should be "close" to each other, and any discrepancy should decrease as the mesh size goes to 0.

A natural question then arises: can one solve the smaller (coarse) system for $P1$-Lagrange pressures defined on the nodes of the mesh, and use the answer as a good initial guess for the solution of the (larger) piecewise-constant pressure unknowns defined on each element? In this chapter, we gain a positive result to the above question, which we state as Corollary 1,

as an application of the Fictitious Space Lemma 8, and verify with numerical examples. This section makes this idea rigorous by defining the notion of auxiliary spaces that, when accompanied with a smoother, can allow for the preconditioning of one finer discretization of a problem with a coarser discretization of the same problem.

5.2.1 Background

Auxiliary space preconditioning is an application of the theory of fictitious spaces, which originated in the work of S.Nepomnyaschikh [74], who was working on robust preconditioners for $P1$-Lagrange discretizations of elliptic boundary value problems over domains Ω with complex geometry, such as non-convex polygons, or unstuctured grids. The core idea was the construction of an *auxiliary* grid on a larger domain $\bar{\Omega} \supset \Omega$ that had much simpler geometry and quasi-uniform mesh size. By splitting the domain Ω into interior nodes that attain a value of 0 on the boundary, and boundary nodes whose value could be interpolated from the trace space $H^{1/2}(\Omega)$ to $H^1(\Omega)$ inside the domain. The idea was that, provided the interpolation operator was extended in a way that preserves the norm, the problem could be discretized on the auxiliary mesh with finite element approximation space $\bar{V}_h$, approximately solved there, and then mapped back to the finite element space V_h on the original domain by using the restriction operator $\Pi : \bar{V} \to V$,to yield a uniform preconditioner for the original problem.

To make this notion rigorous, given a variational problem

$$u \in V : \quad a(u, v) = f(v) \quad \forall v \in V, \tag{5.4}$$

in some real Hilbert space V, if a is SPD, the bilinear form $a(u, v)$ is an inner product which induces a norm $||\cdot||_A$ on V. The building blocks of the fictitious space method are

1. a *fictitious space* $\bar{V}$, that is, another real Hilbert space equipped with another SPD bilinear form $\bar{a}(\bar{u}, \bar{v})$ with induced norm $||\cdot||_{\bar{A}}$, and
2. a continuous and surjective linear transfer operator $\Pi : \bar{V} \to V$.

Indicating the dual of a space with $'$, adjoint operators by $*$, and using angle brackets for duality pairings, one can define the fictitious space preconditioner

$$B = \Pi \cdot \bar{A}^{-1} \cdot \Pi^* : \quad V' \to V, \tag{5.5}$$

with the operators $A : V \to V'$ and $\bar{A} : \bar{V} \to \bar{V}'$ denoting the isomorphisms associated with bilinear forms $a(\cdot, \cdot)$ and $\bar{a}(\cdot, \cdot)$, resp. The operator $\Pi : \bar{V} \to V$ being surjective and the

operator $\bar{A}$ being SPD yield that the preconditioner B is SPD, by the Banach open mapping theorem. Given this, the uniformity of the preconditioner B for A is a result of the Fictitious Space Lemma, first proven by [74]:

Theorem 8 ([75], Thm 2.2)**.** *Assume that $\Pi : \bar{V} \to V$ is surjective and the following two conditions hold:*

$$\exists c_0 > 0 : \quad \forall v \in V : \quad \exists \bar{v} \in \bar{V} : \quad v = \Pi\bar{v} \text{ and } ||\bar{v}||_{\bar{A}} \leq c_0||v||_A, \tag{5.6}$$

$$\exists c_1 > 0 : ||\Pi\bar{v}||_A \leq c_1||\bar{v}||_{\bar{A}} \quad \forall \bar{v} \in \bar{V}. \tag{5.7}$$

Then

$$c_0^{-2}||v||_A^2 \leq a(BAv, v) \leq c_1^2||v||_A \quad \forall v \in V. \tag{5.8}$$

(5.8) immediately gives an estimate on the spectral condition number of the preconditioned system,

$$\kappa(BA) \equiv \frac{\lambda_{\max}(BA)}{\lambda_{\min}(BA)} \leq (c_0c_1)^2. \tag{5.9}$$

In essence, the key to this method working is the transfer operator $\Pi : \bar{V} \to V$ being continuous, but especially, being surjective. In the context of Nepomnyaschikh's choice of fictitious space, surjectivity is essentially given trivially, as the fictitious domain $\bar{\Omega} \supset \Omega$. The continuity along the boundary $\partial\Omega$ is what needed to be paid attention to, but with the right choice of interpolation for the boundary data, $t : H^{1/2}(\partial\Omega) \to H^1(\Omega)$, Nepomnyaschikh was able to preserve the continuity of the norm as required.

In the context of auxiliary spaces, the idea is actually the opposite– the idea was to be able to precondition an SPD bilinear form $A : V \to V'$, with SPD bilinear forms defined on an *auxiliary space* $\bar{A}_j : W_j \to W_j'$, with W_j being *smaller* than V.

In this sense, the surjectivity would have to be imposed by enriching the lower dimension spaces W_j with V itself, by introducing the following space as the fictitious space $\bar{V}$:

$$\bar{V} = V \times W_1, \tag{5.10}$$

where W_1 is a Hilbert space endowed with inner product $\bar{a}_1(\cdot, \cdot)$, and the inner product taken on V is induced by some other SPD operator $S : V \to V'$, denoted the *smoother*. Then we can define the fictitious inner product

$$\bar{a}(\bar{v}, \bar{v}) \equiv s(v_0, v_0) + \bar{a}_1(w_1, w_1), \quad \forall \bar{v} = (v_0, w_1) \in \bar{V}, \text{ with } v_0 \in V, w_1 \in W_1, \tag{5.11}$$

and the surjective operator

$$\Pi \equiv \begin{pmatrix} I & \\ & \Pi_1 \end{pmatrix} : \bar{V} \to V. \tag{5.12}$$

Thus, the auxiliary space preconditioner can be formed:

$$B = S^{-1} + \Pi_1 \cdot \bar{A}_1^{-1} \cdot \Pi_1^*. \tag{5.13}$$

The verification of the assumptions of theorem 8 can be shown in three steps:

1. Find bound $c_1 > 0$ for the norm of the transfer operator Π_1:

$$||\Pi_1 w_1||_A \leq c_1 \bar{a}(w_1, w_1)^{1/2}, \quad w_1 \in W_1.$$

2. Investigate the continuity of S^{-1}:

$$\exists c_s > 0 : \ ||v||_A \leq c_s s(v, v)^{1/2}, \quad \forall v_{\in} V.$$

3. Establish that for every $v \in V$ there are $v_0 \in V$ and $w_1 \in W_1$ such that $v = v_0 + \Pi_1 w_1$ and

$$s(v_0, v_0) + \bar{a}_1(w_1, w_1) \leq c_0^2 ||v||_A^2,$$

where $c_0 > 0$ should be small and independent of v.

In this context, theorem 8 holds, and the estimate on the condition number (5.9) is of form

$$\kappa(BA) \leq c_0^2(c_s^2 + c_1^2). \tag{5.14}$$

This method was initially used in the context of domain decomposition and parallel subspace correction, where multiple subspaces W_j were disjoint subspaces of V_h on subsets of Ω, which could be solved on approximately and interpolated back up, or in the case of nodal approximations with non-conforming elements, could be preconditioned by conforming nodal approximations, ([76], [77], [78]). However, Hiptmair and Xu [75] proposed a different context, in particular for preconditioning $\mathrm{div} - \mathrm{div}$ systems and $\mathrm{curl} - \mathrm{curl}$ systems with $P1$-Lagrange discretizations (i.e, nodal preconditionining). The core of this idea lies in the fact that on a regular enough domain Ω, the spaces $H_0(\mathrm{curl}, \Omega)$ and $H_0(\mathrm{div}, \Omega)$ can be decomposed into stable auxiliary spaces using potential theory. This stable splitting allows the construction of preconditioners that, with the introduction of a "high-frequency"

smoother, are uniform, with the only conditions being convexity of the boundary (so-called "2-regularity"), and quasi-uniformity of the mesh, with the preconditioners being solves of Laplacian-type operators discretized on nodal elements, which can be performed efficiently using AMG and multigrid-like algorithms.

The theory of HX auxiliary space preconditioners has been extended to other problems that can be discretized with mixed formulations, such as Darcy flow [79], flow in fractured porous media [80], time-dependent Maxwell equations [81], and Biot's equations [82]. Our goal was to investigate the potential of auxiliary space preconditioners for Richards' equation, which can be interpreted as a Darcy solve per linear solve, with conductivity and water content functions that are allowed to vary between solves.

5.3 Preconditioning the Saddle Point System

Before introducing the auxiliary space preconditioner, we first introduce our Schur iteration, which is an inexact Uzawa method used to approximately solve the pressure Schur complement of the system,

$$S_R := -(D_\theta + B_{\text{div}} A_{qq}^{-1} B_{\text{div}}^T). \tag{5.15}$$

For a brief overview of inexact Uzawa algorithms for saddle point problems, see [83], [84], [85].

One such iteration for the indefinite problem (4.54) is given in Algorithm 5.3.1.

Algorithm 5.3.1 (Schur iteration)**.** Given initial guess $\left(q_\epsilon^0,\, \Psi_\epsilon^0\right)$, we use the following recurrence relation to define $\left(q_\epsilon^{k+1},\, \Psi_\epsilon^{k+1}\right)$ in terms of the k-th iterates:

1. Solve $A_{qq} u = \widetilde{f} - B_{\text{div}}^T \Psi_\epsilon^k$;
2. Solve,
$$\widetilde{S}_R v = \widetilde{g} + D_\theta \Psi_\epsilon^k - B_{\text{div}} u, \tag{5.16}$$
where $0 < \omega_R$ is properly chosen in advance;
3. Update Ψ_ϵ as $\Psi_\epsilon^{k+1} = \Psi_\epsilon^k + \omega_R v$ and then solve again with A_{qq}, namely, solve
$$A_{qq} w = -B_{\text{div}}^T v;$$
4. Update q_ϵ as $q_\epsilon^{k+1} = u + w$.

We note that while this method requires three linear solves in total (one of the approximate Schur complement $\widetilde{S}_R$, and two of the much larger mass matrix A_{qq}), A_{qq} is already SPD and well-conditioned, so this system can be solved very efficiently using standard iterative techniques. We use Conjugate Gradient with unsmoothed aggregation AMG as preconditioner, and this yields fast solves to small tolerances with little expense. The more difficult problem is solving the $\widetilde{S}_R$ system, which is where we use our proposed aux space preconditioner to precondition a Conjugate Gradient (GMRES for nonsymmetric) solver.

In order for our iterative scheme to be uniformly convergent with respect to both discretization parameters, we need to choose $\widetilde{S}_R$ that is spectrally equivalent to S_R.

It is shown in [86] that A_{qq} and its diagonal D_{qq} are spectrally equivalent, i.e, $\exists \mu_0, \mu_1 > 0$ with

$$\mu_0(D_{qq}v, v) \leq (A_{qq}v, v) \leq \mu_1(D_{qq}v, v), \quad v \in S_h.$$

As both D_{qq} and A_{qq} are SPD, their inverses are as well, and choosing $v = D_{qq}^{-1/2}A_{qq}^{-1/2}w$ and manipulating yields

$$\mu_0(A_{qq}^{-1}w, w) \leq (D_{qq}^{-1}w, w) \leq \mu_1(A_{qq}^{-1}w, w), \quad w \in S_h.$$

Then, taking $w = B_{\text{div}}^T v$ in the above yields

$$\mu_0(B_{\text{div}}A_{qq}^{-1}B_{\text{div}}^T v, v) \leq (B_{\text{div}}D_{qq}^{-1}B_{\text{div}}^T v, v) \leq \mu_1(B_{\text{div}}A_{qq}^{-1}B_{\text{div}}^T v, v), \quad v \in S_h. \tag{5.17}$$

Then adding in D_θ shows that $\widetilde{S}_R$ defined as

$$\widetilde{S}_R := -\left(D_\theta + B_{\text{div}}D_{qq}^{-1}B_{\text{div}}^T\right). \tag{5.18}$$

Is spectrally equivalent to the full Schur complement S_R.

As the inverse of A_{qq} is generally a full matrix, this choice allows us to approximate S_R with a sparse, spectrally equivalent operator whose action can be formulated as sums and products of differences of values across faces, which will be important in the proofs of uniformity of our auxiliary space preconditioner.

We now show that Algorithm 5.3.1 converges for certain choices of ω_R, independent of discretization parameters; this proof is similar in approach to many standard approaches in the literature on iterative methods (see, e.g. Young [87]).

Lemma 5.3.1. *For sufficiently small ω_R, Ψ_ϵ^k and q_ϵ^k obtained by Algorithm 5.3.1 converge to the solution of* (4.54)*.*

Proof. We first consider $\Psi_\epsilon^{k+1} - \Psi_\epsilon^k$. Note that if $\begin{pmatrix} q_\epsilon \\ \Psi_\epsilon \end{pmatrix}$ is a solution to (4.54), then $A_{qq}q_\epsilon + B_{\text{div}}^T\Psi_\epsilon = \tilde{f}$ and $B_{\text{div}}q_\epsilon - D_\theta\Psi_\epsilon = \tilde{g}$. We have,

$$\begin{aligned}
\Psi_\epsilon^{k+1} - \Psi_\epsilon &= \Psi_\epsilon^k + v - \Psi_\epsilon \\
&= \Psi_\epsilon^k + \omega_R\widetilde{S}_R^{-1}\left[\underbrace{B_{\text{div}}q_\epsilon - D_\theta\Psi_\epsilon}_{\tilde{g}} + D_\theta\Psi_\epsilon^k - B_{\text{div}}A_{qq}^{-1}(\tilde{f} - B_{\text{div}}^T\Psi_\epsilon^k)\right] - \Psi_\epsilon \\
&= \Psi_\epsilon^k - \Psi_\epsilon + \omega_R\widetilde{S}_R^{-1}\left[B_{\text{div}}q_\epsilon + D_\theta(\Psi_\epsilon^k - \Psi_\epsilon) - B_{\text{div}}A_{qq}^{-1}(\underbrace{A_{qq}q_\epsilon + B_{\text{div}}^T\Psi_\epsilon}_{\tilde{f}} - B_{\text{div}}^T\Psi_\epsilon^k)\right] \\
&= \Psi_\epsilon^k - \Psi_\epsilon + \omega_R\widetilde{S}_R^{-1}\left[D_\theta(\Psi_\epsilon^k - \Psi_\epsilon) + B_{\text{div}}A_{qq}^{-1}B_{\text{div}}^T(\Psi_\epsilon^k - \Psi_\epsilon)\right] \\
&= (I - \omega_R\widetilde{S}_R^{-1}S_R)(\Psi_\epsilon^k - \Psi_\epsilon).
\end{aligned}$$

Thus, in order for $\Psi_\epsilon^k \to \Psi_\epsilon$ as $k \to \infty$, we need

$$\rho(I - \omega_R\widetilde{S}_R^{-1}S_R) < 1.$$

Elementary manipulations yield a condition for ω_R:

$$0 < \omega_R < \frac{2}{\rho(\widetilde{S}_R^{-1}S_R)}. \tag{5.19}$$

Hence, any choice of ω_R satisfying (5.19) guarantees the convergence $\Psi_\epsilon^k \to \Psi_\epsilon$ as $k \to \infty$, and due to the spectral equivalence of $\widetilde{S}_R$ to S_R, this bound is independent of discretization parameters.

On the other hand, for $q_\epsilon^{k+1} - q_\epsilon$ we have

$$\begin{aligned}
q_\epsilon^{k+1} &= u + w = A_{qq}^{-1}(\tilde{f} - B_{\text{div}}^T\Psi_\epsilon^k) - A_{qq}^{-1}B_{\text{div}}^T v \\
&= A_{qq}^{-1}(A_{qq}q_\epsilon + B_{\text{div}}^T\Psi_\epsilon - B_{\text{div}}^T\Psi_\epsilon^k) - A_{qq}^{-1}B_{\text{div}}^T v \\
&= q_\epsilon + A_{qq}^{-1}B_{\text{div}}^T(\Psi_\epsilon - \Psi_\epsilon^k - v) = q_\epsilon - A_{qq}^{-1}B_{\text{div}}^T(\Psi_\epsilon^{k+1} - \Psi_\epsilon).
\end{aligned}$$

As a result from this relation we get

$$q_\epsilon - q_\epsilon^{k+1} = A_{qq}^{-1}B_{\text{div}}^T(\Psi_\epsilon^{k+1} - \Psi_\epsilon).$$

Thus, since $\Psi_\epsilon^k \to \Psi_\epsilon$ and $A_{qq}^{-1}B_{\text{div}}^T$ is a bounded operator with our choice of finite-element

spaces, $q_\epsilon^k \to q_\epsilon$ as well. $\square$

We note here that while simply using the diagonal of A_{qq} in (5.18) suffices for our purposes, this approximation can be improved on. We refer to [88] for mass lumping techniques approximating A_{qq} by a diagonal matrix with higher order accuracy.

Remark 5.3.1 (On the sign of S_R and $\widetilde{S}_R$)**.** One should note that although $\widetilde{S}_R$ and S_R are actually negative definite in their definitions above, one can simply negate them and their corresponding right hand sides to arrive at positive definite problems, which is what we do for the rest of this chapter.

5.4 Auxiliary space preconditioning the Schur complement

To motivate the auxiliary space preconditioner for (5.16), we note that the bilinear form generated by our particular choice of $\widetilde{S}_R$ can be interpreted on each element as sums of differences of values of Ψ with neighboring elements across each face of the element, with weighting given by values of K at each element. An analogous formulation of convection-diffusion operators as sums of differences along edges can also be derived ([19], [89]). Given this similarity, we decided to adapt the methods of [75] to develop an auxiliary space preconditioner, using the nodal discretization combined with a simple smoother and some interpolation map to precondition the larger system (5.16).

Thus, we introduce an auxiliary space preconditioner for (5.16) and subsequently prove its uniformity on a shape regular mesh for $d = 2$, 3 or 4 (for space-time finite elements). This preconditioner offers two distinct advantages. First, we show that it is uniform with respect to the mesh size and timestep τ. Second, the number of vertices of any triangulation $\mathcal{T}_h$ with no hanging nodes is significantly smaller than the number of simplices forming the triangulation. This can be seen by considering the case of a uniform lattice of size h of the unit square $[0,1]^d$ with $N = h^{-1}$ being an integer. The number of simplices in the mesh is $d!N^d$, while the number of vertices is $(N+1)^d$. Thus, the number of degrees of freedom of our auxiliary space is on the order of $\frac{1}{d!}$ the number of degrees of freedom of the $\widetilde{S}_R$ system that must be solved on each step of our Schur iteration 5.3.1. On the coarse space, multilevel preconditioners can be further used to speed up the computations.

To define the auxiliary space preconditioner, we follow in the spirit of [75] and introduce the following components.

- The ficticious space $\bar{V} = S_h \times V_h$, with $V_h \subset H_0^1(\Omega)$ being the space of piece-wise linear and continuous functions with zero trace on the boundary, and $S_h \subset L^2(\Omega)$ being the space of functions that are piecewise constant on each element $T \in \mathcal{T}_h$.

- The map between the auxiliary space and S_h, $\Pi = \begin{pmatrix} I & \Pi_{VS} \end{pmatrix}$, with $\Pi_{VS} : V_h \to S_h$. The action of Π_{VS} amounts to taking the average per element T of the values of v on its vertices $j \in T$, namely, given $v \in V_h$, $v = \sum_{i=1}^{NV} v_i \varphi_i$, we define

$$[\Pi_{VS}(v)]_T := p_T = \frac{1}{d+1} \sum_{j \in T} v_j, \quad \Pi_{VS}(v) = \sum_T p_T \chi_T, \quad \text{for all} \quad v \in V_h.$$

- Our smoother is the Jacobi smoother, which just uses the diagonal of $\widetilde{S}_R$ (which may be scaled if needed);

$$D_{\widetilde{S}} : S_h \to S_h' = \text{diag}(\widetilde{S}_R).$$

The preconditioner B is then

$$B = \Pi \begin{pmatrix} D_{\widetilde{S}} & 0 \\ 0 & A_{\text{Lgr}} \end{pmatrix}^{-1} \Pi^* = D_{\widetilde{S}}^{-1} + \Pi_J A_{\text{Lgr}}^{-1} \Pi_J^*, \tag{5.20}$$

where A_{Lgr} denotes the $P1$-Lagrange discretization (5.1).

Given a right hand side r, an algorithm for the action of B is as follows.

Algorithm 5.4.1 (Auxiliary space preconditioner B: $z \leftarrow Br$).

1. Transfer the right hand side r to the auxiliary space V_h: $r_{V_h} \leftarrow \Pi r$,
2. Solve the auxiliary problem on V_h: $e_{V_h} \leftarrow A_{\text{Lgr}}^{-1} r_{V_h}$,
3. Transfer the corection e_{V_h} back to S_h: $z \leftarrow \Pi^* e_{V_h}$,
4. Smooth the correction with Jacobi iteration: $z \leftarrow z + D_{\widetilde{S}}^{-1}(r - \widetilde{S}_R z)$.

To discretize the problem on the auxiliary space, we wish to use a discretization on V_h that formulates the diffusion operator on vertices as differences along edges of elements, as that facilitates the analysis we use to prove the uniformity of our preconditioner. To this effect, we elect to discretize using EAFE [19]. We do this for one reason, namely that this discretization provides a monotone discretization on V_h for both symmetric and non-symmetric (i.e, convection-diffusion) problems. Further, the relatively lax requirement

of the triangulation being Delaunay is easy to satisfy if using certain mesh refinement schemes ([60], [90]).

The proof of uniformity of B requires that we use spectrally equivalent forms of A_{Lgr} and $\widetilde{S}_R$:

$$(A_{\text{Lgr}}v, w) = \sum_{E\in\mathcal{T}_h} \omega_E \delta_E v \delta_E w \quad \text{and} \quad (\widetilde{S}_R p, s) = \sum_{f\in\mathcal{T}_h} d_f (p_{T^+} - p_{T^-})(s_{T^+} - s_{T^-}). \tag{5.21}$$

where for an edge E connecting vertices i and j, where we assume $i > j$ given some ordering on the vertices of $\mathcal{T}_h$, we define $\delta_E f = f(i) - f(j)$, and for a face $f \in \mathcal{T}_h$, an assigned ordering of $\mathcal{T}_h$, and a function $f \in S_h$, we take f_+ and f_- to be the value of f at the higher (resp. lower) numbered simplex sharing the face. Direct computations yield

$$d_f = \tau \left[\int_\Omega K^{-1} |\phi_f|^2 \, \mathrm{d}x \right]^{-1}.$$

For the symmetric problem the weight

$$\omega_E = \tau \left[\frac{1}{|\tau_E|} \int_E K^{-1} \, \mathrm{d}s \right]^{-1} \widetilde{\omega}_E,$$

with $|\tau_E| = |\delta_E x|$ being the length of edge E, and $\widetilde{\omega}_E$ being the weights defined as in (4.14), which only uses geometric properties of the individual simplices, namely the angle between faces across edges and lengths of edges opposite these angles. The scaling of both of these coefficients with respect to mesh size is h^{d-2} and, as is immediately seen, the ratio of these terms is only dependent on mesh geometry and harmonic average values of K over edges versus over elements.

Taking $s = p$ and $v = w$, we arrive at

$$(A_E v, v) = \sum_{E\in\mathcal{T}_h} \omega_E (\delta_E v)^2 \quad \text{and} \quad (\widetilde{S}_R p, p) = \sum_{f\in\mathcal{T}_h} d_f (p_{T^+} - p_{T^-})^2. \tag{5.22}$$

Next, we show that (5.20) is a uniform preconditioner for the problem (5.16). To this end, we introduce the following notation,

- Ω_i is the subdomain consisting of simplices sharing vertex i, $\Omega_i = \cup_{T\ni i} T$,
- $\mathcal{F}_i$ is the set of faces containing vertex i, $\mathcal{F}_i = \{f \ni i\}$,
- $\mathcal{N}_i$ is the number of simplices sharing vertex i,

- $\mathcal{N}_i^f$ is the number of faces in $\mathcal{F}_i$,
- For an edge E with vertices i and j, $\mathcal{N}_{i\cup j}$ is the number of simplices in $\Omega_i \cup \Omega_j$, and $\mathcal{N}_{i\cap j}$ is the number of simplices in $\Omega_i \cap \Omega_j$.

Corollary 1 (Uniformity of the preconditioner)**.** *The auxiliary space preconditioner defined in* (5.20) *provides a uniform preconditioner for* $\widetilde{S}_R$*.*

Proof. To prove that this auxiliary space preconditioner is uniform, it is sufficient to prove the following three properties of the transfer operator Π_{VS}, the auxiliary problem A_{Lgr}, and the smoother $D_{\widetilde{S}}$ hold independent of h and τ.

Lemma 5.4.1 (Continuity of Π_{VS})**.** *There exists* $c_J > 0$ *such that*

$$(\widetilde{S}_R \Pi_{VS} v, \Pi_{VS} v)^2 \leq c_J^2 (A_{Lgr} v,\, v), \quad \forall v \in V_h. \tag{5.23}$$

with c_J *independent of mesh size.*

Proof. Let $v \in V_h$. To show (5.23), we use the definition of Π_{Lgr} and the relation (5.22),

$$\begin{aligned}(\widetilde{S}_R \Pi_{VS} v, \Pi_{VS} v)^2 &= \sum_{f\in\mathcal{T}_h} d_f \left(\frac{1}{d+1} \sum_{i\in T^+} v_i - \frac{1}{d+1} \sum_{j\in T^-} v_j \right)^2 \\ &= \sum_{f\in\mathcal{T}_h} \frac{d_f}{(d+1)^2} (v_{f,+} - v_{f,-})^2,\end{aligned}$$

where $v_{f,+}$ and $v_{f,-}$ are the values of v at the vertices in T^+ and T^- opposite face f, respectively. These vertices are not connected by any edge in $\mathcal{T}_h$, but can be connected via two edges $E^+ \in T^+$ and $E^- \in T-$. Using this fact we can relate the two bilinear forms:

$$\begin{aligned}&\sum_{f\in\mathcal{T}_h} \frac{d_f}{(d+1)^2} (v_{f,+} - v_{f,-})^2 = \sum_{f\in\mathcal{T}_h} \frac{d_f}{(d+1)^2} (\delta_{E^+} v + \delta_{E^-} v)^2 \\ &\leq \sum_{f\in\mathcal{T}_h} \frac{2d_f}{(d+1)^2} \left[(\delta_{E^+} v)^2 + (\delta_{E^-} v)^2 \right] \leq \frac{2D\kappa_E}{(d+1)^2} \sum_{E\in\mathcal{T}_h} \omega_e (\delta_E v)^2 \\ &= c_J^2 (A_{\text{Lgr}} v,\, v),\end{aligned}$$

with D being a scaling constant changing from the weights d_f to ω_e (since both weights scale like τh^{d-2}, this scaling is independent of h and τ) and κ_E being an upper bound on the number of times an edge can be used on the sum over faces, $\kappa_E = \max_{E\in\mathcal{T}_h} 2\mathcal{N}_{i\cap j}$. Thus, this c_J is indeed independent of mesh size. □

Remark 5.4.1. Note that the estimate for κ_E is conservative. If $d > 2$, one can almost always use different edges E^+ and E^- to connect the points opposite a face f, so that in practice, κ_E can be made much smaller.

Lemma 5.4.2 (Continuity of the smoother)**.**

$$\exists\, c_{\widetilde{S}_R} > 0 : \quad (\widetilde{S}_R v, v) \leq c^2_{\widetilde{S}_R}(D_{\widetilde{S}} v,\, v), \quad \forall v \in S_h.$$

with $c_{\widetilde{S}_R}$ independent of mesh size.

Proof. Given $\widetilde{S}_R$ is SPD, using the standard Euclidean basis $\{e_i\}$ we use the Cauchy-Schwarz inequality to obtain

$$|\widetilde{S}_R^{ij}| = |(\widetilde{S}_R e_i, e_j)| = |(e_i, e_j)_{\widetilde{S}_R}| \leq \|e_i\|_{\widetilde{S}_R} \|e_j\|_{\widetilde{S}_R} = \sqrt{\widetilde{S}_R^{ii} \widetilde{S}_R^{jj}}.$$

Next, using this inequality we obtain

$$\left|[D_{\widetilde{S}}^{-1/2} \widetilde{S}_R D_{\widetilde{S}}^{-1/2}]_{ij}\right| = \frac{|a_{ij}|}{\sqrt{a_{ii} a_{jj}}} \leq 1. \tag{5.24}$$

Therefore, we have

$$\begin{aligned} c_{\widetilde{S}_R} &= \max_{v \in S_h} \frac{(\widetilde{S}_R v, v)}{(D_{\widetilde{S}} v, v)} = \max_{w = D_{\widetilde{S}}^{1/2} v \in S_h} \frac{(D_{\widetilde{S}}^{-1/2} \widetilde{S}_R D_{\widetilde{S}}^{-1/2} w, w)}{(w, w)} \\ &= \rho\left(D_{\widetilde{S}}^{-1/2} \widetilde{S}_R D_{\widetilde{S}}^{-1/2}\right) \leq \|D_{\widetilde{S}}^{-1/2} \widetilde{S}_R D_{\widetilde{S}}^{-1/2}\|_\infty. \end{aligned}$$

From the inequality (5.24) it follows that $\|D_{\widetilde{S}}^{-1/2} \widetilde{S}_R D_{\widetilde{S}}^{-1/2}\|_\infty$ is bounded by the number of nonzeroes per row in $\widetilde{S}_R$, which can be bounded by the number of faces an element has, $d + 1$. □

Lemma 5.4.3 (Stable splitting)**.** *For every $p \in S_h$, there exist $p_0 \in S_h$ and $w_J \in V_h$ such that $p = p_0 + \Pi_{VS} w_{VS}$ and*

$$(D_{\widetilde{S}} p_0,\, p_0) + (A_{Lgr} w_J,\, w_J) \leq c_0^2 ||p||^2_{\widetilde{S}_R}, \tag{5.25}$$

where $c_0 > 0$ should be small and independent of mesh size.

Proof. To bound the first term on the left hand side of (5.25), we set the value of $w_{VS} \in V_h$ at a vertex i to equal the average of the values of $p \in S_h$ on the simplices T which surround

the vertex i. More precisely,

$$V_h \ni w_{VS} = \sum_i [w_{VS}]_i \varphi_i, \quad [w_{VS}]_i = \frac{1}{\mathcal{N}_i} \sum_{T \in \Omega_i} p_T.$$

By the definition (5.22) and rearranging the decomposition of p, for each $T \in \mathcal{T}_h$

$$\begin{aligned}
[p_0]_T &= [p - \Pi_{VS} w_{VS}]_T = p_T - \frac{1}{d+1} \sum_{i \in T} [w_{VS}]_i \\
&= p_T - \frac{1}{d+1} \sum_{i \in T} \frac{1}{\mathcal{N}_i} \sum_{T' \in \Omega_i} p_{T'} \\
&= \frac{1}{d+1} \sum_{i \in T} \frac{1}{\mathcal{N}_i} \sum_{T' \in \Omega_i} (p_T - p_{T'}).
\end{aligned}$$

We now make the following estimate for this fixed T, using two applications of Cauchy-Schwarz:

$$\begin{aligned}
&\left(p_T - \frac{1}{d+1} \sum_{i \in T} \frac{1}{\mathcal{N}_i} \sum_{T' \in \Omega_i} p_{T'} \right)^2 \\
&= \left(\frac{1}{d+1} \sum_{i \in T} \frac{1}{\mathcal{N}_i} \sum_{T' \in \Omega_i} (p_T - p_{T'}) \right)^2 \\
&\leq \frac{1}{(d+1)^2 n_T^2} \left(\sum_{i \in T} \sum_{T' \in \Omega_i} (p_T - p_{T'}) \right)^2 \\
&\leq \frac{1}{(d+1) n_T} \sum_{i \in T} \sum_{T' \in \Omega_i} (p_T - p_{T'})^2.
\end{aligned}$$

with $n_T = \min_{i \in T} \mathcal{N}_i$. For each simplex $T' \in \Omega_i$, we expand the difference $(p_T - p_{T'})$ as a telescoping sum of differences across faces, $\sum_{f \in \mathcal{F}_i(T,T')} (p_{T+} - p_{T-})$. To define this set of faces $\mathcal{F}_i(T, T')$, we must first define a chain of pairwise–adjacent simplices $\{T_j\}_{j=1}^{J_i} \subset \Omega_i$ of minimal length, with $T_1 = T$ and $T_{J_i} = T'$. Then for this chain,

$$\mathcal{F}_i(T, T') = \{ f \in \mathcal{F}_i | f_j = T_{j+1} \cap T_j, \quad j = 1, \ldots, J_i - 1 \}.$$

Denoting $\mathcal{N}^f_{\mathcal{F}_i(T,T')}$ as the number of faces in $\mathcal{F}_i(T,T')$, we get the following:

$$\begin{aligned}
&\left(p_T - \frac{1}{d+1}\sum_{i\in T}\frac{1}{\mathcal{N}_i}\sum_{T'\in\Omega_i} p_{T'}\right)^2 \\
&\leq \frac{1}{(d+1)n_T}\sum_{i\in T}\sum_{T'\in\Omega_i}(p_T - p_{T'})^2 = \frac{1}{(d+1)n_T}\sum_{i\in T}\sum_{T'\in\Omega_i}\left(\sum_{f\in\mathcal{F}_i(T,T')}(p_{T+} - p_{T-})\right)^2 \\
&\leq \frac{1}{(d+1)n_T}\sum_{i\in T}\sum_{T'\in\Omega_i}\mathcal{N}^f_{\mathcal{F}_i(T,T')}\sum_{f\in\mathcal{F}_i(T,T')}(p_{T+} - p_{T-})^2. \qquad (5.26)
\end{aligned}$$

Since $\{T_j\}_{j=1}^{J_i} \subset \Omega_i$, each simplex chain length J_i is bounded from above by $\mathcal{N}_i/2$; otherwise a shorter chain of simplices connecting T and T' must exist.

Given this observation, we can reorder the two innermost sums in the last inequality above:

$$\sum_{T'\in\Omega_i}\mathcal{N}^f_{\mathcal{F}_i(T,T')}\sum_{f\in\mathcal{F}_i(T,T')}(p_{T+} - p_{T-})^2 \leq \sum_{f\in\mathcal{F}_i}(p_{T+} - p_{T-})^2\sum_{T'\in\Omega_i}\mathcal{N}^f_{\mathcal{F}_i(T,T')}$$

which gives us the bound

$$\sum_{T'\in\Omega_i}\mathcal{N}^f_{\mathcal{F}_i(T,T')}\sum_{f\in\mathcal{F}_i(T,T')}(p_{T+} - p_{T-})^2 \leq \frac{F(F+1)}{2}\sum_{f\in\mathcal{F}_i}(p_{T+} - p_{T-})^2, \qquad (5.27)$$

with $F = \max_{i\in\mathcal{T}_h}\max_{T,\,T'\in\Omega_i}\mathcal{N}^f_{\mathcal{F}_i(T,T')}$. As each $\mathcal{N}^f_{\mathcal{F}_i(T,T')} = J_i - 1$, taking $J = \max_{i\in\mathcal{T}_h} J_i$ it follows that $F \leq \max_{i\in\mathcal{T}_h}\frac{\mathcal{N}_i}{2} - 1$, which gives us a uniform estimate for F independent of mesh size.

Combining (5.26) and (5.27) with the observation that any face $f \in \mathcal{F}_i$, $i \in T$ is shared by at most d vertices in T, and that each face f is globally shared by at most 2 simplices in the mesh, we get the final set of inequalities,

$$\begin{aligned}
(D_{\widetilde{S}}p_0,\, p_0)^2 &= \sum_{T\in\mathcal{T}_h} d_T(p_0)_T^2 = \sum_{T\in\mathcal{T}_h} d_T(p - \Pi_J w_J)^2 \\
&\leq \frac{dF(F+1)D_*}{(d+1)n}\sum_{f\in\mathcal{T}_h} d_f(p_{T+} - p_{T-})^2 = c_D^2||p||^2_{\widetilde{S}_R},
\end{aligned}$$

with $n = \min_{T\in\mathcal{T}_h} n_T$ and $D_* = \max_{T,f\in\mathcal{T}}\frac{d_T}{d_f}$, thus giving us a constant independent of mesh size and τ due to the ratio of the weights d_T and d_f being of the same order in h and τ.

Now we need to bound the second half of the left hand side of (5.25). Taking the same

definition of w_{VS} as above, our goal is to show $(A_{\text{Lgr}} w_{VS}, w_{VS}) \leq c_A^2 ||p||^2_{\widetilde{S}}$. Since

$$(A_{\text{Lgr}} w_{VS}, w_{VS}) = \sum_{E \in \mathcal{T}_h} w_E \left(\frac{1}{\mathcal{N}_i} \sum_{T \in \Omega_i} p_T - \frac{1}{\mathcal{N}_j} \sum_{T' \in \Omega_j} p_{T'} \right)^2,$$

we will fix an edge $E \in \mathcal{T}_h$ to estimate each term of the sum first.

Note that for any constant C, $\left(\frac{1}{\mathcal{N}_i} \sum_{T \in \Omega_i} C - \frac{1}{\mathcal{N}_j} \sum_{T' \in \Omega_j} C \right)^2 = (C - C)^2 = 0$. Then we have,

$$\begin{aligned}
&\left(\frac{1}{\mathcal{N}_i} \sum_{T \in \Omega_i} p_T - \frac{1}{\mathcal{N}_j} \sum_{T' \in \Omega_j} p_{T'} \right)^2 = \left(\sum_{T \in \Omega_i \cup \Omega_j} (p_T - C) \left(\frac{\chi_{\Omega_i}(T)}{\mathcal{N}_i} - \frac{\chi_{\Omega_j}(T)}{\mathcal{N}_j} \right) \right)^2 \\
&\leq \sum_{T \in \Omega_i \cup \Omega_j} (p_T - C)^2 \sum_{T \in \Omega_i \cup \Omega_j} \left(\frac{\chi_{\Omega_i}(T)}{\mathcal{N}_i} - \frac{\chi_{\Omega_j}(T)}{\mathcal{N}_j} \right)^2 \\
&\leq \left(\sum_{T \in \Omega_i \cup \Omega_j} \left(\frac{\chi_{\Omega_i}(T)}{\mathcal{N}_i} \right)^2 + \sum_{T \in \Omega_i \cup \Omega_j} \left(\frac{\chi_{\Omega_j}(T)}{\mathcal{N}_j} \right)^2 \right) \inf_{C \in \mathbb{R}} ||\boldsymbol{p}_{\Omega_i \cup \Omega_j} - C\mathbb{1}||^2_{\ell^2},
\end{aligned}$$

where the first inequality is the Cauchy-Schwarz inequality and the second is due to both the Cauchy-Schwarz inequality and the non-negativity of the terms in the second summand. Here $\chi_{\Omega_k}(T)$ is the characteristic function on set Ω_k, the vector $\boldsymbol{p}_{\Omega_i \cup \Omega_j} = (p_{T_1}, \ldots, p_{T_{\mathcal{N}_{i \cup j}}})^T$ denotes the values of p on each simplex in $\Omega_i \cup \Omega_j$, and $\mathbb{1}$ is the vector of the same size as $\boldsymbol{p}_{\Omega_i \cup \Omega_j}$ with ones on each entry. It is well known that the C that will minimize the ℓ^2–norm in this scenario is the average of p over all the simplices in the union, $\bar{p} = \frac{1}{\mathcal{N}_{i \cup j}} \sum_{T \in \Omega_i \cup \Omega_j} p_T$. Using this fact and that both $\frac{\chi_{\Omega_k}(T)}{\mathcal{N}_k} \leq 1$ and $\sum_{T \in \Omega_i \cup \Omega_j} \frac{\chi_{\Omega_k}(T)}{\mathcal{N}_k} = 1$ for $k = i, j$, we can continue our estimates,

$$\begin{aligned}
&\left(\sum_{T \in \Omega_i \cup \Omega_j} \left(\frac{\chi_{\Omega_i}(T)}{\mathcal{N}_i} \right)^2 + \sum_{T \in \Omega_i \cup \Omega_j} \left(\frac{\chi_{\Omega_j}(T)}{\mathcal{N}_j} \right)^2 \right) \inf_{C \in \mathbb{R}} ||\boldsymbol{p}_{\Omega_i \cup \Omega_j} - C\mathbb{1}||^2_{\ell^2} \\
&\leq 2||\boldsymbol{p}_{\Omega_i \cup \Omega_j} - \bar{p}\mathbb{1}||^2_{\ell^2} \leq 2\mathcal{N}_{i \cup j} \text{Diam}(\Omega_i \cup \Omega_j)(Lp, p)_{\Omega_i \cup \Omega_j} \\
&= \gamma^2_{P,E} \sum_{f \in \Omega_i \cup \Omega_j} w_{f,L} (p_{T+} - p_{T-})^2.
\end{aligned}$$

The second inequality is due to the Poincaré inequality, with $||\nabla p||^2_{\ell^2}$ expressed as the bilinear form $(Lp, p)_{\Omega_i \cup \Omega_j}$, which is the local action of the graph Laplacian. The weights $w_{f,L}$ are the weights required to form the local Laplacian bilinear form across faces.

Finally, we incorporate the sum over all edges,

$$\begin{aligned}(A_{\text{Lgr}} w_J, w_J) &= \sum_{E \in \mathcal{T}_h} w_E (\delta_E w_J)^2 \\ &\leq \gamma_P^2 \sum_{E \in \mathcal{T}_h} \sum_{f \in \Omega_i \cup \Omega_j} w_{f,L} (p_{T+} - p_{T-})^2 \\ &\leq D \gamma_P^2 \sum_{f \in \mathcal{T}_h} d_f (p_{T+} - p_{T-})^2 = c_A^2 ||p||_{\widetilde{S}},\end{aligned}$$

where $\gamma_P = \max_{E \in \mathcal{T}_h} \gamma_{P,E}$, $D = d\,\alpha\,\widetilde{D}$, with $\alpha = \max_{i \in \mathcal{T}_h}\{\#\text{ vertices} \in \Omega_i\}$ and $\widetilde{D}$ is a scaling factor used to change from the weights $w_{f,L}$ to d_f, again independent of mesh size due to both weights being of the same order in h.

Taking the max of c_D and c_A as c_0 in (5.25) gives us the required uniform bound. □

As shown in [75] and [76], the last three Lemmas guarantee that our preconditioner is uniform, as Lemma 5.4.1 and 5.4.2 prove that our scheme fulfills the second assumption of Theorem 8, and Lemma 5.4.3 shows that the first assumption is satisfied. □

5.5 Numerical tests

To verify the robustness of the combined preconditioners for symmetric (5.2), we solved the linear system (4.54) for the first modified Picard iteration as outlined in (4.42) with an outer CG iteration that had a relative residual stopping criterion of less than 5×10^{-8}. We precondition this full system solve by the operator whose action is defined in 5.3.1. The inner solve of the $\widetilde{S}_R$ system (5.16) are done using CG, preconditioned with auxiliary space method described in the previous section. The inner iterations were stopped when the relative residual was smaller than 10^{-9}. To solve the auxiliary problem to machine precision, we used unsmoothed aggregation AMG, with 20 5-level V-cycles.

For numerical illustration of the performance of the preconditioner, we used an analytic solution $\Psi_e(x,t)$, $\Omega = [0,2]^3$, with the source term being determined analytically by plugging in the solution into Richards' equation. We used Dirichlet data for $\Psi(x,t)$ for all $t > 0$, $\Psi(x,t)|_{\partial\Omega} = \Psi_e(x,t)$, and initial condition $\Psi_0 = \Psi_e(x,0)$.

5.5.1 Example 1: Continuously varying K

For our first example, we chose $\Psi_e(x,t) = -10t|x|^2$,

$$\theta(\Psi) = \exp(\Psi), \quad K(\theta) = (K_{\max} - K_{\min})\theta + K_{\min},$$

with $K_{\min} = 1 \times 10^{-6}$ and $K_{\max} = 1$. Thus, K varies continuously by several orders of magnitude from the bottom of the cube to the top.

p	2	3	4	5	6
Outer/Inner	5/13	5/15	5/15	5/15	5/16

Table 5.1. Number of outer PCG/average inner PCG iterations (rounded to nearest integer) for solving the linearization of the mixed form of RE, using the analytic K and θ as described in example 1. Here the mesh size is $h = 2^{-p}$ and timestep $\tau = 1$.

Table 5.5.1 lists the number of outer PCG and average inner PCG iterations for the preconditioned linear solve of the first Modified Picard step for this problem. To give some perspective on the size of the respective mesh sizes used to test, for $h = 2^{-6}$, the size of the full system is over 4.5 million DOF, with over 1.5 million pressure unknowns; the size of the auxiliary system used to precondtion (5.16) in the Schur iteration is around 262,000 unknowns, which is a reduction in degrees of freedom of roughly 1/6. It should be noted that since the auxiliary system is as poorly conditioned as (5.16), a preconditioner can (and should in practice) be used to solve the auxiliary problem efficiently, as the performance of the auxiliary space preconditioner depends on solving the auxiliary problem to low tolerance.

5.5.2 Example 2: Van Genuchten-Mualem (VGM) model

For the second test related to the VGM model, we considered the same problem setup, but use the VGM K and θ as discussed in chapter 1. The test runs for values of α, n, and K_S for three different media: Beit Netofa clay ($\alpha = 0.152, n = 1.17, K_S = 8.2 \times 10^{-4}$), silt loam ($\alpha = 0.423, n = 2.06, K_S = 5 \times 10^{-2}$), and clay loam ($\alpha = 1.9, n = 1.31, K_S = 6.2 \times 10^{-2}$). The Beit Netofa clay and the silt loam examples are from in [13], and the clay loam example is from [45]. Due to the sharper gradients introduced by considering VGM parameters, we needed to decrease the time step size to $\tau = 1/32$ to ensure convergence of the linearization procedure to the solution at the next timestep.

Note that K reduces roughly 4 orders of magnitude from the bottom of the cube to the silt loam, roughly 8 orders for the Beit Netofa clay, and roughly 10 orders for the clay loam.

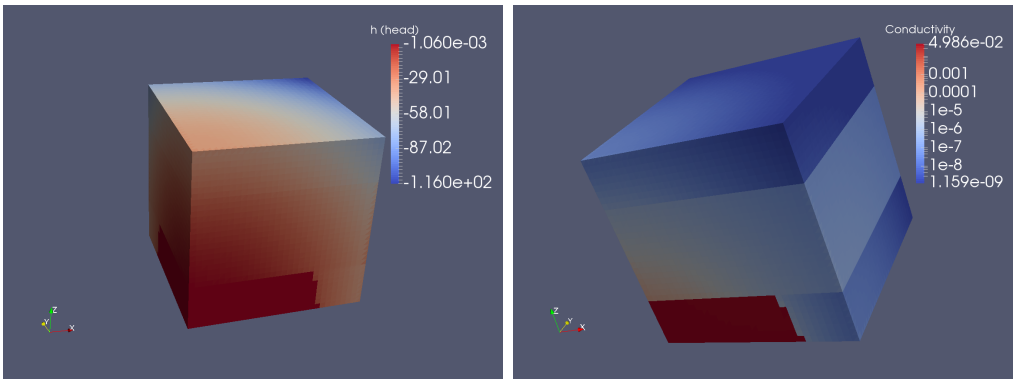

Figure 5.1. 3D profiles of Ψ (left) and K (right) for homogeneous boundary condition, with source term $f = 1$ for the layered VGM problem.

Due to the high contrast, we increased the number of V cycles to 60 for the coarse space solve.

p	2	3	4	5	6
Beit Netofa clay	8	12	15	15	15
Silt Loam	11	14	15	15	15
Clay Loam	9	13	16	17	17

Table 5.2. Average inner PCG iterations for Stilde solve for each of three different media after preconditioning with aux space preconditioner. Here the mesh size is $h = 2^{-p}$ and timestep $\tau = 1/32$.

As table 5.5.2 shows, even for high contrast K, our preconditioner maintains its robustness.

5.5.3 Example 3: VGM Layered media test

Finally, to measure the effectiveness of this method for more complex simulations, we wanted to run an unsaturated test with layered media, akin to the layered media example of Lehmann and Ackerer [44]. Using the same setup and boundary/initial conditions and same analytic solution, we split the domain into three layers, $\Omega_1 = [0,2] \times [0,2] \times [0, \frac{1}{2}]$, $\Omega_2 = [0,2] \times [0,2] \times [\frac{1}{2}, \frac{3}{2}]$, $\Omega_3 = [0,2] \times [0,2] \times [\frac{3}{2}, 2]$ In $\Omega_1 \cup \Omega_3$, we used the VGM parameters for silt loam, and in Ω_2 we used that for Beit Netofa clay; see figures 5.1 and 5.2 for views of the layered problem. A particular point of interest is how quickly the conductivity changes as a function of pressure head. This combined with the discontinuity at the boundary interfaces make this problem very difficult to solve numerically.

as the results show in table 5.5.3, despite discontinuities at both interfaces, the preconditioner is still asymptotically uniform, as is typical for HX preconditioners. Note however, that

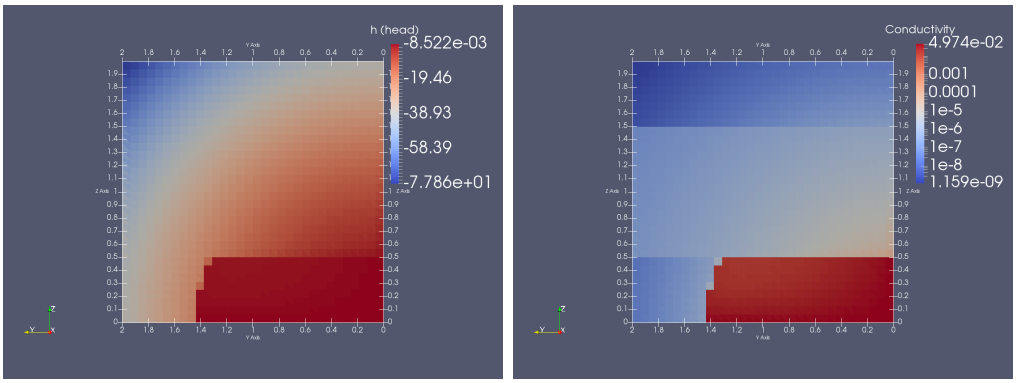

Figure 5.2. Vertical slice (with normal e_x) profiles of Ψ (left) and K (right) for homogeneous boundary condition, with source term $f = 1$ for the layered VGM problem.

we do Not incorporate transmission conditions when solving these systems, as our focus in this test was merely preconditioning the linear systems; such terms correspond to manipulations of the right hand side, and should not affect the structure of the linear systems that need to be solved; for some consideration of transmission conditions for layered media with Richards' equation, see [51], and [52].

m	2	3	4	5	6
$\widetilde{S}_R$ PCG iterations	10	13	16	17	18

Table 5.3. Average number of inner PCG iterations for exact solving of $\widetilde{S}_R$ preconditioned by aux space exact solve (first mod Picard iteration). Here the characteristic mesh size is $h = 2^{-m}$.

5.5.4 Auxiliary space preconditioning on nonsymmetric linearizations

In this case of nonsymmetric linearizations, the standard theory of auxiliary spaces cannot be applied, as the bilinear forms considered are no longer symmetric, i.e, they do not define inner products on the spaces on which they are defined, and the ficititious space Lemma can no longer be applied. Moreover, as noted above, when K gradients are steep, the linear convection-diffusion problem that needs to be solved will have areas where the convection term dominates the diffusion and vice-versa, which implies that Jacobi or Gauss-Seidel smoothers will not be able to capture the high frequency errors. However, heuristically speaking, as both (5.1) and (5.2) are nonsymmetric and are simply different finite element discretizations of the same nonsymmetric weak problem (3.8), we wanted to investigate the performance of a nonsymmetric variant of our Auxiliary space preconditioner combined with the nonlinear variant of the Schur iteration 5.3.1 on the full Newton linearization, in which the B_{div}^T term is

replaced with $B_{\text{div}}^T - B_{K'}$, for problems with mild convection terms (i.e, small K gradients).

To numerically test this, we used the same setup as our first example using the Newton-Raphson linearization, with $K_{\min} = 10^{-2}$, and a small timestep $\tau = 1/128$ to ensure convergence of the linear scheme. Due to the inner system being nonsymmetric, we switched to GMRES for the $\widetilde{S}_R$ solve.

m	2	3	4	5	6
$\widetilde{S}_R$ GMRES iterations	14	16	17	18	18
Outer GMRES iters	10	9	11	10	8

Table 5.4. Average number of inner GMRES and outer GMRES iterations for the first full Newton iteration, with exact solving of $\widetilde{S}_R$ preconditioned by aux space exact solve. Here the characteristic mesh size is $h = 2^{-m}$.

Multigrid-based preconditioning techniques have been applied in the literature for Richards equation; in particular, Woodward and Jones [91] compare two different multigrid algorithms, with one using a pointwise red/black Gauss-Seidel algorithm, and another using a more expensive plane-smoother that takes anisotropies into account, however both preconditioners use a symmetric approximation by ignoring the nonsymmetric contributions to the Jacobian. Jenkins et. al. [92] consider the full Newton linearization and solve each Newton step using BiCG-stab, a nonsymmetric variant of Conjugate Gradient, preconditioned with a two-level Schwarz domain decomposition method, in which they use aggregation based domain decomposition to split the domain into subdomains with minimal overlap, transfer errors to each subspace, restrict and solve and approximately solve the problem on each sumdomain with BiCG-stab, and interpolate back. Their method scales mildly with mesh size, but isn't uniform.

As can be seen, for relatively small K gradients, the preconditioner still works, though we verified experimentally that allowing $K_{\min}$ to be any smaller would result in the number of iterations increasing more sharply with mesh size decrease.

Conclusions

In this thesis, we gave a comprehensive review and showed new results on the mathematical and numerical study of Richards' equation. On the analytic side, we were able to give a set of conditions that guarantee the convergence of a Picard linearization on the continuous level, which lends some insight into when similar linearization techniques may fail. We also introduced a preconditioner for the linear systems resulting from a perfectly mass conservative discretization that is adaptible to a wide class of meshes, with the intent being application to domains with complex geometry and unstructured meshes. We were able to prove the robustness of this preconditioner with respect to discretization parameters for popular symmetric linearizations used in the field to simulate unsaturated groundwater flow, and verified this uniformity for various models of the physical parameters K and θ, including the popular VGM model. We were able to show numerically similar behavior for a particular example using the nonsymmetric Newton-Raphson linearization.

There are many avenues for extension of the results in this work. The first is in extending the class of K in which our proposed preconditioner retains robustness for the Newton-Raphson linearization. One of the main challenges that this preconditioner faces is in the choice of smoother used for the auxiliary space preconditioner. In the context of convection-diffusion equations, particularly if the Jacobian ∂K becomes zero in certain areas of the domain and very large in others, a smoother that is capable of robustly smoothing out error is hard to define. This is primarily due to the fact that in the regions of the domain where the diffusion term dominates, a local averaging smoother like Jacobi will be effective in eliminating high frequency errors, whereas in the parts of the domain where convection dominates, a sweeping relaxation scheme in the direction of convection, such as Gauss-Seidel with cross-wind block (Tarjan) ordering [93, 94], will effectively eliminate high frequency errors. This choice of good smoother is made even more difficult by the fact that for the nonlinear problems, these regions of convection versus diffusion dominance may change on each iteration. For linear problems, Kim et al. [95] propose a robust multigrid preconditioner for the GMRES iterative solver given the typical $P1$-Lagrange discretization of convection-diffusion problems of the form

$$Au = f.$$

The action of multilevel preconditioner on operator A_k, $B_k : r \to z$, is defined by first applying an ordered Gauss-Seidel smoothing step some fixed (small) number of times using the normal equation on the finest grid,

$$A^T Au = A^T f,$$

Then inputting the smoothed output to a V-cycle that pre and post smooths using some fixed small number of Gauss-Seidel iterations with special ordering of unknowns, and then finally solving on the coarsest grid. Using this method, they were able to show numerically that GMRES preconditioned with this special multigrid preconditioner is uniform and robust with respect to the coarsest grid size for variably convection-dominated problems.

Additionally, the use of EAFE as a stand-alone discretization for Richards' equation is worth a more in-depth investigation, particularly due to the automatic upwinding it provides. This in effect generalizes the results of manual upwinding schemes for finite element discretizations, such as those in [27], and has been shown to be equivalent to special finite volume discretizations [96]. In this perspective, EAFE can be seen as a good candidate in the design of robust monotonicity-preserving numerical schemes, particularly for mixed unsaturated-saturated simulations, as other monotonicity preserving schemes tend to use special finite volume or finite difference approximations whose setup is highly dependent on mesh configuration, and are also more difficult to analyze than EAFE.

Finally, another research direction is developing rigorous, consistent criteria under which higher order linearizations of Richards' equation are guaranteed to converge. In our work, we showed some conditions under which convergence was guaranteed for the lowest order linearization. This result, while true for specific forms of K and θ, matches well with numerical experiments in the literature [11], where one finds examples with divergent Picard iteration and convergent higher order methods, like Newton-Raphson. This result can also lend credence to the fact that the Picard iteration is seldom used due to its inability to converge for strongly nonlinear K and θ, which are emblematic of most physically realistic standard models of these parameters. A continuation of this analysis would focus on determining conditions on K, θ and initial guesses under which the higher order modified Picard linearization would converge, and more importantly, information on how these parameters affect the rate of convergence. Such criteria remain elusive to the community in general [38], with the exception of the works of [14] and [13], who themselves work with the L-scheme, which is a simplified relaxation method. Such an investigation might lead to the development of reliable selection and switching criteria for modern numerical codes used by federal and private entities to simulate groundwater flow, which could increase the efficiency of these codes significantly.

Bibliography

[1] Davis, T. A. and Y. Hu (2011) "The University of Florida sparse matrix collection," *ACM Trans. Math. Software*, **38**(1), pp. Art. 1, 25.
URL http://dx.doi.org/10.1145/2049662.2049663

[2] van Genuchten, M., F. Leij, and S. Yates (1991) "The RETC code for quantifying the hydraulic functions of unsaturated soils," *Tech. Rep. IAG-DW12933934*.

[3] Nordbotten, J. and C. M.A. (2012) *Geological Sotrage of CO2: Modeling approaches for Large-Scale Simulation*, Wiley.

[4] Forsyth, P. (1995) "Simulation of nonaqueous phase groundwater contamination," *Advances in Water Resources*, **18**, pp. 74 – 83.

[5] Arbogast, T., M. F. Wheeler, and N.-Y. Zhang (1996) "A nonlinear mixed finite element method for a degenerate parabolic equation arising in flow in porous media," *SIAM J. Numer. Anal.*, **33**(4), pp. 1669–1687.
URL http://dx.doi.org/10.1137/S0036142994266728

[6] Arbogast, T., M. Obeyesekere, and M. F. Wheeler (1993) "Numerical methods for the simulation of flow in root-soil systems," *SIAM J. Numer. Anal.*, **30**(6), pp. 1677–1702.
URL http://dx.doi.org/10.1137/0730086

[7] Velten, K. (2009) *Mathematical Modeling and Simulation: Introduction for Scientists and Engineers*, Wiley.

[8] Forsyth, P. (1988) "Comparison of the single-Phase and two-Phase numerical model formulation for saturated-unsaturated groundwater flow," *Com. Meth. Appl. Mech. Eng.*, **69**, pp. 243 – 259.

[9] Alt, H. W. and S. Luckhaus (1983) "Quasilinear elliptic-parabolic differential equations," *Math. Z.*, **183**(3), pp. 311–341.
URL http://dx.doi.org/10.1007/BF01176474

[10] Otto, F. (1996) "L1-Contraction and Uniqueness for Quasilinear Elliptic-Parabolic Equations," *Journal of Diff. Equ.*, **131**(0155), pp. 20–38.

[11] Paniconi, C. and M. Putti (1994) "A comparison of Picard and Newton iteration in the numerical solution of multidimensional variably saturated flow problems," *Water Resour. Res.*, **30**(12), pp. 3357–3374.

[12] Celia, M. A., E. T. Bouloutas, and R. L. Zarba (1990) "A general mass-conservative numerical solution for the unsaturated flow equation," *Water Resources Research*, **26**(7), pp. 1483–1496.
URL http://dx.doi.org/10.1029/WR026i007p01483

[13] List, F. and F. A. Radu (2016) "A study on iterative methods for solving Richards' equation," *Comput. Geosci.*, **20**(2), pp. 341–353.
URL http://dx.doi.org/10.1007/s10596-016-9566-3

[14] Slodička, M. (2002) "A robust and efficient linearization scheme for doubly nonlinear and degenerate parabolic problems arising in flow in porous media," *SIAM J. Sci. Comput.*, **23**(5), pp. 1593–1614.

[15] Pop, I. S., F. Radu, and P. Knabner (2004) "Mixed finite elements for the Richards' equation: linearization procedure," *J. Comput. Appl. Math.*, **168**(1-2), pp. 365–373.
URL http://dx.doi.org/10.1016/j.cam.2003.04.008

[16] Radu, F., I. S. Pop, and P. Knabner (2004) "Order of convergence estimates for an Euler implicit, mixed finite element discretization of Richards' equation," *SIAM J. Numer. Anal.*, **42**(4), pp. 1452–1478.
URL http://dx.doi.org/10.1137/S0036142902405229

[17] Bergamaschi, L. and M. Putti (1999) "Mixed finite elements and Newton-type linearizations for the solution of Richards' equation," *Internat. J. Numer. Methods Engrg.*, **45**(8), pp. 1025–1046.
URL http://dx.doi.org/10.1002/(SICI)1097-0207(19990720)45:8<1025::AID-NME615>3.3.CO;2-7

[18] Radu, F., I. Pop, and P. Knabner (2006) "On the convergence of the Newton method for the mixed finite element discretization of a class of degenerate parabolic equation," *Numerical Mathematics and Advanced Applications, A. Bermudez de Castro et al.(editors), Springer*, pp. 1194–1200.

[19] Xu, J. and L. Zikatanov (1999) "A monotone finite element scheme for convection-diffusion equations," *Math. Comp.*, **68**(228), pp. 1429–1446.
URL https://doi.org/10.1090/S0025-5718-99-01148-5

[20] Radu, F. A., I. S. Pop, and P. Knabner (2008) "Error estimates for a mixed finite element discretization of some degenerate parabolic equations," *Numer. Math.*, **109**(2), pp. 285–311.
URL http://dx.doi.org/10.1007/s00211-008-0139-9

[21] Van Genuchten, M. T. (1980) "A closed-form equation for predicting the hydraulic conductivity of unsaturated soils," *Soil science society of America journal*, **44**(5), pp. 892–898.

[22] Morel-Seytoux, H. and J. Billica (1985) "A two phase numerical model for prediction of infiltration: applications to a semi-infinite column," *Water resour. Res.*, **21**, pp. 607–615.

[23] Green, D., H. Dabiri, and C. Weinaug (1970) "Numerical modelling of unsaturated groundwater flow and comparison of the model to a field experiment," *Water Resour. Res.*, **6**, pp. 862–874.

[24] Forsyth, P., Y. Wu, and K. Pruess (1995) "Robust numerical methods for saturated-unsaturated flow with dry initial conditions in heterogeneous media," *Advances in Water Resources*, **18**(1), pp. 25 – 38.
URL `http://www.sciencedirect.com/science/article/pii/030917089500020J`

[25] Wu, Y., J. Kool, and J. McCord (1992) "An evaluation of alternative numerical formulations for two-phase air water flow in unsaturated soils," in *Proceedings of the American Geophysical Union Spring meeting, Montreal*.

[26] Rubin, B. and P. Sammon (1983) "Practical control of timestep selection in thermal simulation," in *Proc. of 1983 Society of Petroleum Engineers Reservoir Simulation Symposium*.

[27] Forsyth, P. and M. Kropinski (1997) "Monotonicity considerations for saturated-unsaturated subsurface flow," *SIAM J. Sci. Comput.*, **18**, pp. 1328–1354.

[28] Rogers, C., M. Stallybrass, and D. Clements (1983) "On two phase filtration under gravity and with boundary infiltration: Application of a Backlund transformation." *Nonlin. Anal. Theory Meth. Appl.*, **7**, pp. 785–799.

[29] Broadbridge, P. and I. White (1988) "Modelling solute transport, chemical adsorption and cation exchange," in *Int. Hydrology and Water Resources Symp. Nat. Conf. Publ.*, pp. 924–929.

[30] Santer, G. e. (1988) "Exact nonlinear solution for constant flux infiltration," *Jour. Hydrology*, **97**, pp. 341–346.

[31] Miller, C. T. and et al. (1998) "Multiphase flow and transport modeling in heterogeneous porous media: Challenges and approaches," *Adv. Water Resour.*, **21**, pp. 77–120.

[32] Miller, C. T., C. Kelley, and M. D. Tocci (1997) "Accurate and economical solution of the pressure-head form of Richards' equation by the method of lines," *Adv. Water Resour.*, **20**, pp. 1–14.

[33] Pop, I. (2002) "Error estimates for a time discretization method for the Richards equation," *Comp. Geos.*, **6**, pp. 141–160.

[34] Assouline, S. (2002) "Infiltration into soils: Conceptual approaches and solutions," *Water Resour. Res.*, **6**, pp. 1755–1772.

[35] Averyanov, S. (1950) "About permeability of subsurface soils in case of incomplete saturation," *Engineering Collection*, **7**.

[36] Buckingham, E. (1907) *Studies on the Movement of Soil Moisture*, *Tech. Rep. 38*.

[37] Childs, E. and N. Collis-George (1950) "The permeability of porous materials," *Proc. Roy. Sov. Ser.*, **201**, pp. 392–405.

[38] Farthing, M. W. and F. L. Ogden (2017) "Numerical Solution of Richards' Equation: A review of Advances and Challenges," *Soil Sci. Soc. of Am. J.*, **81**, pp. 1257–1269.

[39] Kozeny, J. (1927) "Ueber kapillare Leitung des Wassers im Boden," *Sitzungsber Akad. Wien.*, **136**, pp. 271–306.

[40] Burdine, N. T. (1953) "Relative permeability calculation from size distribution data," *Trans. AIME*, **198**, pp. 71–78.

[41] Mualem, Y. (1976) "A new model for predicting the hydraulic conductivity of unsaturated porous media," *Water Resour. Res.*, **12**, pp. 513–522.

[42] Brooks, R. and A. Corey (1966) "Properties or porous media affecting fluid flow," *J. Irrig. Drain. Div. Am. Soc. Civil Eng.*, **92**, pp. 61–88.

[43] Celia, M., A. Lajpat, and G. Pinder (1987) "Orthogonal collocation and alternating-direction procedures for unsaturated flow problems," *Adv. Water Resources*, **10**, pp. 178–187.

[44] Lehmann, F. and P. Ackerer (1998) "Comparison of Iterative Methods for Improved Solutions of the Fluid Flow Equation in Partially Saturated Porous Media," *Transport in Porous Media*, **31**(3), pp. 275–292.
URL `http://dx.doi.org/10.1023/A:1006555107450`

[45] Miller, C. T., G. A. Williams, C. Kelley, and M. D. Tocci (1998) "Robust solution of Richards' equation for nonuniform porous media," *Water Resour. Res.*, **34**, pp. 2599–2610.

[46] Evans, L. C. (2002) *Partial Differential Equations*, American Mathematical Society.

[47] Kirchhoff, G. (1883) "Zur Theorie der Lichtstrahlen," *Ann. d. Physik.*, **18**, pp. 663–695.

[48] Berninger, e. a., H.B. (2014) "A multidomain discretization of the Richards equation in layered soil," *Comput. Geosci.*, **19**, pp. 213–232.

[49] Brindt, N. and R. Wallach (2017) "Thr moving-boundary approach for modeling gravity-driven stable and unstable flow in soils," *Water Resourc. Res.*, **53**, pp. 344–350.

[50] Kružkov, S. (1970) "First Order Quasilinear Equations in Several Independent Variables," *USSR Sb. 10 217*, **12**(2), pp. 217–241.

[51] Hills, R., I. Porro, D. Hudson, and P. Wierenga (1989) "Modeling One-Dimensional Infiltration Into Very Dry Soils 1. Model Development and Evaluation," *Water Resour. Res.*, **25**(6), pp. 1259–1269.

[52] Hills, R., M. Kirkland, and P. Wierenga (1992) "Algorithms for solving Richards' equation for variably saturated soils," *Water Resour. Res.*, **28**(8), pp. 2049–2058.

[53] Simunek, J., M. van Genuchten, and M. Sejna (2008) "Development and applications of the HYDRUS and STANMOD software packages and related codes," *Vadose Zone J.*, **7**, pp. 587–600.

[54] Lott, P., H. Walker, C. Woodward, and U. Yang (2012) "An accelerated Picard method for nonlinear systems related to variably saturated flow," *Advances in Water Resources*, **38**, pp. 92 – 101.
URL `http://www.sciencedirect.com/science/article/pii/S0309170811002569`

[55] Walker, H. and P. Ni (2011) "Anderson acceleration for fixed-point iterations," *SIAM J. Num. Anal.*, **49**, pp. 1715 – 1735.

[56] Nédélec, J.-C. (1980) "Mixed finite elements in $\mathbf{R}^3$," *Numer. Math.*, **35**(3), pp. 315–341.
URL `http://dx.doi.org/10.1007/BF01396415`

[57] Ciarlet, P. (1978) *The Finite Element Method for Elliptic Problems*, Oxford, New York.

[58] Knabner, P. (1987) *Finite Element Simulation of Saturated-Unsaturated Flow Through Porous Media*, Birkhäuser Boston, Boston, MA, pp. 83–93.
URL `http://dx.doi.org/10.1007/978-1-4684-6754-3_6`

[59] Draganescu, A., T. Dupont, and L. Ridgeway Scott (2004) "Failure of the Discrete Maximum Principle for an elliptic finite element problem," *Math. Comput.*, **74**(249), pp. 1–23.

[60] Barth, T. (1992) *Aspects of unstructured grids and finite-volume solvers for the Euler and Navier-Stokes equations*, *Tech. Rep. 787*, AGARD, special course on unstructured grids methods for advection dominated flows.

[61] Zaidel, J. and D. Russo (1992) "Estimation of finite difference interblock conductivities for simulation of infiltration into initially dry soils," *Water Resour. Res.*, **28**(9), pp. 2285–2295.

[62] Markowich, P. and M. Zlamal (1989) "Inverse-average-type finite element discretizations of self-adjoint second order elliptic problems," *Math. Comp.*, **51**, pp. 431–449.

[63] Brezzi, F., L. Marini, and P. Pietra (1989) "2-dimensional exponential fitting and applications to drift-diffusion models," *SIAM J. Numer. Anal.*, **26**(6), pp. 1342–1355.

[64] Putti, M. and C. Cordes (1998) "Finite element approximation of the diffusion operator on tetrahedra," *SIAM J. Sci. Comput.*, **19**(4), pp. 1154–1168.

[65] Aizinger, V., C. Dawson, B. Cockburn, and P. Castillo (2000) "The local discontinuous Galerkin method for contaminant transport," *Advances in Water Resources*, **24**(1), pp. 73 – 87.
URL `http://www.sciencedirect.com/science/article/pii/S0309170800000221`

[66] Brezzi, F. and M. Fortin (1991) *Mixed and Hybrid Finite Element Methods*, Springer:Berlin.

[67] Saad, Y. (2003) *Iterative methods for Sparse Linear Systems*, SIAM.

[68] Saad, Y. and M. Schultz (1986) "GMRES: A Generalized Minimal Residual Algorithm for solving Nonsymmetric Linear Systems," *SIAM J. Sci. Stat. Comput.*, **7**(3), pp. 856–869.

[69] Brezzi, F. and M. Fortin (1991) *Mixed and hybrid finite element methods*, Springer, New York.

[70] Brandt, A., S. F. McCormick, and J. W. Ruge (1982) *Algebraic Multigrid (AMG) for Automatic Multigrid Solution With Application To Geodetic Computations*, *Tech. Rep.*, Institute for Computational Studies, Colorado State University.

[71] Brandt, A., S. McCormick, and J. Ruge (1985) "Algebraic multigrid (AMG) for sparse matrix equations," in *Sparsity and its applications (Loughborough, 1983)*, Cambridge Univ. Press, Cambridge, pp. 257–284.

[72] Bramble, J. H. (1993) *Multigrid methods*, vol. 294 of *Pitman Research Notes in Mathematics Series*, Longman Scientific & Technical, Harlow.

[73] Kim, H., J. Xu, and L. Zikatanov (2003) "A multigrid method based on graph matching for convection diffusion equations," *Numerical linear algebra with applications*.
URL `http://onlinelibrary.wiley.com/doi/10.1002/nla.317/abstract`

[74] Nepomnyaschikh, S. (1992) "Decomposition and fictitious domain methods for elliptic boundary value problems," in *Proceedings of the Fifth International Symposium on Domain Decomposition Methods for Partial Differential Equations* (D. Keyes, T. Chan, G. Meurant, J. Scroggs, and R. Voigt, eds.), SIAM, pp. 62–72.

[75] Hiptmair, R. and J. Xu (2007) "Nodal Auxiliary Space Preconditioning in H(curl) and H(div) Spaces," *SIAM J. Numer. Anal.*, **45**, pp. 2483–2509.

[76] Xu, J. (1996) "The auxiliary space method and optimal multigrid preconditioning techniques for unstructured grids," *Computing*, **56**, pp. 215–235.

[77] Xu, J. and L. Zikatanov (2002) "The method of alternating projections and the method of subspace corrections in Hilbert space," *J. Amer. Math. Soc.*, **15**(3), pp. 573–597.
URL `http://dx.doi.org/10.1090/S0894-0347-02-00398-3`

[78] Xu, J. (1992) "Iterative methods by space decomposition and subspace correction," *SIAM Rev.*, **34**(4), pp. 581–613.
URL `http://dx.doi.org/10.1137/1034116`

[79] Tuminaro, R. S., J. Xu, and Y. Zhu (2009) "Auxiliary Space Preconditioners for Mixed Finite Element Methods," in *Domain Decomposition Methods in Science and Engineering XVIII* (M. Bercovier, M. J. Gander, R. Kornhuber, and O. Widlund, eds.), Lecture Notes in Computational Science and Engineering, Springer Berlin Heidelberg, pp. 99–109.

[80] Budiša, A. and X. Hu (2019) "Block Preconditioners for Mixed-dimensional Discretization of Flow in Fractured Porous Media," *arXiv preprint arXiv:1905.13513*.

[81] Adler, J., X. Hu, and L. Zikatanov (2017) "Robust Solvers for Maxwell's Equations with Dissipative Boundary Conditions," *SIAM Journal on Scientific Computing*, **39**(5), pp. S3–S23.

[82] Adler, J. H., F. J. Gaspar, X. Hu, C. Rodrigo, and L. T. Zikatanov (2018) "Robust Block Preconditioners for Biot's Model," in *Lecture Notes in Computational Science and Engineering*, vol. 125, Springer, pp. 3–16, `1705.08842`.

[83] Elman, H. and G. Golub (1994) "Inexact and Preconditioned Uzawa Algorithms for Saddle Point Problems," *SIAM J. Numer. Anal.*, **31(6)**, pp. 1645–1661.

[84] Rusten, T., P. Vassilevski, and R. Winther (1996) "Interior Penalty Preconditioners for mixed einite element approximations of elliptic problems," *Math. of Comput.*, **65(214)**, pp. 447–466.

[85] Batista, J., X. Hu, and L. T. Zikatanov (2019) "Auxiliary space preconditioners for mixed finite element discretizations of Richards' equation," *Comp. and Math. with Appl.*

[86] Hiptmair, R. (2002) "Finite elements in computational electromagnetism," *Acta Numer.*, **11**, pp. 237–339.

[87] Young, D. M. (1971) *Iterative solution of large linear systems*, Academic Press, New York-London.

[88] Brezzi, F., M. Fortin, and L. Marini (2006) "Error analysis of piecewise constant pressure approximations of Darcy's law," *Comp. Meth. in Appl. Mech. and Engr.*, **195**, pp. 1547–1559.

[89] Xu, J. and L. Zikatanov (2017) "Algebraic multigrid methods," *Acta Numer.*, **26**, pp. 591–721.
URL `https://doi.org/10.1017/S0962492917000083`

[90] Bern, M. and D. Eppstein (1992) "Mesh generation and optimal triangulation," *Computing in Euclidean Geometry*, **World Scientific**, pp. 23–90.

[91] Woodward, C. S. and J. Jones (2001) "Newton–Krylov–multigrid solvers for large–scale, highly heterogeneous, variably saturated flow problems," *Adv. Water Resourc.*, **24**, pp. 763–774.

[92] Jenkins, E., C. Kees, C. Kelley, and C. Miller (2001) "An Aggregation–Based Domain Decomposition Preconditioner for Groundwater Flow," *SIAM J. Sci. Comput.*, **23**(2), pp. 430–441.

[93] Wang, F. and J. Xu (1999) "A cross-wind block iterative method for convection-dominated problems," *SIAM J. Comput.*, **21**(2), pp. 620–645.

[94] Tarjan, R. E. (1972) "Depth-first seach and linear graph algorithms," *SIAM J. Comput.*, **1**, pp. 146–160.

[95] Kim, H., J. Xu, and L. Zikatanov (2004) "Uniformly convergent multigrid methods for convection-diffusion problems without any constraint on coarse grids," *Advances in Computational Mathematics*, **20**, pp. 385–399.

[96] Bank, R., P. Vassilevski, and L. Zikatanov (2017) "Arbitrary dimension convection-diffusion schemes for space–time discretizations," *Jour. of Comp. and Appl. Math.*, **310**, pp. 19–31.

Vita

Juan Batista

148 Longmeadow Lane, State College, PA 16803 email: jxb3641@gmail.com
office: +1 814 863 9126 cell: +1 347 724 0009

Education

- **PhD in Mathematics** 09/2015 - 12/2019
  Department of Mathematics, Eberly College of Science
  Pennsylvania State University, University Park, PA, USA
- **M.A. in Mathematics** 09/2013 - 05/2015
  Department of Mathematics, Eberly College of Science
  Pennsylvania State University, University Park, PA, USA
- **BS in Applied Mathematics** 09/2008 - 05/2012
  Department of Mathematical Sciences
  Rochester Institute of Technology, Rochester, NY

Theses and yearly projects

- *Auxiliary space preconditioning for mixed finite element discretizations of Richards' equation*
  Published, Journal of CAMWA 9/2019
  Authors: Juan Batista, Xiaozhe Hu (Tufts University), Ludmil Zikatanov (PSU)
- *Solving a pressure-perturbed finite element discretization of Biot's consolidation model in two and three dimensions using MFEM* 8/2017
  Summer internship project for MSGI 2017 advisor Dr. Bin Zheng (PNNL)
- *Computational schemes for one dimensional p-Laplacian* 6/2014
  M.A. paper, advisor Dr. Ludmil Zikatanov
- *Subway population flow model* 5/2011
  2011-2012 yearly project, advisor Dr. David Ross

Current research direction

Models of groundwater flow using Richards' equation, robust preconditioners for sparse linear systems

Research Interests

Methods of computation for PDE; multigrid; mixed Finite Elements; Auxiliary space preconditioning; Groundwater flow; contamination problems.

Programming Experience

Mathematical packages: MatLab, Maple, MFEM, HAZMATH (4 yrs), Java, Python (1 yr), C++ (1 yr), C (2 yrs), SQL (6 mo)