Freie Universität Berlin
Fachbereich Mathematik und Informatik
Institut für Mathematik

Solvers for Saddle Point Problems Arising from Finite Element Discretizations of the Darcy Equations

Master Thesis

Selin Saydan

supervised by
Prof. Dr. Volker John
Dr. Alfonso Caiazzo

Berlin, December 4, 2019

Eigenständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ohne unerlaubte fremde Hilfe und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

Ort, Datum Unterschrift

Contents

1 Introduction

The Darcy equations describe the behavior of fluids in porous media. They consist of a system of partial differential equations arising form Darcy's law for the flow of a fluid in a porous medium and the law of conservation of mass.

Boundary value problems for the Darcy equations can be used to model e.g. groundwater contamination and other problems of practical importance in civil, geotechnical and petroleum engineering involving fluid flow through porous media.

It is usually not possible to find an analytic solution to boundary value problems for the Darcy equations, such that numerical methods have to be employed for approximating the solution. (Mixed) finite element methods are a popular technique to discretize the equations in order to obtain an approximate solution. They belong to the class of numerical methods based on a variational formulation of the problem in which pressure and velocity are variables.

Discretization of the Darcy equations with finite element methods results in algebraic linear systems of saddle point type, that are large and sparse. Due to their indefiniteness and often poor spectral properties, such linear systems are challenging to solve. The efficient solution of these systems leads to the overall efficient simulation of flow problems in porous media using finite element methods.

The purpose of this thesis is to study solution methods for the arising saddle point problems, with an emphasis on iterative methods for large and sparse problems.

In Chapter 2 the mathematical foundations for the Darcy equations and the corresponding saddle point problems are introduced. Chapter 3 discusses the choice of appropriate finite element spaces that are used to approximate the two variables.

The subsequent chapter presents basic algebraic properties of the saddle point matrices and strategies for preconditioning of the saddle point system arising from the Darcy equations, namely the Least Squares Commutator (LSC) preconditioner, a block preconditioner based on the Schur complement approximation, the Vanka preconditioner, which can be considered as a block Gauss-Seidel method and the incomplete LU factorization.

The numerical results can be found in Chapter 5. Finally a summary of results is presented.

2 Darcy Equations

This chapter introduces the Darcy equations and describes the derivation of the weak formulation. Subsequently existence and uniqueness of a weak solution and the finite element discretization are discussed. The presentation of this chapter mostly follows [2], [3] and [5].

Consider the following system of first-oder partial differential equations for the flow of a fluid through a porous medium

$$\mathbb{K}\mathbf{u} + \nabla p = 0 \quad \text{in } \Omega \tag{2.1}$$

$$\nabla \cdot \mathbf{u} = f \quad \text{in } \Omega \tag{2.2}$$

where

$\Omega \subset \mathbb{R}^n$ is a bounded connected flow domain, which is a porous medium saturated with a fluid, with Lipschitz boundary $\partial\Omega$

$\mathbf{u} : \Omega \to \mathbb{R}^n$ is the fluid velocity

$p : \Omega \to \mathbb{R}$ is the fluid pressure

$f \in L^2(\Omega)$ is a given source term that represents the density of potential sources (or sinks) in the medium

∇ is the gradient operator

$\nabla\cdot$ is the divergence operator

$\mathbb{K}$ is the tensor of hydraulic permeability of the medium.

Assume that $\mathbb{K}$ is symmetric and uniformly bounded from below and above on Ω, i.e. there exist positive constants κ_0 and κ_1 such that the inequalities

$$\kappa_0|\mathbf{w}|^2 \leq \mathbf{w}^T\mathbb{K}(\mathbf{x})\mathbf{w} \leq \kappa_1|\mathbf{w}|^2$$

hold for all $\mathbf{x} \in \Omega$ and $\mathbf{w} \in \mathbb{R}^n$, where $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^n$.

The first equation (2.1) represents Darcy's law for the velocity field $\mathbf{u}$ and the second equation (2.2) derives from the law of conservation of mass and is called continuity equation.

In order to obtain a well-posed problem, the equations need to be equipped with boundary conditions. Consider the following **boundary conditions**

$$p = p_D \quad \text{on } \Gamma_D \quad \text{Dirichlet boundary condition} \tag{2.3}$$
$$\mathbf{u} \cdot \mathbf{n} = u_N \quad \text{on } \Gamma_N \quad \text{Neumann boundary condition} \tag{2.4}$$

where Γ_D and Γ_N are subsets of the boundary $\Gamma = \partial\Omega$ with $\Gamma = \overline{\Gamma}_D \cup \overline{\Gamma}_N$ and $\Gamma_D \cap \Gamma_N = \emptyset$; $\mathbf{n}$ is the outward normal vector defined (a.e.) on Γ.

2.1 Weak Formulation

The numerical solution of the Darcy problem (2.1) - (2.4) with finite element methods is based on its variational or equivalently called weak formulation.

Function spaces In order to derive a weak formulation of problem (2.1) - (2.4) function spaces that are based on the space of square (Lebesgue-) integrable functions on Ω,

$$L^2(\Omega) = \{w : \Omega \to \mathbb{R} \,|\, \int_\Omega |w(\mathbf{x})|^2 \, d\mathbf{x} = \|w\|^2_{L^2(\Omega)} < \infty\}$$

are used. To be precise, instead of functions, $L^2(\Omega)$ consists of classes of measurable functions, meaning that a class is made of functions that differ form each other only on a subset of Ω of zero Lebesgue measure. One keeps calling them functions, that are defined almost everywhere.

Consider the Hilbert space given for any positive integer m by

$$H^m(\Omega) = \{w \in L^2(\Omega) \,|\, D^\alpha w \in L^2(\Omega) \quad \forall\, |\alpha| \le m\}$$

with the norm

$$\|w\|^2_{H^m(\Omega)} = \sum_{|\alpha| \le m} \|D^\alpha w\|^2_{L^2(\Omega)}$$

where

$$\alpha = (\alpha_1, ..., \alpha_n), \quad |\alpha| = \alpha_1 + ... + \alpha_n \quad \text{and} \quad D^\alpha w = \frac{\partial^{|\alpha|} w}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}$$

and the derivatives $D^\alpha w$ are taken in a weak sense. In particular the space $H^1(\Omega)$ is of importance for the derivation of a weak formulation.

A function $w \in H^1(\Omega)$ is defined a.e. in Ω, i.e. everywhere except on a subset of Ω of zero Lebesgue measure. The boundary Γ has n-dimensional Lebesgue measure zero, since it is a $n-1$ dimensional subset of $\mathbb{R}^n$. Assigning boundary values along Γ to a function $w \in H^1(\Omega)$ is possible with the notion of a trace operator.

Theorem 2.1. *Assume Ω is bounded and the boundary $\Gamma = \partial\Omega$ is C^1. Then there exists a bounded linear operator $\gamma_0 : H^1(\Omega) \to L^2(\Gamma)$ such that*

$$\gamma_0 w = w|_\Gamma \quad \text{if } w \in C^1(\overline{\Omega}) \cap H^1(\Omega) \quad \text{and}$$
$$\|\gamma_0 w\|_{L^2(\Gamma)} \le C_0 \|w\|_{H^1(\Omega)} \quad \text{for each } w \in H^1(\Omega).$$

Proof: [7], chapter 5.5, Theorem 1.

Then $\gamma_0 w$ is called the trace of w on Γ and denoted by $w|_\Gamma$ even if w is a general function in $H^1(\Omega)$ that might not be in $C^1(\overline{\Omega})$.

The traces of functions in $H^1(\Omega)$ span a Hilbert space, denoted $H^{\frac{1}{2}}(\Gamma)$, that is a proper subspace of $L^2(\Gamma)$. Hence

$$H^{\frac{1}{2}}(\Gamma) = \gamma_0(H^1(\Omega))$$

with norm

$$\|g\|_{H^{\frac{1}{2}}(\Gamma)} = \inf_{v \in H^1(\Omega), \gamma_0 v = g} \|v\|_{H^1(\Omega)}.$$

The trace operator $\gamma_0 : H^1(\Omega) \to H^{\frac{1}{2}}(\Gamma)$ is surjective. The dual space of $H^{\frac{1}{2}}(\Gamma)$ is denoted by $H^{-\frac{1}{2}}(\Gamma)$ and the dual pairing between $H^{\frac{1}{2}}(\Gamma)$ and $H^{-\frac{1}{2}}(\Gamma)$ by $\langle \cdot, \cdot \rangle$. For an introduction to the spaces $H^s(\Gamma)$ (defined for all $s \in \mathbb{R}$) see [9], p. 8.

To write an appropriate weak formulation of problem (2.1) - (2.4) we introduce the Hilbert space

$$H(\mathrm{div}, \Omega) = \{\mathbf{v} \in L^2(\Omega)^n \,|\, \nabla \cdot \mathbf{v} \in L^2(\Omega)\}$$

with the norm

$$\|\mathbf{v}\|^2_{H(\mathrm{div},\Omega)} = \|\mathbf{v}\|^2_{L^2(\Omega)} + \|\nabla \cdot \mathbf{v}\|^2_{L^2(\Omega)}.$$

Functions in $H(\mathrm{div}, \Omega)$ admit traces of the normal component on Γ, namely there exists a bounded linear and surjective operator $\gamma_n : H(\mathrm{div}, \Omega) \to H^{-\frac{1}{2}}(\Gamma)$ such that $\gamma_n \mathbf{v} = \mathbf{v} \cdot \mathbf{n}|_\Gamma$, for every smooth $\mathbf{v}$.

Lemma 2.2. *For $\mathbf{v} \in H(div, \Omega)$ one can define $\mathbf{v} \cdot \mathbf{n}|_\Gamma \in H^{-\frac{1}{2}}(\Gamma)$, the normal trace of $\mathbf{v}$ on Γ and Green's formula holds,*

$$\int_\Omega (\nabla \cdot \mathbf{v}) p \, d\mathbf{x} + \int_\Omega \nabla p \cdot \mathbf{v} \, d\mathbf{x} = \langle \mathbf{v} \cdot \mathbf{n}, p \rangle \quad \forall p \in H^1(\Omega)$$

where one can write $\int_\Gamma (\mathbf{v} \cdot \mathbf{n}) p \, ds$ instead of $\langle \mathbf{v} \cdot \mathbf{n}, p \rangle$, to denote the duality between $H^{\frac{1}{2}}(\Gamma)$ and $H^{-\frac{1}{2}}(\Gamma)$.

Proof: [3], Lemma 2.1.1.

The restriction of $\mathbf{v} \cdot \mathbf{n}$ to Γ_N, however, may not lie in $H^{-\frac{1}{2}}(\Gamma_N)$. The subspace of $H(\mathrm{div}, \Omega)$ consisting of functions with zero normal trace on Γ_N is given by

$$H_{0,\Gamma_N}(\mathrm{div}, \Omega) = \{\mathbf{v} \in H(\mathrm{div}, \Omega) \,|\, \langle \mathbf{v} \cdot \mathbf{n}, w \rangle = 0 \quad \forall w \in H^1_{0,\Gamma_D}(\Omega)\} \tag{2.5}$$

where

$$H^1_{0,\Gamma_D}(\Omega) = \{w \in H^1(\Omega) \,|\, w|_{\Gamma_D} = 0\}.$$

Homogeneous problem Consider the case $u_N = 0$. A weak formulation of problem (2.1) - (2.4) is obtained by multiplying eq. (2.1) with a test function $\mathbf{v} \in H_{0,\Gamma_N}(\mathrm{div}, \Omega)$ and the continuity equation (2.2) with test function $q \in L^2(\Omega)$. Then both equations are integrated over Ω,

$$\int_\Omega \mathbb{K}\mathbf{u} \cdot \mathbf{v}\, d\mathbf{x} + \int_\Omega \nabla p \cdot \mathbf{v}\, d\mathbf{x} = 0 \quad \forall \mathbf{v} \in H_{0,\Gamma_N}(\mathrm{div}, \Omega)$$
$$\int_\Omega (\nabla \cdot \mathbf{u}) q\, d\mathbf{x} = \int_\Omega f q\, d\mathbf{x} \quad \forall q \in L^2(\Omega).$$

Assume that the functions $(\mathbf{u}, p)$ are sufficiently smooth and apply Green's formula to the second integral in the first equation

$$\int_\Omega \nabla p \cdot \mathbf{v}\, d\mathbf{x} = -\int_\Omega (\nabla \cdot \mathbf{v}) p\, d\mathbf{x} + \int_\Gamma (\mathbf{v} \cdot \mathbf{n}) p\, ds$$

and insert $p|_{\Gamma_D} = p_D$ in the integral over the boundary Γ_D,

$$\int_\Gamma (\mathbf{v} \cdot \mathbf{n}) p\, ds = \int_{\Gamma_N} (\mathbf{v} \cdot \mathbf{n}) p\, ds + \int_{\Gamma_D} (\mathbf{v} \cdot \mathbf{n}) p_D\, ds.$$

The second term makes sense if $p_D \in H^{\frac{1}{2}}(\Gamma_D)$ and the boundary integral reads as the duality product between $H^{-\frac{1}{2}}$ and $H^{\frac{1}{2}}$. Since $\mathbf{v} \in H_{0,\Gamma_N}(\mathrm{div}, \Omega)$ the first term vanishes.

Then the weak formulation of the Darcy problem (2.1) - (2.4) with inhomogeneous Dirichlet and homogeneous Neumann boundary conditions reads as follows:

Given $f \in L^2(\Omega)$ and $p_D \in H^{\frac{1}{2}}(\Gamma_D)$ find $(\mathbf{u}, p) \in H_{0,\Gamma_N}(\mathrm{div}, \Omega) \times L^2(\Omega)$ such that

$$\begin{aligned} \int_\Omega \mathbb{K}\mathbf{u} \cdot \mathbf{v}\, d\mathbf{x} - \int_\Omega (\nabla \cdot \mathbf{v})\, p\, d\mathbf{x} &= -\int_{\Gamma_D} (\mathbf{v} \cdot \mathbf{n}) p_D\, ds \quad \forall \mathbf{v} \in H_{0,\Gamma_N}(\mathrm{div}, \Omega) \\ \int_\Omega (\nabla \cdot \mathbf{u}) q\, d\mathbf{x} &= \int_\Omega f q\, d\mathbf{x} \qquad\qquad \forall q \in L^2(\Omega). \end{aligned} \tag{2.6}$$

The Dirichlet boundary condition is implicit in the weak formulation. This type of boundary condition is called natural. The Neumann boundary condition has to be imposed on the velocity space and is called essential boundary condition.

Introducing two continuous bilinear forms

$$a(\mathbf{u}, \mathbf{v}) = \int_\Omega \mathbb{K}\mathbf{u} \cdot \mathbf{v}\, d\mathbf{x} \tag{2.7}$$

and

$$b(\mathbf{v}, p) = -\int_\Omega (\nabla \cdot \mathbf{v})\, p\, d\mathbf{x} \tag{2.8}$$

on $H_{0,\Gamma_N}(\mathrm{div}, \Omega) \times H_{0,\Gamma_N}(\mathrm{div}, \Omega)$ and $H_{0,\Gamma_N}(\mathrm{div}, \Omega) \times L^2(\Omega)$, respectively, (2.6) can be written in the following form:

Given $f \in L^2(\Omega)$ and $p_D \in H^{\frac{1}{2}}(\Gamma_D)$ find $(\mathbf{u}, p) \in H_{0,\Gamma_N}(\mathrm{div}, \Omega) \times L^2(\Omega)$ such that

$$\begin{aligned} a(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) &= -\langle \mathbf{v} \cdot \mathbf{n}, p_D \rangle \qquad \forall \mathbf{v} \in H_{0,\Gamma_N}(\mathrm{div}, \Omega) \\ b(\mathbf{u}, q) &= -\langle f, q \rangle \qquad\qquad\qquad \forall q \in L^2(\Omega), \end{aligned} \tag{2.9}$$

where $\langle f, q \rangle = \int_\Omega f q \, d\mathbf{x}$ and $\langle \cdot, \cdot \rangle$ denotes the dual pairing between a space and its dual space.

Non-homogeneous problem Let $u_N \neq 0$. By taking the divergence of eq. (2.1), $\nabla \cdot \mathbf{u} + \nabla \cdot \mathbb{K}^{-1} \nabla p = 0$ and substituting in eq. (2.2), one obtains an alternative formulation of the Darcy equations

$$-\nabla \cdot \mathbb{K}^{-1} \nabla p = f \quad \text{in } \Omega.$$

Assume $\mathbb{K}^{-1} \nabla p \cdot \mathbf{n}\,|_{\Gamma_N} = u_N$, $p\,|_{\Gamma_D} = 0$ and $f = 0$. Multiplication with $q \in H^1_{0,\Gamma_D}(\Omega)$, integration and integration by parts yields

$$\int_\Omega \mathbb{K}^{-1} \nabla p \cdot \nabla q \, d\mathbf{x} = \int_{\Gamma_N} u_N q \, ds. \tag{2.10}$$

One can show with Lax- Milgram theorem that for $u_N \in H^{-\frac{1}{2}}(\Gamma_N)$ problem (2.10) has a unique solution $p \in H_{0,\Gamma_D}(\Omega)$, see [3].

Thus it is possible to consider any $\tilde{\mathbf{u}}$ such that $\tilde{\mathbf{u}} \cdot \mathbf{n} = u_N$ on Γ_N by considering a solution to problem (2.10) and taking $\tilde{\mathbf{u}} = \mathbb{K}^{-1} \nabla p$. Then look for $\mathbf{u} = \tilde{\mathbf{u}} + \mathbf{u_0}$ with $\mathbf{u_0} \in H_{0,\Gamma_N}(\mathrm{div}, \Omega)$. This leads to the following problem

$$\begin{aligned} a(\mathbf{u_0}, \mathbf{v}) + b(\mathbf{v}, p) &= -\langle \mathbf{v} \cdot \mathbf{n}, p_D \rangle - a(\tilde{\mathbf{u}}, \mathbf{v}) \qquad \forall \mathbf{v} \in H_{0,\Gamma_N}(\mathrm{div}, \Omega) \\ b(\mathbf{u_0}, q) &= -\langle f, q \rangle - b(\tilde{\mathbf{u}}, q) \qquad\qquad\qquad\quad \forall q \in L^2(\Omega). \end{aligned} \tag{2.11}$$

This means that considering $u_N \neq 0$ can be reduced to changing the right-hand side of (2.9).

Problem (2.9) is a particular case of a general class of problems described as follows: Let V and Q be two Hilbert spaces and their corresponding dual spaces given by V' and Q'. Suppose that $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ are two continuous bilinear forms on $V \times V$ and $V \times Q$, respectively. For given $g \in V'$ and $f \in Q'$ find $(\mathbf{u}, p) \in V \times Q$ such that

$$\begin{aligned} a(\mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) &= \langle g, \mathbf{v} \rangle \quad \forall \mathbf{v} \in V \\ b(\mathbf{u}, q) &= \langle f, q \rangle \quad \forall q \in Q. \end{aligned} \tag{2.12}$$

System (2.12) is called **linear saddle point problem**.

2.2 Existence and Uniqueness

In the following conditions on the bilinear forms $a(\cdot, \cdot)$ and $b(\cdot, \cdot)$ for the well-posedness of the linear saddle point problem (2.12), i.e. the existence of a unique solution of (2.12), are given.

Theorem 2.3. *Assume that the bilinear form $a(\cdot,\cdot)$ is coercive on the subspace*

$$W = \{\mathbf{v} \in V \mid b(\mathbf{v}, q) = 0 \quad \forall q \in Q\} \subset V,$$

i.e. there exists a constant $\alpha \geq 0$ such that

$$a(\mathbf{v}, \mathbf{v}) \geq \alpha \|\mathbf{v}\|_V^2 \quad \forall \mathbf{v} \in W.$$

Then problem (2.12) has a unique solution $(\mathbf{u}, p) \in V \times Q$ if and only if $b(\cdot,\cdot)$ satisfies the inf-sup condition, i.e.

$$\exists \beta > 0 : \qquad \inf_{q \in Q \setminus \{0\}} \sup_{\mathbf{v} \in V \setminus \{0\}} \frac{b(\mathbf{v}, q)}{\|\mathbf{v}\|_V \|q\|_Q} \geq \beta. \tag{2.13}$$

Proof: [9], Corollary 4.1.

Apply the theorem 2.3 to problem (2.9), where

$$V = H_{0,\Gamma_N}(\mathrm{div}, \Omega) \quad \text{and} \quad Q = L^2(\Omega)$$

and the bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ are given by (2.7) and (2.8).

Coercivity Let $\mathbf{v} \in W$, i.e.

$$b(\mathbf{v}, q) = -\int_\Omega (\nabla \cdot \mathbf{v})\, q \, d\mathbf{x} = 0 \quad \forall q \in L^2(\Omega).$$

Since $\mathbf{v} \in H_{0,\Gamma_N}(\mathrm{div}, \Omega)$ it holds $\nabla \cdot \mathbf{v} \in L^2(\Omega)$. Choosing q to be $\nabla \cdot \mathbf{v}$,

$$0 = \int_\Omega (\nabla \cdot \mathbf{v})(\nabla \cdot \mathbf{v})\, d\mathbf{x} = ||\nabla \cdot \mathbf{v}||^2_{L^2(\Omega)}.$$

Thus the $H(\mathrm{div}, \Omega)$-norm coincides on W with the $L^2(\Omega)$-norm. Therefore coercivity holds

$$a(\mathbf{v}, \mathbf{v}) = \int_\Omega \mathbf{v}^T \mathbb{K} \mathbf{v}\, d\mathbf{x} \geq \kappa_0 ||\mathbf{v}||^2_{L^2(\Omega)} = \kappa_0 ||\mathbf{v}||^2_V \quad \forall \mathbf{v} \in W.$$

Inf-sup condition The bilinear form $b(\cdot,\cdot)$ given by (2.8) satisfies the inf-sup condition on $(H^1_{0,\Gamma_N}(\Omega))^n \times L^2(\Omega)$, ([12], theorem 3.46). On $H^1_{0,\Gamma_N}(\Omega)$ the norms $\|\cdot\|_{H^1_0(\Omega)}$ and $\|\cdot\|_{H^1(\Omega)}$ are equivalent. Thus the inf-sup condition still holds, when replacing $\|\cdot\|_{H^1_0(\Omega)}$ by $\|\cdot\|_{H^1(\Omega)}$.

From the estimate

$$||\nabla \cdot \mathbf{v}||_{L^2(\Omega)} \leq \sqrt{d} ||\nabla \mathbf{v}||_{L^2(\Omega)} \quad \text{for } \mathbf{v} \in (H^1(\Omega))^n$$

(proof: [12], Lemma 3.34) it follows

$$||\mathbf{v}||^2_{H^1(\Omega)} \geq \frac{1}{d} ||\mathbf{v}||^2_{H(\mathrm{div},\Omega)}.$$

Since $H_{0,\Gamma_N}(\mathrm{div},\Omega)$ contains $H^1_{0,\Gamma_N}(\Omega)$, the supremum over the larger space $H_{0,\Gamma_N}(\mathrm{div},\Omega)$ will be greater. Using the last inequality

$$\inf_{q\in L^2(\Omega)\setminus\{0\}} \sup_{\mathbf{v}\in H_{0,\Gamma_N}(\mathrm{div},\Omega)\setminus\{0\}} \frac{b(\mathbf{v},q)}{\|\mathbf{v}\|_{H(\mathrm{div},\Omega)}\|q\|_{L^2(\Omega)}} \geq$$
$$\inf_{q\in L^2(\Omega)\setminus\{0\}} \sup_{\mathbf{v}\in H^1_{0,\Gamma_N}(\Omega)\setminus\{0\}} \frac{b(\mathbf{v},q)}{\sqrt{d}\|\mathbf{v}\|_{H^1(\Omega)}\|q\|_{L^2(\Omega)}} \geq \frac{\beta}{\sqrt{d}}.$$

Continuity of the bilinear forms follow from Hölder inequality

$$|a(\mathbf{u},\mathbf{v})| \leq \int_\Omega |\mathbf{u}^T\mathbb{K}\mathbf{v}|\,dx \leq \kappa_1\|\mathbf{u}\|_{H(\mathrm{div},\Omega)}\|\mathbf{v}\|_{H(\mathrm{div},\Omega)}$$
$$|b(\mathbf{v},p)| \leq \int_\Omega |(\nabla\cdot\mathbf{v})\,p|\,dx \leq \|\mathbf{v}\|_{H(\mathrm{div},\Omega)}\|p\|_{L^2(\Omega)}.$$

2.3 Finite Element Discretization

The main idea of using finite element methods consists in replacing the infinite-dimensional spaces V and Q by finite-dimensional spaces V^h and Q^h, respectively, and to apply Galerkin method. The use of two different finite element spaces for the approximation of the two variables $(\mathbf{u},p)$ is denoted by mixed finite element method. If $V^h\subset V$ and $Q^h\subset Q$ the finite element method is called conforming, otherwise non-conforming. In the following only conforming finite element methods are considered. For conforming spaces the discrete bilinear forms are the restrictions of $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ from $V\times V$ to $V^h\times V^h$ and $V\times Q$ to $V^h\times Q^h$, respectively.

Finite element formulation: Let $V^h\subset V$ be a velocity finite element space and let $Q^h\subset Q$ be a pressure finite element space. The finite element discretization of (2.12) reads as follws:

Find $(\mathbf{u}^h,p^h)\in V^h\times Q^h$ such that

$$\begin{aligned} a(\mathbf{u}^h,\mathbf{v}^h)+b(\mathbf{v}^h,p^h) &= \langle g,\mathbf{v}^h\rangle \quad \forall\mathbf{v}^h\in V^h \\ b(\mathbf{u}^h,q^h) &= \langle f,q^h\rangle \quad \forall q^h\in Q^h. \end{aligned} \tag{2.14}$$

In order to have the finite element approximation well defined we need to know that there exists a unique solution $(\mathbf{u}^h,p^h)\in V^h\times Q^h$ of problem (2.14).

Introduce the space

$$W^h=\{\mathbf{v}^h\in V^h\,|\,b(\mathbf{v}^h,q^h)=0\quad\forall\,q^h\in Q^h\}\subset V^h \tag{2.15}$$

and assume $a(\cdot,\cdot)$ is coercive on W^h. From Theorem 2.3 it follows, that there exists a unique finite element solution if and only if the discrete inf-sup condition

$$\exists\beta^*>0:\qquad \inf_{q^h\in Q^h\setminus\{0\}}\sup_{\mathbf{v}^h\in V^h\setminus\{0\}}\frac{b(\mathbf{v}^h,q^h)}{\|\mathbf{v}^h\|_V\|q^h\|_Q}\geq\beta^* \tag{2.16}$$

holds.

A useful criterion to check the discrete inf-sup condition is the following result due to Fortin.

Theorem 2.4. *(Fortin, 1977, [8]) Assume the continuous inf-sup condition* (2.13) *holds. Then the discrete inf-sup condition holds with a constant* $\beta^* > 0$ *independent of* h*, if and only if, there exists an operator* $\Pi_h : V \to V^h$ *such that*

$$b(\mathbf{v} - \Pi_h \mathbf{v}, q^h) = 0 \quad \forall\, \mathbf{v} \in V, \forall\, q^h \in Q^h \tag{2.17}$$

and

$$\|\Pi_h \mathbf{v}\|_V \leq C \|\mathbf{v}\|_V \quad \forall \mathbf{v} \in V \tag{2.18}$$

with a constant $C > 0$ *independent of* h*.*
Proof: [5], Theorem 5.7.

2.4 Matrix-Vector Form

In order to derive an algebraic system from (2.9) or (2.11), the spaces V^h and Q^h are equipped with a basis. Let $\{\boldsymbol{\varphi}_i^h\}_{i=1,...,n_V}$ be a basis of V^h and $\{\psi_j^h\}_{j=1,...,n_Q}$ be a basis of Q^h, where $n_V = \dim V^h$ and $n_Q = \dim Q^h$. Then $\mathbf{u}^h \in V^h$ and $p^h \in Q^h$ have the unique representations

$$\mathbf{u}^h = \sum_{i=1}^{n_V} u_i^h \boldsymbol{\varphi}_i^h, \quad p^h = \sum_{j=1}^{n_Q} p_j^h \psi_j^h \tag{2.19}$$

with unknown coefficients u_i^h, $i = 1, ..., n_V$ and p_j^h, $j = 1, ..., n_Q$. Inserting (2.19) in (2.14) and testing with each basis function separately lead to the linear system of equations

$$\begin{aligned} \sum_{i=1}^{n_V} a(\boldsymbol{\varphi}_i^h, \boldsymbol{\varphi}_l^h) u_i^h + \sum_{j=1}^{n_Q} b(\boldsymbol{\varphi}_l^h, \psi_j^h) p_j^h &= \langle \tilde{g}, \boldsymbol{\varphi}_l^h \rangle \quad l = 1, ..., n_V \\ \sum_{i=1}^{n_V} b(\boldsymbol{\varphi}_i^h, \psi_k^h) u_i^h &= \langle \tilde{f}, \psi_k^h \rangle \quad k = 1, ..., n_Q, \end{aligned}$$

which can be expressed in matrix-vector form as

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \cdot \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix} = \begin{pmatrix} \underline{g} \\ \underline{f} \end{pmatrix} \tag{2.20}$$

with finite element matrices

$$(A)_{li} := a(\boldsymbol{\varphi}_i^h, \boldsymbol{\varphi}_l^h) = \int_\Omega \mathbb{K} \boldsymbol{\varphi}_i^h \cdot \boldsymbol{\varphi}_l^h \, d\mathbf{x}, \quad A \in \mathbb{R}^{n_V \times n_V}$$

$$(B)_{ki} := b(\boldsymbol{\varphi}_i^h, \psi_k^h) = -\int_\Omega (\nabla \cdot \boldsymbol{\varphi}_i^h) \psi_k^h \, d\mathbf{x}, \quad B \in \mathbb{R}^{n_Q \times n_V}$$

coefficient vectors

$$(\underline{u})_i := u_i^h, \quad \underline{u} \in \mathbb{R}^{n_V}$$

$$(\underline{p})_j := p_j^h, \quad \underline{p} \in \mathbb{R}^{n_Q},$$

and right-hand sides including eventually non-homogeneous boundary conditions as in (2.11),

$$(\underline{g})_l := \langle \tilde{g}, \boldsymbol{\varphi}_l^h \rangle = -\int_{\Gamma_D} (\boldsymbol{\varphi}_l^h \cdot \mathbf{n}) p_D \, ds - a(\tilde{\mathbf{u}}, \boldsymbol{\varphi}_l^h), \quad \underline{g} \in \mathbb{R}^{n_V}$$

$$(\underline{f})_k := \langle \tilde{f}, \psi_k^h \rangle = -\int_\Omega f \psi_k^h \, d\mathbf{x} - b(\tilde{\mathbf{u}}, \psi_k^h), \quad \underline{f} \in \mathbb{R}^{n_Q}.$$

The algebraic systems (2.20) that arise in the finite element discretizations of the Darcy equations belong to the type of linear systems in saddle point form and have block structure. The symmetric positive definite matrix A is a discretization of the linear operator 'multiplication by $\mathbb{K}$', a zeroth-order differential operator. The conditioning properties of A are independent of the discretization parameter h (for most discretizations), and depend only on properties of the hydraulic permeability tensor $\mathbb{K}$. The matrix $-B$ represents a discrete divergence operator and B^T a discrete gradient operator. Properties of saddle point matrices such as solvability condition, spectral properties and conditioning will be considered in section 4.1.

3 The Choice of Finite Element Spaces

This chapter introduces appropriate finite element spaces for the discretization of problem (2.9) and follows [3], [5], and [12].

The construction of finite-dimensional subspaces of the velocity and pressure spaces is based on the decomposition of the domain Ω, in which the problem is posed, into polyhedrons. This decomposition is called triangulation $\mathcal{T}_h$. The polyhedrons T are called mesh cells and are usually triangles, quadrilaterals, tetrahedra, hexahedra, e.t.c. The union of the polyhedrons is called grid or mesh and $h = \max_{T\in\mathcal{T}_h} h_T$ is the mesh size, with h_T denoting the diameter of T. Let $\{\mathcal{T}_h\}$ be a familiy of triangulations. Assume that

- it holds $\overline{\Omega} = \cup_{T\in\mathcal{T}_h} T$
- each mesh cell $T \in \mathcal{T}_h$ is closed, the boundary is Lipschitz continuous and the interior $\mathring{T}$ in nonempty,
- for $T_1 \neq T_2$ it holds $\mathring{T_1} \cap \mathring{T_2} = \emptyset$
- the intersection of two elements in $\mathcal{T}_h$ is either empty or a common m-face, $m \in \{0, ..., n-1\}$
- the family of triangulations is regular, i.e there is a constant $\sigma > 0$ independent of h such that $\frac{h_T}{p_T} \leq \sigma$ for all $T \in \mathcal{T}_h$, where p_T is the diameter of the largest ball inscribed in T.

The velocity and pressure spaces are then approximated by piecewise (usually) polynomial functions defined on each mesh cell T or functions obtained from polynomials by a change of variable.

Let $P(T) \subset C^s(T)$, $s \in \mathbb{N}$, be a finite-dimensional space defined on the mesh cell T. For the definition of finite element spaces one has to specify linear functionals defined on $P(T)$. Let $\Phi_{T,1}, ..., \Phi_{T,\dim P(T)} : C^s(T) \to \mathbb{R}$ be linear and continuous functionals that are linearly independent, where the smoothness parameter s has to be chosen in such a way that the functionals $\Phi_{T,1}, ..., \Phi_{T,\dim P(T)}$ are continuous.

Assume that $P(T)$ is **unisolvent** with respect to $\Phi_{T,1}, ..., \Phi_{T,\dim P(T)}$, i.e. for each $\mathbf{a} \in \mathbb{R}^{\dim P(T)}$ there exists exactly one $p \in P(T)$ such that

$$\Phi_{T,i}(p) = a_i \quad \forall i = 1, ..., \dim P(T).$$

Unisolvence means that every function $p \in P(T)$ is uniquely determined by values under the functionals, **the degrees of freedom**. Choosing in particular the Cartesian unit

vectors for $\mathbf{a}$, then it follows from the unisolvence that a set $\{\phi_{T,i}\}_{i=1}^{\dim P(T)}$ in $P(T)$ exists, such that

$$\Phi_{T,i}(\phi_{T,j}) = \delta_{ij}, \quad i,j = 1,...,\dim P(T).$$

Consequently, the set $\{\phi_{T,i}\}_{i=1}^{\dim P(T)}$ forms a basis of $P(T)$.

Let $\Phi_i : C^s(\overline{\Omega}) \to \mathbb{R}$, $i = 1,...,N$ be continuous linear functionals whose restiction to $C^s(T)$ are the local functionals $\Phi_{T,1},...,\Phi_{T,\dim P(T)}$. The subdomain w_i denotes the union of all mesh cells T_j for which there is a $v \in P(T_j)$ such that $\Phi_i(v) \neq 0$.

A function v defined on Ω with $v|_T \in P(T)$ for all $T \in \mathcal{T}_h$ is called continuous with respect to Φ_i if

$$\Phi_i(v|_{T_1}) = \Phi_i(v|_{T_2}) \quad \forall\, T_1, T_2 \in w_i.$$

Then the space

$$S = \{v \in L^\infty(\Omega) \,|\, v|_T \in P(T) \text{ and } v \text{ is continuous with respect to } \Phi_i,\ i = 1,...,N\} \tag{3.1}$$

is called **finite element space**. Recall that $L^\infty(\Omega) = \{v : \Omega \to \mathbb{R} \,|\, \text{ess sup}_{\mathbf{x}\in\Omega}|v(\mathbf{x})| < \infty\}$.

The global basis $\{\phi_j\}_{j=1}^N$ of S is defined by the condition

$$\phi_j \in S, \quad \Phi_i(\phi_j) = \delta_{ij}, \quad i,j = 1,...,N.$$

3.1 Approximations of $H(\mathbf{div}, \Omega)$ on simplicial grids

In order to define finite element approximations to the solution $(\mathbf{u}, p)$ of (2.9) we need to introduce finite-dimensional subspaces of $H(\text{div}, \Omega)$ and $L^2(\Omega)$ made of piecewise polynomial functions. First consider the case of simplicial grids and the associated Raviart-Thomas spaces which are the best-known spaces approximating $H(\text{div}, \Omega)$.

3.1.1 Raviart-Thomas spaces

Given a simplex $T \in \mathbb{R}^n$ the **local Raviart-Thomas space of order** $k \geq 0$ is defined by

$$\mathcal{RT}_k(T) = \mathcal{P}_k(T)^n + \mathbf{x}\mathcal{P}_k(T) \tag{3.2}$$

where $\mathcal{P}_k$ is the space of polynomials of degree less than or equal to k. Denote with $F_i, i = 1,...,n+1$, the faces of the simplex T and with $n_i, i = 1,...,n+1$ their corresponding exterior unit normals. Local Raviart-Thomas spaces have the following properties:

Lemma 3.1. *a) dim*$\mathcal{RT}_k(T) = n\binom{k+n}{k} + \binom{k+n-1}{k}$.

b) If $\mathbf{v} \in \mathcal{RT}_k(T)$, *then* $\mathbf{v} \cdot \mathbf{n}_i \in \mathcal{P}_k(F_i)$ *for* $i = 1,...,n+1$.

Proof: The proof can be found in [5], Lemma 3.1.

The degrees of freedom of $\mathcal{RT}_k(T)$ define a local interpolation operator

$$\Pi_T : H^1(T)^n \to \mathcal{RT}_k(T).$$

Lemma 3.2. *Given* $\mathbf{v} \in H^1(T)^n$ *there exists a unique* $\Pi_T\mathbf{v} \in \mathcal{RT}_k(T)$ *such that*

$$\int_{F_i} \Pi_T\mathbf{v} \cdot \mathbf{n}_i \, p_k \, ds = \int_{F_i} \mathbf{v} \cdot \mathbf{n}_i \, p_k \, ds \quad \forall p_k \in \mathcal{P}_k(F_i), i = 1, ..., n+1 \tag{3.3}$$

and if $k \geq 1$

$$\int_T \Pi_T\mathbf{v} \cdot \mathbf{p}_{k-1} \, d\mathbf{x} = \int_T \mathbf{v} \cdot \mathbf{p}_{k-1} \, d\mathbf{x} \quad \forall \mathbf{p}_{k-1} \in \mathcal{P}_{k-1}(T)^n. \tag{3.4}$$

Proof: The proof can be found in [5], Lemma 3.2

In the proof it is first shown that the number of conditions defining $\Pi_T\mathbf{v}$ equals the dimension of $\mathcal{RT}_k(T)$. Since $\dim\mathcal{P}_k(F_i) = \binom{k+n-1}{k}$, the number of conditions in (3.3) is

$$\text{the number of faces} \times \dim\mathcal{P}_k(F_i) = (n+1)\binom{k+n-1}{k}.$$

On the other hand, the number of conditions in (3.4) is $\dim(\mathcal{P}^n_{k-1}(T)) = n\binom{k+n-1}{k-1}$. Then the total number of conditions defining $\Pi_T\mathbf{v}$ is

$$(n+1)\binom{k+n-1}{k} + n\binom{k+n-1}{k-1}.$$

After rewriting one gets that the number of conditions defining $\Pi_T\mathbf{v}$ is precisely the dimension of $\mathcal{RT}_k(T)$. Therefore, in order to show existence of $\Pi_T\mathbf{v}$, it is enough to prove uniqueness, i.e. for $\mathbf{v} \in \mathcal{RT}_k(T)$ such that

$$\int_{F_i} \mathbf{v} \cdot \mathbf{n}_i \, p_k \, ds = 0 \quad \forall p_k \in \mathcal{P}_k(F_i), i = 1, ..., n+1$$

and

$$\int_T \mathbf{v} \cdot \mathbf{p}_{k-1} \, d\mathbf{x} = 0 \quad \forall \mathbf{p}_{k-1} \in \mathcal{P}_{k-1}(T)^n$$

it follows $\mathbf{v} = \mathbf{0}$ (see [5], p. 13). The proof implies that $\mathcal{RT}_k(T)$ is unisolvent with respect to the degrees of freedom definining the interpolation operator. Figure 3.1 shows the degrees of freedom for $k = 0$ and $k = 1$ in the two dimensional case. The arrows indicate values of normal components and the filled circle values of $\mathbf{v}$ (and so it corresponds to two degrees of freedom).

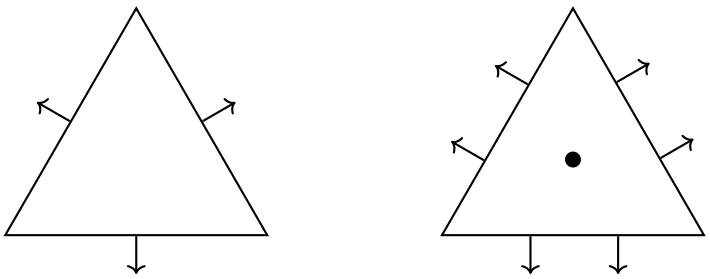

Figure 3.1: Degrees of freedom for $\mathcal{RT}_0$ and $\mathcal{RT}_1$ in $\mathbb{R}^2$

The degrees of freedom are choosen in such a way that a piecewise local Raviart-Thomas function belongs to $H(\mathrm{div}, \Omega)$.

Lemma 3.3 (**Sufficient and Necessary Condition for a Finite Element Function to be in** $H(\mathbf{div}, \Omega)$)**.** *Let $\mathcal{T}_h$ be a regular triangulation of Ω. A finite element function $\mathbf{v}^h \in (L^2(\Omega))^n$, i.e. piecewise polynomial vector function belongs to $H(div, \Omega)$ if and only if $\mathbf{v}^h \cdot \mathbf{n}_F$ is continuous for all faces F of the triangulation.*

Proof: [12], Chapter 3, Lemma 3.66

Let $\mathbf{v}^h$ be a piecewise local Raviart-Thomas function. From the definition of the finite element space (3.1) it follows that $\mathbf{v}^h$ is continuous with respect to the degrees of freedom, i.e. it holds

$$\int_F \mathbf{v}^h|_{T_1} \cdot \mathbf{n}_F \, p_k \, ds = \int_F \mathbf{v}^h|_{T_2} \cdot \mathbf{n}_F \, p_k \, ds \quad \forall \, p_k \in \mathcal{P}_k(F), \, \forall \, T_1, T_2 \in \mathcal{T}_h, \; F = T_1 \cap T_2.$$

Therefore the jumps $\left[|\mathbf{v}^h \cdot \mathbf{n}_F|\right]_F := \mathbf{v}^h|_{T_1} \cdot \mathbf{n}_F - \mathbf{v}^h|_{T_2} \cdot \mathbf{n}_F$ vanish on all interior faces, which is equivalent to the continuity of the normal component of $\mathbf{v}^h$ across all faces of the mesh cells in $\mathcal{T}_h$.

The **global Raviart-Thomas finite element space** associated with the triangulation $\mathcal{T}_h$ can be defined by

$$\mathcal{RT}_k(\mathcal{T}_h) = \{\mathbf{v} \in H(\mathrm{div}, \Omega) \mid \mathbf{v}|_T \in \mathcal{RT}_k(T) \quad \forall \, T \in \mathcal{T}_h\}.$$

The global interpolation operator

$$\Pi_h : H(\mathrm{div}, \Omega) \cap \prod_{T \in \mathcal{T}_h} H^1(T)^n \to \mathcal{RT}_k(\mathcal{T}_h)$$

is defined by setting

$$\Pi_h \mathbf{v}|_T = \Pi_T \mathbf{v} \quad \forall T \in \mathcal{T}_h.$$

Since by definition $\Pi_h \mathbf{v}|_T = \Pi_T \mathbf{v} \in \mathcal{RT}_k(T)$, it remains to see that $\Pi_h \mathbf{v} \in H(\mathrm{div}, \Omega)$. A piecewise polynomial vector function $\Pi_h \mathbf{v}$ is in $H(\mathrm{div}, \Omega)$ if and only if it has continuous normal components across the mesh cells. Since $\mathbf{v} \in H(\mathrm{div}, \Omega)$, the continuity of the normal component of $\Pi_h \mathbf{v}$ follows from lemma 3.1 b), in view of the degrees of freedom (3.3) in the definition of Π_T.

The finite element space for the approximation of the scalar variable p is the standard space of piecewise polynomials of degree k, namely

$$\mathcal{P}_k^d(\mathcal{T}_h) = \{q \in L^2(\Omega) \, | \, q|_T \in \mathcal{P}_k(T)\}.$$

where d stands for 'discontinuous'. Since no derivative of the scalar variable appears in the weak formulation (2.9), no continuity in the approximation space for this variable is required.

Lemma 3.4. *The operator* Π_h *satisfies*

$$\int_\Omega \nabla \cdot (\mathbf{v} - \Pi_h \mathbf{v}) q^h \, d\mathbf{x} = 0 \tag{3.5}$$

$\forall \mathbf{v} \in H(div, \Omega) \cap \prod_{T \in \mathcal{T}_h} H^1(T)^n$ *and* $\forall q^h \in \mathcal{P}_k^d$. *Moreover,*

$$\nabla \cdot \mathcal{RT}_k = \mathcal{P}_k^d \tag{3.6}$$

Proof: From (3.3) and (3.4) it follows, that for any $\mathbf{v} \in H^1(T)^n$ and any $q^h \in \mathcal{P}_k^d(T)$

$$\int_T \nabla \cdot (\mathbf{v} - \Pi_T \mathbf{v}) q^h \, d\mathbf{x} = -\int_T (\mathbf{v} - \Pi_T \mathbf{v}) \cdot \nabla q^h \, d\mathbf{x} + \int_{\partial T} (\mathbf{v} - \Pi_T \mathbf{v}) \cdot \mathbf{n} \, q^h \, ds = 0$$

thus (3.5) holds. It is easy to see that $\nabla \cdot \mathcal{RT}_k \subset \mathcal{P}_k^d$. To see the other inclusion recall that $\operatorname{div} : H^1(\Omega)^n \to L^2(\Omega)$ is surjective (see [5], Lemma 2.4). Therefore, given $q^h \in \mathcal{P}_k^d$ there exists $\mathbf{v} \in H^1(\Omega)^n$ such that $\nabla \cdot \mathbf{v} = q^h$. From (3.5) it follows that $\nabla \cdot \Pi_h \mathbf{v} = q^h$.

3.1.2 Application to Darcy Equations

In order to discretize problem (2.9) or (2.11) using the Raviart-Thomas space $\mathcal{RT}_k(\mathcal{T}_h)$ define

$$V^h := \{\mathbf{v}^h \in \mathcal{RT}_k(\mathcal{T}_h) \,|\, \mathbf{v}^h \cdot \mathbf{n}|_{\Gamma_N} = 0\}$$

in the sense of $\mathbf{v}^h \cdot \mathbf{n}|_{\Gamma_N} = 0$ being defined as in (2.5). Such a definiton is possible if the triangulation is made in such a way that there is no mesh cell across the interface between Γ_D and Γ_N on Γ. Having chosen V_h the approximation of the pressure space $L^2(\Omega)$ is then implicitly done

$$Q^h := \mathcal{P}_k^d(\mathcal{T}_h).$$

Consider the discrete problem:

Find $(\mathbf{u}^h, p^h) \in V^h \times Q^h$ such that

$$\begin{aligned} a(\mathbf{u}^h, \mathbf{v}^h) + b(\mathbf{v}^h, p^h) &= \langle \tilde{g}, \mathbf{v}^h \rangle \qquad \forall \mathbf{v}^h \in V^h \\ b(\mathbf{u}^h, q^h) &= \langle \tilde{f}, q^h \rangle \qquad \forall q^h \in Q^h \end{aligned} \tag{3.7}$$

where $\tilde{f}$ and $\tilde{g}$ eventually include non-homogeneous boundary conditions as in problem (2.11), that is,

$$\begin{aligned} \langle \tilde{g}, \mathbf{v}^h \rangle &= -\langle \mathbf{v}^h \cdot \mathbf{n}, p_D \rangle - a(\tilde{\mathbf{u}}, \mathbf{v}^h) \\ \langle \tilde{f}, \mathbf{v}^h \rangle &= -\langle f, q^h \rangle - b(\tilde{\mathbf{u}}, q^h). \end{aligned}$$

In section 2.2 it was shown that $b(\cdot, \cdot)$ satisfies the continuous inf-sup condition (2.13) on $H_{0,\Gamma_N}(\operatorname{div}, \Omega) \times L^2(\Omega)$. The interpolation operators Π_h are uniformly bounded from $T \subset H(\operatorname{div}, \Omega) = V$ to V^h, i.e.

$$\|\Pi_h \mathbf{v}\|_{H(\operatorname{div},\Omega)} \leq C \|\mathbf{v}\|_T \quad \forall \mathbf{v} \in T \tag{3.8}$$

(since they are linear and continuous mappings $T \to V^h$), where $T = H(\mathrm{div}, \Omega) \cap \prod_{T \in \mathcal{T}_h} H^1(T)^n$. Using (3.5) and (3.8) one obtains from theorem 2.4 that the discrete inf-sup condition (2.16) is satisfied with a constant independent of h.

Equation (3.6) shows that $W^h \subset W$: Let $\mathbf{v}^h \in W^h \subset V^h$, from (3.6) it follows that $\nabla \cdot \mathbf{v}^h \in Q^h \subset Q$. From the definition (2.15) of W^h it follows that $b(\mathbf{v}^h, \nabla \cdot \mathbf{v}^h) = 0$ and thus $\mathbf{v}^h \in W$.

$W^h \subset W$ implies that coercivity of $a(\cdot, \cdot)$ on W^h follows directly from coercivity of $a(\cdot, \cdot)$ on W. The discrete problem (3.7) has a unique solution due to theorem 2.3.

The explicit construction of the operator Π_h requires regularity assumptions which do not hold for a general function in $V = H(\mathrm{div}, \Omega)$. But existence of the operator Π_h on a subspace $T \subset V$ verifying (2.17) and (2.18) for $\mathbf{v} \in T$ and the norm on the right hand side of (2.18) replaced by that of the space T is enough to obtain optimal error estimates.

Error estimates

Theorem 3.5. *If the solution $(\mathbf{u}, p)$ of problem (2.9) belongs to $H^m(\Omega)^n \times H^m(\Omega)$, $1 \leq m \leq k+1$, and if $(\mathbf{u}^h, p^h)$ is the solution of (3.7), then there exist constants C_1, C_2 depending on n, k, the regulartiy constant σ and the coefficient $\mathbb{K}$, such that*

$$\|\mathbf{u} - \mathbf{u}^h\|_{L^2(\Omega)} \leq C_1 h^m \|\nabla^m \mathbf{u}\|_{L^2(\Omega)} \tag{3.9}$$

$$\|p - p^h\|_{L^2(\Omega)} \leq C_2 h^m \Big(\|\nabla^m \mathbf{u}\|_{L^2(\Omega)} + \|\nabla^m p\|_{L^2(\Omega)} \Big). \tag{3.10}$$

Proof: [5], Theorem 3.8, Theorem 3.10.

3.1.3 $\mathcal{BDM}$ - spaces

The following spaces introduced by Brezzi, Douglas and Marini use different order approximations for each variable in order to reduce the degrees of freedom (thus reducing the computational cost) while preserving the same order of convergence for $\mathbf{u}$ provided by $\mathcal{RT}_k$ spaces.

In the following the local spaces for each variable $\mathbf{u}$ and p are defined. It can be checked that the degrees of freedom defning the spaces approximating the vector variable guarantee the continuity of the normal component and therefore the global spaces are subspaces of $H(\mathrm{div}, \Omega)$.

For $n = 2$, $k \geq 1$ and a triangle T define

$$\mathcal{BDM}_k(T) = (\mathcal{P}_k(T))^2$$

for the approximation of the vector variable. And let

$$\mathcal{P}_{k-1}(T)$$

be the corresponding space for the scalar variable. Observe that

$$\dim \mathcal{BDM}_k(T) = (k+1)(k+2).$$

For example $\dim\mathcal{BDM}_1(T) = 6$ and $\dim\mathcal{BDM}_2(T) = 12$. Figure 3.2 shows the degrees of freedom for these two spaces. The arrows correspond to degrees of freedom of normal components while the circles indicate the internal degrees of freedom corresponding to the second and third conditions in the definition of Π_T below.

The parametrization of the triangle T with a convex combination of $\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3$, the corners of T, reads as follows

$$T = \{\mathbf{x} \in \mathbb{R}^2 \,:\, \mathbf{x} = \lambda_1\mathbf{a}_1 + \lambda_2\mathbf{a}_2 + \lambda_3\mathbf{a}_3,\ 0 \le \lambda_1, \lambda_2, \lambda_3 \le 1,\ \lambda_1 + \lambda_2 + \lambda_3 = 1\}.$$

The coefficients $\lambda_1, \lambda_2, \lambda_3$ are called the barycentric coordinates of $\mathbf{x} \in T$. The function $b_T = \lambda_1\lambda_2\lambda_3$ is called "bubble" function. For $\phi \in H^1(\Omega)$

$$\mathbf{curl}\phi = (\frac{\partial\phi}{\partial y}, -\frac{\partial\phi}{\partial x}).$$

Let $l_i, i = 1, 2, 3$ be the sides of T. The interpolation operator Π_T is defined as follows:

$$\int_{l_i} \Pi_T\mathbf{v} \cdot \mathbf{n}_i\, p_k\, ds = \int_{l_i} \mathbf{v} \cdot \mathbf{n}_i\, p_k\, ds \quad \forall p_k \in \mathcal{P}_k(F_i), i = 1, 2, 3$$

$$\int_T \Pi_T\mathbf{v} \cdot \nabla p_{k-1}\, d\mathbf{x} = \int_T \mathbf{v} \cdot \nabla p_{k-1}\, d\mathbf{x} \quad \forall p_{k-1} \in \mathcal{P}_{k-1}(T)$$

and when $k \ge 2$

$$\int_T \Pi_T\mathbf{v} \cdot \mathbf{curl}(b_T p_{k-2})\, d\mathbf{x} = \int_T \mathbf{v} \cdot \mathbf{curl}(b_T p_{k-2})\, d\mathbf{x} \quad \forall p_{k-2} \in \mathcal{P}_{k-2}(T)$$

Figure 3.2: Degrees of freedom for $\mathcal{B}DM_1$ and $\mathcal{B}DM_2$ in $\mathbb{R}^2$

One can check that property (3.5) follows from the definition of Π_T and proof of its existence is similiar to that of Lemma 3.2. The same error estimate for the approximation of $\mathbf{u}$ holds. For p the best order of convergence is reduced by one with respect to the estimate obtained for the Raviart-Thomas approximation, see [5].

3.2 Approximations of $H(\mathbf{div}, \Omega)$ on rectangular grids

First spaces introduced by Raviart and Thomas are defined. For nonnegative integers k, m the space of polynomials of the form

$$q(x, y) = \sum_{i=1}^{k} \sum_{j=1}^{m} a_{ij} x^i y^j$$

is denoted by $\mathcal{Q}_{k,m}$, then the Raviart-Thomas space on a rectangle R is given by

$$\mathcal{RT}_k(R) = \mathcal{Q}_{k+1,k}(R) \times \mathcal{Q}_{k,k+1}(R)$$

and the space for the scalar variable is $\mathcal{Q}_{k,k}(R)$. It can be checked that

$$\dim \mathcal{RT}_k(R) = 2(k+1)(k+2).$$

Figure 3.3 shows the degrees of freedom for $k = 0$ and $k = 1$. Denoting with l_i, $i = 1, 2, 3, 4$ the four sides of R, the degrees of freedom defining the operator Π_T for this case are

$$\int_{l_i} \Pi_T \mathbf{v} \cdot \mathbf{n}_i \, p_k \, ds = \int_{l_i} \mathbf{v} \cdot \mathbf{n}_i \, p_k \, ds \quad \forall p_k \in \mathcal{P}_k(l_i), i = 1, 2, 3, 4$$

and for $k \geq 1$

$$\int_R \Pi_T \mathbf{v} \cdot \phi_k \, d\mathbf{x} = \int_R \mathbf{v} \cdot \phi_k \, d\mathbf{x} \quad \forall \phi_k \in \mathcal{Q}_{k-1,k}(R) \times \mathcal{Q}_{k,k-1}(R).$$

Figure 3.3: Degrees of freedom for $\mathcal{RT}_0(R)$ and $\mathcal{RT}_1(R)$

Spaces introduced by Brezzi, Douglas and Marini on rectangular elements are defined for $k \geq 1$ as

$$\mathcal{BDM}_k(R) = (\mathcal{P}_k(R))^2 + \langle \mathbf{curl}(x^{k+1} y) \rangle + \langle \mathbf{curl}(x y^{k+1}) \rangle$$

and the associated scalar space is $\mathcal{P}_{k-1}(R)$. It can be checked that

$$\dim \mathcal{BDM}_k(R) = (k+1)(k+2) + 2.$$

The degrees of freedom for $k = 1$ and $k = 2$ are shown in fig. 3.4.

The operator Π_T is defned by

$$\int_{F_i} \Pi_T \mathbf{v} \cdot \mathbf{n}_i \, p_k \, ds = \int_{F_i} \mathbf{v} \cdot \mathbf{n}_i \, p_k \, ds \quad \forall p_k \in \mathcal{P}_k(l_i), i = 1, 2, 3, 4$$

and for $k \geq 2$

$$\int_R \Pi_T \mathbf{v} \cdot \mathbf{p}_{k-2} \, d\mathbf{x} = \int_R \mathbf{v} \cdot \mathbf{p}_{k-2} \, d\mathbf{x} \quad \forall \mathbf{p}_{k-2} \in (\mathcal{P}_{k-2}(R))^2.$$

Figure 3.4: Degrees of freedom for $\mathcal{B}DM_1(R)$ and $\mathcal{B}DM_2(R)$

The $\mathcal{RT}_k$ as well as the $\mathcal{BDM}_k$ spaces on rectangles have analogous properties to those on triangles. Therefore the same error estimates obtained for triangular grids are valid in both cases.

More generally, one can consider quadrilateral grids. Given a convex quadrilateral Q, the spaces are defined using the Piola transform from a reference rectangle R to Q. Define for example the Raviart-Thomas spaces $\mathcal{RT}_k(Q)$.

Let $R = [0, 1] \times [0, 1]$ be the reference rectangle and $F : R \to Q$ a bilinear transformation taking the vertices of R into the vertices of Q. Then define the local space $\mathcal{RT}_k(Q)$ by using the Piola transform, i.e. if $\mathbf{x} = F(\hat{\mathbf{x}})$, DF is the Jacobian matrix of F and $J = |\text{det}DF|$,

$$\mathcal{RT}_k(Q) = \{\mathbf{v} : Q \to \mathbb{R}^2 : \mathbf{v}(\mathbf{x}) = \frac{1}{J(\hat{\mathbf{x}})} DF(\hat{\mathbf{x}})\hat{\mathbf{v}}(\hat{\mathbf{x}}) \quad \text{with } \hat{\mathbf{v}} \in \mathcal{RT}_k(R)\}.$$

Similar error estimates to those obtained for triangular elements can be proved under appropriate regularity assumptions on the quadrilaterals.

3d extensions of the spaces defined above: For tetrahedral grids the spaces are defined in an analogous way, the construction of the interpolation operator Π_T requires a different analysis. In the case of 3d rectangular grids, the extensions of $\mathcal{RT}_k$ are defined in an analogous way and the extensions of $\mathcal{BDM}_k$ can be defined for a 3d rectangle R by

$$\begin{aligned}\mathcal{BDDF}_k(R) = \mathcal{P}_k^3 &+ \langle\{\mathbf{curl}(0, 0, xy^{i+1}z^{k-i}), i = 0, ...k\}\rangle \\ &+ \langle\{\mathbf{curl}(0, x^{k-i}yz^{i+1}, 0), i = 0, ...k\}\rangle \\ &+ \langle\{\mathbf{curl}(x^{i+1}y^{k-i}z, 0, 0), i = 0, ...k\}\rangle\end{aligned}$$

where now the notation $\mathbf{curl}\,\mathbf{v}$ is used for the rotational of a three dimensional vector field $\mathbf{v}$. All the convergence results obtained in 2d can be extended for the 3d spaces mentioned here.

4 Solvers for Linear Saddle Point Problems

Finite element discretizations of the Darcy problem, (2.1) - (2.4), lead to linear systems of equations in saddle point form

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \cdot \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix} = \begin{pmatrix} \underline{g} \\ \underline{f} \end{pmatrix}, \quad \mathcal{A} = \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \tag{4.1}$$

with

$$A \in \mathbb{R}^{n_V \times n_V}, B \in \mathbb{R}^{n_Q \times n_V}, \qquad \underline{u}, \underline{g} \in \mathbb{R}^{n_V}, \underline{p}, \underline{f} \in \mathbb{R}^{n_Q},$$

where n_V is the number of velocity degrees of freedom and n_Q is the number of pressure degrees of freedom. The matrix A represents the discrete analog of the linear operator 'multiplication by $\mathbb{K}$', i.e. $(A)_{ij} = \int_\Omega \mathbb{K}\boldsymbol{\varphi}_j \cdot \boldsymbol{\varphi}_i \, d\mathbf{x}, i, j = 1, ..., n_V$ is a weighted mass matrix. B and B^T are matrix representations of discrete analogs of the negative divergence operator and the gradient operator. The assumptions on $\mathbb{K}$ being symmetric and bounded from below imply that A is symmetric positive definite. Therefore the saddle point matrix $\mathcal{A} \in \mathbb{R}^{(n_V+n_Q)\times(n_V+n_Q)}$ of system (4.1) is symmetric and indefinite. All matrix blocks are sparse such that $\mathcal{A}$ is sparse too.

Thus efficient simulations of flow problems in porous media, using inf-sup stable finite element spaces, require the efficient solution of sparse linear saddle point problems (4.1).

The subsequent section follows [2] and investigates properties of the saddle point matrix $\mathcal{A}$ such as solvability condition, spectral properties and conditioning. Preconditioners, that are used to accelerate the speed of convergence of iterative solvers for linear saddle point problems (4.1), are presented afterwards following [1], [6], [12] and [13].

4.1 Properties of Saddle Point Matrices

Solvability Condition If A is nonsingular, then the saddle point matrix $\mathcal{A}$ admits the following block triangular factorization

$$\mathcal{A} = \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} = \begin{pmatrix} I & 0 \\ BA^{-1} & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & -BA^{-1}B^T \end{pmatrix} \begin{pmatrix} I & A^{-1}B^T \\ 0 & I \end{pmatrix}. \tag{4.2}$$

The matrix

$$S := -BA^{-1}B^T$$

is called Schur complement of $\mathcal{A}$. It follows from the block factorization (4.2) that $\mathcal{A}$ is nonsingular, if and only if S is, because then all factors in (4.2) are nonsingular.

We consider the case where A is symmetric positve definite. Then A^{-1} is also symmetric positve definite and the matrix $BA^{-1}B^T$ is symmetric positiv semidefinite. Indeed, one has

$$\underline{x}^T BA^{-1}B^T\underline{x} = (B^T\underline{x})^T A^{-1}(B^T\underline{x}) \geq \kappa_0 |B^T\underline{x}|^2$$

and one can see that $BA^{-1}B^T$ is symmetric positive definite if and only if $\ker B^T = \{0\}$ (hence, if and only if B^T has full column rank, i.e. $\text{rank}(B) = n_Q$).

The Schur complement S, and thus $\mathcal{A}$, is invertible if and only if B^T has full column rank, since in this case the Schur complement S is symmetric negative definite. Then problem (4.1) has a unique solution.

It is proved in [12] that B has full rank if and only if

$$\exists \beta > 0 : \qquad \inf_{\underline{q}\in\mathbb{R}^{n_Q}\setminus\{0\}} \sup_{\underline{v}\in\mathbb{R}^{n_V}\setminus\{0\}} \frac{\underline{v}^T B^T \underline{q}}{|\underline{v}||\underline{q}|} \geq \beta.$$

In the finite element context the nonsingularity of $\mathcal{A}$ is not sufficient to ensure meaningful computed solutions. It is important that the used finite element spaces satisfy the discrete inf-sup condition (2.16) with a parameter $\beta^* > 0$ that does not depend on the mesh parameter h. This is the case since the inverse of the discrete inf-sup parameter β^* enters the finite element error estimates. The error bounds depend on inverse of powers of β^*. Thus, a behavior of the form $\beta^* \to 0$ for successive refinements ($h \to 0$) leads to a deterioration of the order of convergence in the error estimates.

Spectral Properties Spectral properties of saddle point matrices are relevant when solving the linear system of equations (4.1) by iterative methods. The following result from Rusten and Winther (1992) establishes eigenvalue bounds for the considered class of saddle point matrices.

Theorem 4.1. *(Rusten and Winther, 1992, [15]. Eigenvalue bounds for the symmetric case) Assume A is symmetric positive definite and B has full rank. Let μ_1 and μ_n denote the largest and smallest eigenvalues of A and let σ_1 and σ_m denote the largest and smallest singular values of B. Let $\sigma(\mathcal{A})$ denote the spectrum of $\mathcal{A}$. Then*

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

where

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

and

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

These bounds can be used to obtain estimates for the condition number of $\mathcal{A}$. In turn, these estimates can be used to predict the rate of convergence of iterative methods.

Conditioning Issues Saddle point systems that arise in practice can be very poorly conditioned. In some cases the special structure of the saddle point matrix $\mathcal{A}$ can be exploited to avoid or mitigate the effect of ill-conditioning.

Consider a saddle point problem where A is symmetric positive definite and B has full rank. In this case $\mathcal{A}$ is symmetric and its spectral condition number is given by

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

From the previous Theorem 4.1 one can see that the condition number of $\mathcal{A}$ grows unboundedly as either $\mu_n = \lambda_{min}(A)$ or $\sigma_m = \sigma_{min}(B)$ goes to zero (assuming that $\lambda_{max}(A)$ and $\sigma_{max}(B)$ are kept constant). This growth of the condition number of $\mathcal{A}$ means that the rate of convergence of most iterative solvers (like Krylov subspace methods) deteriorates as the problem size increases. Preconditioning may be used to reduce or even eliminate this dependency on h in many cases.

4.2 Preconditioners for Iterative Solvers

The performance of iterative solvers for algebraic linear saddle point problems (4.1) can be improved if preconditioners are used. Preconditioners are approximations of $\mathcal{A}^{-1}$ that can be computed comparatively efficiently. These approximations might be a fixed matrix or an iterative method or a combination of both.

One distinguishes between left and right preconditioners. Denote the preconditioner by $\mathcal{M}^{-1}$. Then, for left preconditioning, one considers instead of $\mathcal{A}\underline{x} = \underline{b}$ the system

$$\mathcal{M}^{-1}\mathcal{A}\underline{x} = \mathcal{M}^{-1}\underline{b}$$

in the iterative solver and for right preconditioning the problem

$$\mathcal{A}\mathcal{M}^{-1}\underline{y} = \underline{b}, \quad \underline{x} = \mathcal{M}^{-1}\underline{y}.$$

The preconditioner $\mathcal{M}^{-1}$ should satisfy two requirements:

- The convergence of the iterative method for the preconditioned system with $\mathcal{M}^{-1}\mathcal{A}$ or $\mathcal{A}\mathcal{M}^{-1}$ should be faster than for the original system with $\mathcal{A}$. That means, $\mathcal{M}^{-1}$ should be good approximation to $\mathcal{A}^{-1}$.
- The action of $\mathcal{M}^{-1}$ should be inexpensive.

In general, one has to find a compromise between these two requirements.

The Krylov subspace methods compute the solution of (4.1) in at most $n_V + n_Q$ iterations (in exact arithmetic) by construction. However, this property is useless if $n_V + n_Q$ is large. When the linear system represents a discretized partial differential equation, then an approximate solution $\underline{x}^k$ of the linear system with an error norm on the level of the discretization error is often sufficient. Once this error level is reached, the

iterative method can be stopped. The question is how fast can a given Krylov subspace method reach a given accuracy level. Starting point for the convergence analysis of Krylov subspace methods based on the minimization of the residual is the interpretation of the kth residual in terms of the initial residual multiplied by a certain polynomial in the matrix $\mathcal{A}$,

$$\|\underline{b} - \mathcal{A}\underline{x}^k\|_2 = \|\underline{r}^k\|_2 = \min_{\underline{z} \in \underline{x}^0 + K_k(\underline{r}^0, \mathcal{A})} \|\underline{b} - \mathcal{A}\underline{z}\|_2 = \min_{p_k \in \mathcal{P}_k, p_k(0)=1} \|p_k(\mathcal{A})\underline{r}^0\|_2.$$

If $\mathcal{A}$ is normal then

$$\|\underline{r}^k\|_2 \leq \|\underline{r}^0\|_2 \min_{p_k \in \mathcal{P}_k, p_k(0)=1} \|p_k(\mathcal{A})\|_2 = \|\underline{r}^0\|_2 \min_{p_k \in \mathcal{P}_k, p_k(0)=1} \max_{\lambda \text{ is eigenvalue of } \mathcal{A}} |p_k(\lambda)|. \quad (4.3)$$

Hence $\|\underline{r}^k\|_2 / \|\underline{r}^0\|_2$, the kth relative Euclidean residual norm, is bounded by the value of a polynomial approximation problem on the eigenvalues of $\mathcal{A}$. This bound provides some intuition of how the eigenvalue distribution influences the worst-case convergence behavior of minimal residual methods. For example, a single eigenvalue cluster far away from the origin implies fast convergence (measured by the relative Euclidean residual norm).

If $\mathcal{A}$ is symmetric positive definite, the standard approach for estimating the right-hand side of (4.3) is to replace the min-max problem on the discrete set of eigenvalues by a min-max approximation problem on its convex hull (i.e., on an interval from the smallest eigenvalue $\lambda_{\min}$ to the largest eigenvalue $\lambda_{\max}$ of $\mathcal{A}$). The latter is solved by scaled and shifted Chebyshev polynomials of the first kind, giving the bound

$$\min_{p_k \in \mathcal{P}_k, p_k(0)=1} \max_{\lambda \text{ is eigenvalue of } \mathcal{A}} |p_k(\lambda)| \leq 2\Big(\frac{\sqrt{\kappa(\mathcal{A})} - 1}{\sqrt{\kappa(\mathcal{A})} + 1}\Big)^k, \text{ where } \kappa(\mathcal{A}) = \frac{\lambda_{\max}}{\lambda_{\min}}. \quad (4.4)$$

The bounds (4.3)-(4.4) show that in this case a small condition number of $\mathcal{A}$ is sufficient (but not necessary) for fast convergence.

In the case of a nonsingular symmetric indefinite matrix $\mathcal{A}$, the min-max approximation problem on the matrix eigenvalues in (4.3) cannot be replaced by the min-max problem on their convex hull, as eigenvalues lie on both sides of the origin. Here one may replace the discrete set of eigenvalues by the union of two intervals containing all of them and excluding the origin, say $I^- \cup I^+ \equiv [\lambda_{\min}, \lambda_s] \cup [\lambda_{s+1}, \lambda_{\max}]$ with $\lambda_{\min} \leq \lambda_s < 0 < \lambda_{s+1} \leq \lambda_{\max}$.

When both intervals are of the same length, i.e. $\lambda_{\max} - \lambda_{s+1} = \lambda_s - \lambda_{\min}$, the solution of the corresponding min-max approximation problem

$$\min_{p_k \in \mathcal{P}_k, p_k(0)=1} \max_{\lambda \in I^- \cup I^+} |p_k(\lambda)| \quad (4.5)$$

leads to the bound

$$\min_{p_k \in \mathcal{P}_k, p_k(0)=1} \max_{\lambda \text{ is eigenvalue of } \mathcal{A}} |p_k(\lambda)| \leq 2\Big(\frac{\sqrt{|\lambda_{\min}\lambda_{\max}|} - \sqrt{|\lambda_s\lambda_{s+1}|}}{\sqrt{|\lambda_{\min}\lambda_{\max}|} + \sqrt{|\lambda_s\lambda_{s+1}|}}\Big)^{[k/2]}, \quad (4.6)$$

where $[k/2]$ denotes the integer part of $k/2$. For an illustration of this bound suppose that $|\lambda_{\min}| = \lambda_{\max} = 1$ and $|\lambda_s| = \lambda_{s+1}$. Then $\kappa(\mathcal{A}) = \lambda_{s+1}^{-1}$ and the right-hand side of (4.6) reduces to

$$2\Big(\frac{1/\lambda_{s+1} - 1}{1/\lambda_{s+1} + 1}\Big)^{[k/2]}. \tag{4.7}$$

Note that (4.7) corresponds to the value of the right-hand side of (4.4) at step $[k/2]$ for a symmetric positive definite matrix having all its eigenvalues in the interval $[\lambda_{s+1}^2, 1]$, and thus a condition number of λ_{s+1}^{-2}. Hence the convergence bound for an indefinite matrix with condition number κ needs twice as many steps to decrease to the value of the bound for a definite matrix with condition number κ^2. This indicates that solving indefinite problems represents a significant challenge.

In the general case when the two intervals are not of the same length, the explicit solution of (4.5) becomes quite complicated and no simple and explicit bound on the min-max value is known. An alternative to give relevant information about the actual convergence behavior is to consider the asymptotic behavior of the min-max value (4.5), and in particular the asymptotic convergence factor

$$\rho(I^- \cup I^+) \equiv \lim_{k\to\infty} \Big(\min_{p_k \in \mathcal{P}_k, p_k(0)=1} \max_{\lambda \in I^- \cup I^+} |p_k(\lambda)| \Big)^{\frac{1}{k}}.$$

Asymptotic convergence results can be useful in the convergence analysis of minimal residual methods for sequences of linear systems of growing dimension, e.g., when studying the dependence of the convergence behavior on the mesh size in a discretized differential equation.

4.2.1 Least Squares Commutator (LSC) Preconditioner

The straightforward application of many standard schemes for the solution or preconditioning of linear systems of equations becomes difficult because of the block structure of the saddle point matrix $\mathcal{A}$ from (4.1). For this reason preconditioners have been developed where individual systems of equations for the pressure and for each component of the velocity have to be solved. These individual systems do not possess a block structure, which enables the application of standard solvers and preconditioners. The LSC preconditioner belongs to this class of methods that solve equations connected with the first and second rows of blocks separately.

The LSC preconditioner is derived from the LU decomposition of the matrix $\mathcal{A}$ and the approximation of the Schur complement by keeping a certain operator commutator error small.

Multiplying the second and third factor of (4.2) gives the LU decomposition

$$\mathcal{A} = \begin{pmatrix} I & 0 \\ BA^{-1} & I \end{pmatrix} \begin{pmatrix} A & B^T \\ 0 & S \end{pmatrix} = LU \tag{4.8}$$

where $S = -BA^{-1}B^T$ is the Schur complement of $\mathcal{A}$.

From (4.8) it follows that $\mathcal{A}U^{-1} = L$ which suggests to use the matrix U^{-1} as a right-oriented preconditioner, since the preconditioned matrix L has perfectly clustered eigenvalues.

Consider the eigenvalue problem for the preconditioned system with the right-oriented preconditioner U^{-1},

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} A & B^T \\ 0 & S \end{pmatrix}^{-1} \begin{pmatrix} \underline{v} \\ \underline{q} \end{pmatrix} = \lambda \begin{pmatrix} \underline{v} \\ \underline{q} \end{pmatrix}.$$

Setting

$$\begin{pmatrix} A & B^T \\ 0 & S \end{pmatrix}^{-1} \begin{pmatrix} \underline{v} \\ \underline{q} \end{pmatrix} = \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix}$$

it follows

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix} = \lambda \begin{pmatrix} A & B^T \\ 0 & S \end{pmatrix} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix}. \tag{4.9}$$

From the first row of (4.9) one obtains

$$(1 - \lambda)(A\underline{u} + B^T\underline{p}) = 0.$$

There are two possibilites: $\lambda = 1$ and $A\underline{u} + B^T\underline{p} = 0$. In the first case $\lambda = 1$ is an eigenvalue of multiplicity n_V. For the second case, inserting

$$\underline{u} = -A^{-1}B^T\underline{p}$$

in the second block equation $B\underline{u} - \lambda S\underline{p} = 0$ gives

$$-BA^{-1}B^T\underline{p} = \lambda S\underline{p} \tag{4.10}$$

thus $\lambda = 1$ is an eigenvalue of multiplicity n_Q. The Schur complement is generally not explicitly available and is a dense matrix in case it is available, if A^{-1} is dense. Thus in practice it is not feasible to use the Schur complement as part of the preconditioner U^{-1}. But eq. (4.10) shows that a good approximation of the Schur complement will influence the good convergence of the preconditioned system with U^{-1}. The construction of an approximation to the Schur complement leads to the LSC preconditioner.

The basic idea is to search for a regular matrix $A_p \in \mathbb{R}^{n_Q \times n_Q}$ acting on coefficients of the pressure space such that

$$B^T A_p = AB^T. \tag{4.11}$$

Multiplying (4.11) with $-BA^{-1}$ from left and A_p^{-1} from right gives

$$-BA^{-1}B^T = -BB^T A_p^{-1}.$$

Using this form of the Schur complement when applying U^{-1} as a preconditioner requires approximating the action of $(-BB^TA_p^{-1})^{-1}$, which is more easily done, as A_p is known and BB^T is symmetric positive definite and represents a discretization of a pressure Poisson problem.

Since $B^T \in \mathbb{R}^{n_Q \times n_V}$, $n_V > n_Q$, is a full rank rectangular matrix, (4.11) is in general an overdetermined system and can only be solved in a minimizing sense

$$\min_{A_p} = \|AB^T - B^TA_p\| \tag{4.12}$$

for some matrix norm $\|\cdot\|$.

The appearing matrices in (4.12) are discrete counterparts of the underlying continuous operators from the Darcy equations. The matrix B^T stems from the finite element discretization of the gradient operator and A from the linear operator 'multiplication by $\mathbb{K}$'. The unknown matrix A_p is now assumed to originate from the discretization of the linear operator acting on the pressure space. The minimization problem can be interpreted as minimizing the discrete commutation error

$$AB^T - B^TA_p.$$

To support the interpretation one has to specify the concrete choice of the finite element spaces and introduce appropriate weights by multiplying with the inverses of the velocity and pressure mass matrices $M_v \in \mathbb{R}^{n_V \times n_V}$ and $M_p \in \mathbb{R}^{n_Q \times n_Q}$ and consider the minimization problem

$$\min_{A_p} \|M_v^{-1}AM_v^{-1}B^T - M_v^{-1}B^TM_p^{-1}A_p\|. \tag{4.13}$$

Multiplying from left with $BA^{-1}M_v$ and from right with $A_p^{-1}M_p$ the term inside the norm gives a formula for the approximation of the Schur complement

$$S = -BA^{-1}B^T \approx -BM_v^{-1}B^TA_p^{-1}M_p. \tag{4.14}$$

Specifying the minimization problem (4.13) as minimizing columnwise in a M_v-weighted vector norm

$$\|\mathbf{v}\|_{M_v} = \langle M_v\mathbf{v}, \mathbf{v}\rangle^{\frac{1}{2}}$$

leads to the least squares problems

$$\min_{[a_p]_j} \|[M_v^{-1}AM_v^{-1}B^T]_j - M_v^{-1}B^TM_p^{-1}[a_p]_j\|_{M_v} \quad j = 1, ..., n_Q, \tag{4.15}$$

where the unknowns $[a_p]_j$ are the columns of A_p. The first order optimality conditions read

$$M_p^{-1}BM_v^{-1}B^TM_p^{-1}[a_p]_j = [M_p^{-1}BM_v^{-1}AM_v^{-1}B^T]_j \quad j = 1, ..., n_Q$$

which leads to the following representation of A_p,

$$A_p = M_p(BM_v^{-1}B^T)^{-1}(BM_v^{-1}AM_v^{-1}B^T). \tag{4.16}$$

Inserting this expression in (4.14) gives an approximation of the Schur complement

$$S \approx -(BM_v^{-1}B^T)(BM_v^{-1}AM_v^{-1}B^T)^{-1}(BM_v^{-1}B^T). \tag{4.17}$$

It is not practical to work with M_v^{-1} in (4.17), since it is a dense matrix. A practical algorithm is obtained by replacing M_v with $\mathrm{diag}(M_v) = D_v$ from which the sparse discrete Laplacian $BD_v^{-1}B^T$ can be constructed. The LSC preconditioner is obtained by replacing M_v^{-1} with $(\mathrm{diag}(M_v))^{-1} = D_v^{-1}$ everywhere in (4.17)

$$S_{LSC} := -(BD_v^{-1}B^T)(BD_v^{-1}AD_v^{-1}B^T)^{-1}(BD_v^{-1}B^T). \tag{4.18}$$

The application of the LSC preconditioner requires to solve as preconditioning step a problem of the form

$$\begin{pmatrix} A & B^T \\ 0 & S_{LSC} \end{pmatrix} \begin{pmatrix} \underline{v} \\ \underline{q} \end{pmatrix} = \begin{pmatrix} \underline{b}_v \\ \underline{b}_q \end{pmatrix}$$

for a given vector $(\underline{b}_v, \underline{b}_q)^T$. In the first step, one solves

$$S_{LSC}\underline{q} = \underline{b}_q$$

which requires the solution of two discrete Poisson-type problems for the pressure with the same matrix $BD_v^{-1}B^T$ and matrix-vector products with the matrices B, B^T, A and (the diagonal matrix) D_v^{-1} since

$$S_{LSC}^{-1} := -(BD_v^{-1}B^T)^{-1}(BD_v^{-1}AD_v^{-1}B^T)(BD_v^{-1}B^T)^{-1}.$$

After having computed $\underline{q}$ one finds $\underline{v}$ by solving

$$A\underline{v} = \underline{b}_v - B^T\underline{q}$$

which requires the solution of a problem for the velocity unknowns.

4.2.2 Vanka Preconditioner

Standart preconditioners like Jacobi method or SOR method cannot be applied for systems of type (4.1) because of the zero block in the diagonal of $\mathcal{A}$ and special solvers need to be designed. A popular class of iterative solvers or preconditioners for problems of type (4.1) are Vanka-type solvers. They can be understood as block Gauss-Seidel methods.

Let $\mathcal{V}^h$ and $\mathcal{Q}^h$ be the set of velocity and pressure degrees of freedom, respectively. These sets are decomposed into

$$\mathcal{V}^h = \cup_{j=1}^J \mathcal{V}_j^h, \quad \mathcal{Q}^h = \cup_{j=1}^J \mathcal{Q}_j^h. \tag{4.19}$$

These subsets are not required to be disjoint. Let $\mathcal{A}_j$ be the block of the matrix $\mathcal{A}$ that contains those entries of $\mathcal{A}$ whose indices belong to $\mathcal{W}_j^h = \mathcal{V}_j^h \cup \mathcal{Q}_j^h$, i.e. the intersection of rows and columns of $\mathcal{A}$ with the global indices belonging to $\mathcal{W}_j^h$,

$$\mathcal{A}_j = \begin{pmatrix} A_j & B_j^T \\ B_j & 0 \end{pmatrix} \in \mathbb{R}^{\dim\mathcal{W}_j^h \times \dim\mathcal{W}_j^h}.$$

Similarly, denote by $(\cdot)_j$ the restriction of a vector to the rows corresponding to the degrees of freedom in $\mathcal{W}_j^h$. Each preconditioning step with a Vanka-type preconditioner consists in a loop over the sets $\mathcal{W}_j^h, j = 1, ..., J$, where for each $\mathcal{W}_j^h$ a local system of equations connteced with the degrees of freedom in this set is solved. The local solutions are updated in a Gauss-Seidel manner. The Vanka preconditioner computes new velocity and pressure values by

$$\begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix}_j := \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix}_j + \mathcal{A}_j^{-1} \left(\begin{pmatrix} \underline{g} \\ \underline{f} \end{pmatrix} - \mathcal{A} \begin{pmatrix} \underline{u} \\ \underline{p} \end{pmatrix} \right)_j$$

The local systems of equations are usually solved with a direct solver.

A general strategy for choosing the sets $\mathcal{V}_j^h$ and $\mathcal{Q}_j^h$ is as follows:

- First, pick some pressure degrees of freedom that define $\mathcal{Q}_j^h$.
- Second, $\mathcal{V}_j^h$ is formed by all velocity degrees of freedom that are connected with the pressure degrees of freedom from $\mathcal{Q}_j^h$ by non-zero entries in the matrix B.

With this strategy a Vanka-type preconditioner is determined by the particular choice of the sets $\mathcal{Q}_j^h$, $j = 1, ..., J$.

For discontinuous pressure approximation consider the mesh-cell-oriented Vanka preconditioner. This preconditioner takes for $\mathcal{Q}_j^h$ all pressure degrees of freedom that belong to one mesh cell. Then, the corresponding velocity degrees of freedom are all those which belong to the same mesh cell. The number of local systems J to be solved in one preconditioning step then equals the number of cells in the mesh $\mathcal{T}_h$ and all local systems are of the same size.

Damping of the Iterate Sometimes it is beneficial to damp the iterate of the Vanka preconditioner. Let $(\underline{u}, \underline{p})$ be the current iterate and $(\delta\underline{u}, \delta\underline{p})$ the update computed by one iteration of the preconditioner. Then the new iterate is computed by $(\underline{u}, \underline{p}) + \omega(\delta\underline{u}, \delta\underline{p})$, where $\omega \in \mathbb{R}$, $\omega > 0$.

4.2.3 Incomplete LU Factorization

In the application of direct solvers for linear systems of equations with a sparse matrix the additional fill-in that occurs, if a factorization of the matrix, like the LU factorization is computed, appears to be a drawback. Applying a LU factorization to a sparse matirx A, the factors L and U are generally considerably denser than A.

A main part of direct solvers for sparse linear systems, like UMFPACK, [4] is a reordering of the unknowns such that the fill-in is reduced.

In the context of preconditioning, the LU factorization can be modified such that it respects the sparsity pattern or zero pattern of the matrix A, i.e. L is stored with a sparsity pattern that corresponds to the strict lower triangle of A and U with a sparsity pattern that corresponds to the upper triangle of A. Performing the algorithm of the

standard LU factorization (without pivoting), one neglects all entries that do not fit in the prescribed pattern. In this way one obtains an incomplete LU (ILU) factorization

$$A = LU + E,$$

with the error matrix E.

Avoiding pivoting in the application of the ILU factorization of $\mathcal{A}$ does not lead necessarily to a failure of this method. Considering a system of the form (4.1), the degrees of freedom are ordered in a way that the pressure degrees of freedom come last and the entries of the (2,2) block of $\mathcal{A}$ are zero. Under these conditions a division by zero does not occur because the zero entries at the main diagonal vanish during the incomplete factorization.

The application of ILU as preconditioner requires the solution of two sparse linear systems of equations with triangular matrices:

1. solve the lower triangular system $L\underline{w} = \underline{r}$,
2. solve the upper triangular system $U\underline{z} = \underline{w}$.

For using ILU as preconditioner, it is essential that the diagonal entries do not belong to the prescribed zero pattern since a linear system of equations with matrix U has to be solved.

The main costs of unpreconditioned iterative methods are the multiplication of the sparse matrix with a vector. If the zero pattern is appropriately given, then the costs for applying the ILU preconditioner are proportional to the costs of the matrix-vector multiplication.

5 Numerical Studies

In the following numerical studies of different solvers for the Darcy equations in 2d and 3d are presented. The studies were performed on problems involving singularities (the five-spot problem) and discontinuous coefficients associated with a checkerboard domain. The considered examples can be found in [10] and [11].

The simulations were performed with the finite element code ParMooN, [18]. For discretizing the Darcy equations the Galerkin finite element method with inf-sup stable pairs of finite element spaces was used. As Krylov subspace method, the flexible GMRES (FGMRES) method is used for the iterative solution of the arising algebraic linear saddle point systems. The purpose of the computational analysis is to compare the performance of the following solvers:

- UMFPACK, sparse direct solver [4],
- FGMRES with least squares commutator (LSC) preconditioner with direct solver UMFPACK to solve the Poisson subproblems and velocity systems,
- FGMRES with Vanka preconditioner with direct solver LAPACK to solve the local systems of equations,
- FGMRES with LSC preconditioner provided by PETSc, [19]-[20], a library providing iterative solvers together with preconditioners for linear saddle point problems, with direct solver UMFPACK for all linear subsystems,
- FGMRES with Euclid preconditioner provided by PETSc.

The Euclid preconditioner is a scalable implementation of the Parallel ILU (incomplete LU) algorithm, available in hypre, [17], a software library of high performance preconditioners and solvers for the solution of large, sparse linear systems of equations, which PETSc uses. Scalable means that the factorization (setup) and application (triangular solve) timings remain nearly constant when the global problem size is scaled in proportion to the number of processors. As with all ILU preconditioning methods, the number of iterations is expected to increase with global problem size. Experimental results have shown that PILU preconditioning is in general more effective than Block Jacobi preconditioning for minimizing total solution time. For scaled problems, the relative advantage appears to increase as the number of processors is scaled upwards. For details see [17]. For the Vanka preconditioner a damping of the update is applied. The damping parameter was set to $1,5$. The iteration of FGMRES is terminated when the maximum number of iterations is achieved (it has been set to 5000). FGMRES is restarted after 20 iterations and stopped when the Euclidean norm of the residual vector was smaller than 10^{-8}.

5.1 Analytic Example

First consider a problem where the exact solution is known. The domain under consideration is $\Omega = [0,1] \times [0,1]$ and the exact pressure solution is given by

$$p = \sin 2\pi x \sin 2\pi y.$$

The velocity field is computed from Darcy's law, eq. (2.1), with permeability tensor $\mathbb{K} = \mathbb{I}$,

$$\mathbf{u} = -\nabla p = \begin{pmatrix} -2\pi \cos(2\pi x) \sin(2\pi y) \\ -2\pi \sin(2\pi x) \cos(2\pi y) \end{pmatrix}.$$

the source term f is calculated from (2.2) by taking the divergence of the velocity field,

$$f = \nabla \cdot \mathbf{u} = \partial_x u_1 + \partial_y u_2 = 8\pi^2 \sin(2\pi x) \sin(2\pi y)$$

and the boundary data u_N is calculated by taking its normal components,

$$\begin{aligned} \mathbf{u} \cdot \mathbf{n}_1 &= u_1(1,y) = -2\pi \sin(2\pi y) && \text{on } \{1\} \times [0,1] \\ \mathbf{u} \cdot (-\mathbf{n}_1) &= -u_1(0,y) = 2\pi \sin(2\pi y) && \text{on } \{0\} \times [0,1] \\ \mathbf{u} \cdot \mathbf{n}_2 &= u_2(x,1) = -2\pi \sin(2\pi x) && \text{on } [0,1] \times \{1\} \\ \mathbf{u} \cdot (-\mathbf{n}_2) &= -u_2(x,0) = 2\pi \sin(2\pi x) && \text{on } [0,1] \times \{0\} \end{aligned}$$

where $\mathbf{n}_1 = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ and $\mathbf{n}_2 = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$.

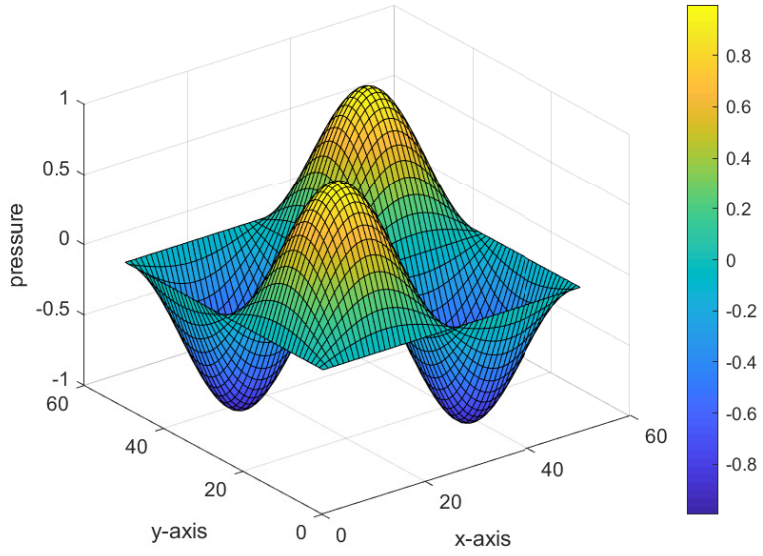

Figure 5.1: Analytic example. Elevation plot of the exact pressure field.

For a numerical simulation of the described boundary value problem the standard Galerkin method was used with the velocity finite element spaces $\mathcal{RT}_k$ and $\mathcal{BDM}_k$ and corresponding pressure finite element spaces $\mathcal{P}_k^d$ on uniform triangular and quadrilateral grids. Table 5.1 and fig. 5.2 illustrate the number of cells and the number of degrees of freedom on quadrilateral grids obtained by uniform refinement of the domain.

Table 5.1: Number of cells and number of degrees of freedom on quadrilateral grids.

grid level	number of cells	number of degrees of freedom						
		$\mathcal{BDM}_1$	$\mathcal{BDM}_2$	$\mathcal{BDM}_3$	$\mathcal{RT}_0$	$\mathcal{RT}_1$	$\mathcal{RT}_2$	$\mathcal{RT}_3$
0	64	352	752	1344	208	800	1776	3136
1	256	1344	2912	5248	800	3136	7008	12416
2	1024	5248	11456	20736	3136	12416	27840	49408
3	4096	20736	45440	82432	12416	49408	110976	197120

Number of degrees of freedom (# dof) on the used quadrilateral grids

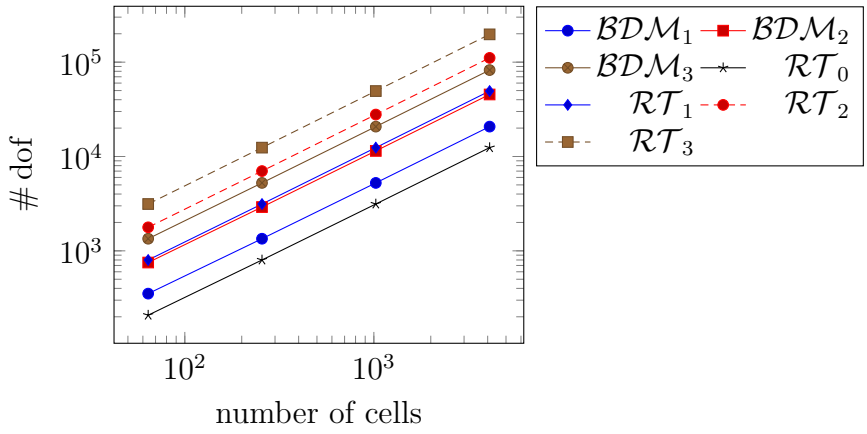

Figure 5.2: Plot of number of degrees of freedom on quadrilateral grids

Results on triangular (right column) and quadrilateral (left column) grids are presented in fig. 5.4.

At first glance it can be seen that the direct solver UMFPACK and PETSc LSC preconditioner performed best and Vanka and LSC preconditioners performed worst for all discretizations on both triangular and quadrilateral grids.

The Vanka preconditioner was slightly faster than the LSC preconditioner (except for $\mathcal{RT}_0$ discretizations on triangular grids), in particular for high order discretizations. For $\mathcal{RT}_2$ discretizations on quadrilateral grids the LSC preconditioner did only converge on the coarsest grid (level 0) and for $\mathcal{BDM}_2$ discretizations the LSC preconditioner did not

converge on the finest triangular grid (level 3) within the prescribed number of iteration steps.

PETSc Euclid preconditioner showed inferior efficiency to the direct solver UMFPACK and PETSc LSC preconditioner but superior efficiency to Vanka and LSC preconditioners.

All solvers needed in general less computing time for lower order discretizations $\mathcal{RT}_0$, $\mathcal{BDM}_1$ (linear finite element spaces) than for higher order discretizations $\mathcal{RT}_1$, $\mathcal{BDM}_2$, $\mathcal{RT}_2$ (quadratic and cubic finite element spaces), due to larger number of degrees of freedom for higher order discretizations, in particular on fine grids. For the solution of the systems on the coarsest grids UMFPACK, PETSc LSC and Euclid approaches had similar computing times for all discretizations, whereas Vanka and LSC preconditioners had worse results for higher than for lower order discretiaztions.

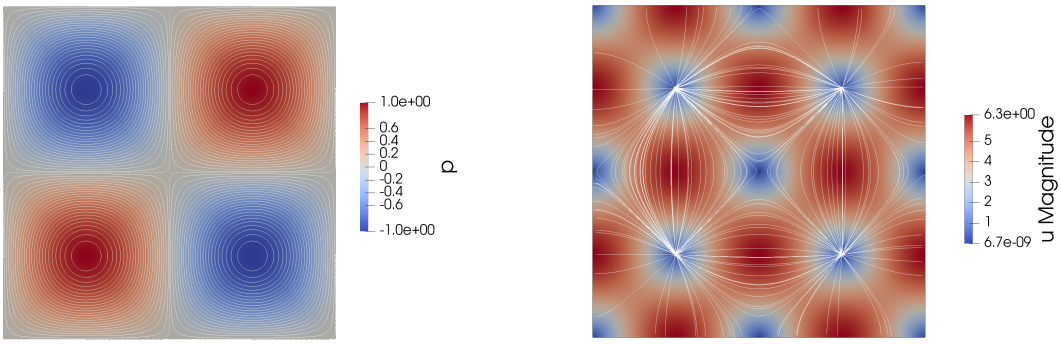

Figure 5.3: Analytic example. Isolines (left) and streamlines (right).

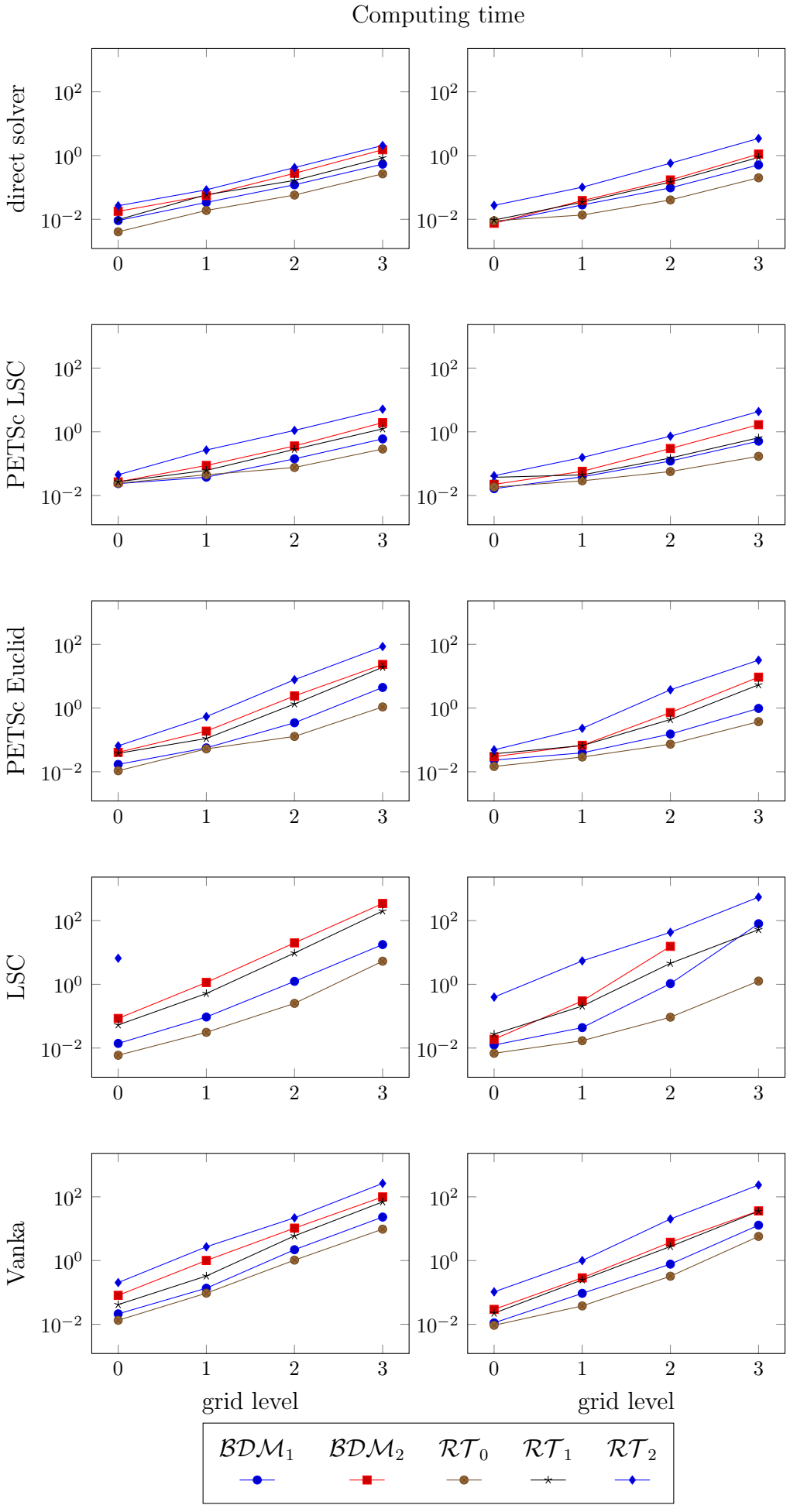

Figure 5.4: Analytic example. Computing times of solvers on triangular (right column) and quadrilateral (left column) grids.

5.2 Five-Spot Problem

Consider a square domain as shown in fig. 5.5 which has prescribed velocity at the lower left-hand corner and the upper right-hand corner and zero normal flow prescribed along the boundaries. Assume the divergence of the velocity field, f, consists of Dirac delta functions acting at the lower left-hand and upper right-hand corner, with strength $\frac{1}{4}$ and $-\frac{1}{4}$, respectively. The lower left-hand corner represents the source, or injection well, while the upper right-hand corner represents the sink, or production well.

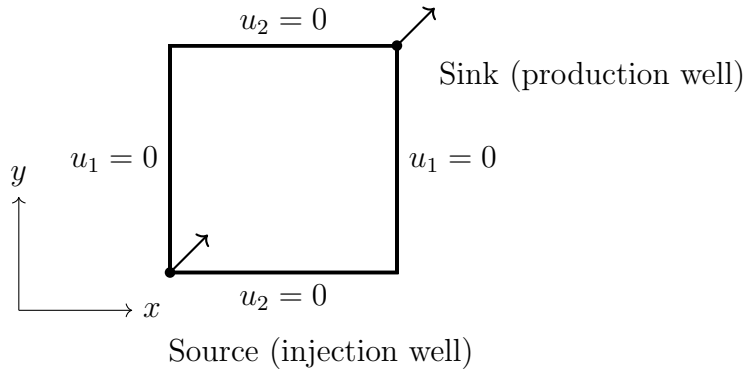

Figure 5.5: Schematic diagram of the quarter five-spot problem.

The divergence of the velocity $\mathbf{u}$ is given by

$$\nabla \cdot \mathbf{u} = \frac{1}{4}\,\delta(x,y) - \frac{1}{4}\,\delta(x-1,y-1) \quad \text{in } [0,1]^2.$$

Inserting Darcy's law, $\mathbb{K}\mathbf{u} = -\nabla p$, with $\mathbb{K} = \mathbb{I}$, leads to the Poisson equation for the pressure p,

$$\nabla \cdot (-\nabla p) = -\Delta p = \frac{1}{4}\,\delta(x,y) - \frac{1}{4}\,\delta(x-1,y-1). \tag{5.1}$$

Recall that the function

$$\Phi(x,y) = -\frac{1}{2\pi}\log\sqrt{x^2+y^2}$$

defined for $\sqrt{x^2+y^2} \neq 0$ is called fundamental solution of Laplace's equation and satisfies

$$-\Delta\Phi = \delta(x,y) \quad \text{in } \mathbb{R}^2.$$

Therefore the solution of eq. (5.1) in $\mathbb{R}^2$ is given by

$$p(x,y) = -\frac{1}{8\pi}\log\sqrt{x^2+y^2} + \frac{1}{8\pi}\log\sqrt{(x-1)^2+(y-1)^2}$$

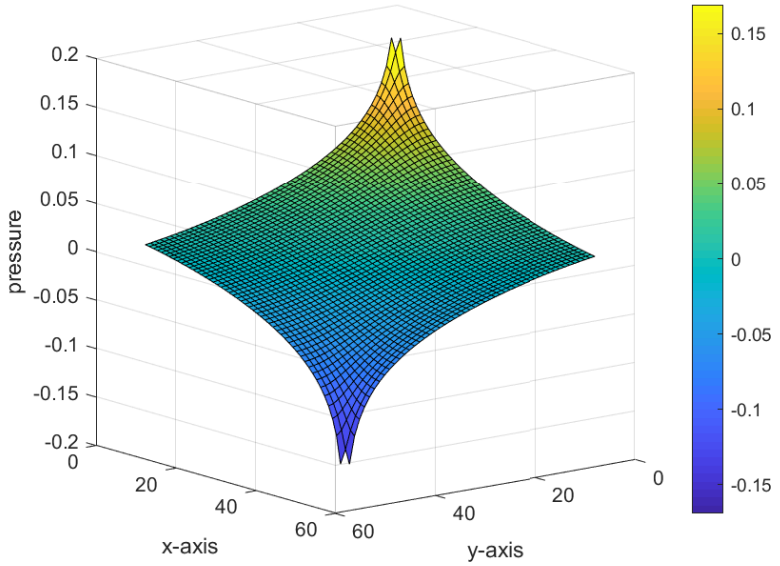

Figure 5.6: Five-spot problem. Elevation plot of the exact pressure field.

and velocity is obtained by Darcy's law,

$$u_1(x,y) = -\frac{\partial}{\partial x}p(x,y) = \frac{1}{8\pi}\left(\frac{x}{x^2+y^2} - \frac{x-1}{(x-1)^2+(y-1)^2}\right)$$

$$u_2(x,y) = -\frac{\partial}{\partial y}p(x,y) = \frac{1}{8\pi}\left(\frac{y}{x^2+y^2} - \frac{y-1}{(x-1)^2+(y-1)^2}\right).$$

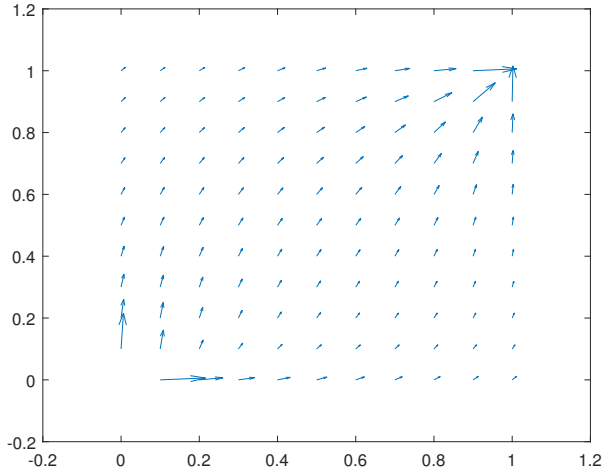

Figure 5.7: Five-spot problem. The velocity field **u** for the quarter five-spot problem.

The Poisson equation (5.1) on the unit square with homogeneous Neumann boundary conditions models two unit points sources of opposite strength and located on a

diagonal of the square. It is a standard porous media problem known as the quarter five-spot problem, which is a popular test case scenario in oil reservoir simulation. The analytical solution of the quarter five-spot problem is derived in [14], where a method for analytically solving boundary value problems stated for the planar Poisson equation in a rectangular domain has been developed.

For the simulation of the problem an equivalent distribution of normal velocity, u_N, was calculated and the simulation was performed with u_N, setting $f = 0$.

Consider a decomposition of the domain $\Omega = [0,1]^2$ into uniform quadrilaterals or triangles. The mesh parameter $h > 0$ is taken to be the edge length for quadrilaterals, and the short-edge length for triangles.

In the case of piecewise linear finite element functions assume a linear distribution of the normal velocity u_N along the external edges of the corner mesh cells. Considering the corner mesh cell at the injection well $(x, y) = (0, 0)$,

$$\begin{aligned} u_N = \mathbf{u} \cdot \mathbf{n}_1 = u_1(0, y) = ay + b \qquad & \text{on } \{0\} \times [0, h] \\ u_N = \mathbf{u} \cdot \mathbf{n}_2 = u_2(x, 0) = ax + b \qquad & \text{on } [0, h] \times \{0\} \end{aligned}$$

where $\mathbf{n}_1 = \begin{pmatrix} -1 \\ 0 \end{pmatrix}$, $\mathbf{n}_2 = \begin{pmatrix} 0 \\ -1 \end{pmatrix}$ and $a, b \in \mathbb{R}$. Assume u_N to be zero at the nodes adjacent to the corner nodes, i.e.

$$u_N(h) = ah + b = 0$$

considering the corner node $(x, y) = (0, 0)$. The divergence theorem gives

$$\int_{\partial[0,1]^2} u_N(x, y)\, ds = 2 \int_0^h ax + b\, dx = ah^2 + 2bh = \frac{1}{4} = \int_{[0,1]^2} \nabla \cdot \mathbf{u}\, dxdy$$

since the divergence of the velocity is the Dirac delta funtion acting at the injection well with strength $\frac{1}{4}$. One can determine a, b by solving

$$\begin{pmatrix} h^2 & 2h \\ h & 1 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \end{pmatrix} = \begin{pmatrix} \frac{1}{4} \\ 0 \end{pmatrix}.$$

It follows

$$-\frac{1}{h^2} \begin{pmatrix} 1 & -2h \\ -h & h^2 \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{4} \\ 0 \end{pmatrix} = \begin{pmatrix} a \\ b \end{pmatrix}$$

and $u_N(x) = -\frac{1}{4h^2}x + \frac{1}{4h}$. The distribution of u_N at the production well is derived similarly. Thus the linear distribution of u_N is uniquely defined on the edges (see fig. 5.9 left).

In the case of piecewise quadratic finite element functions assume a parabolic distribution of the normal velocity u_N along the external edges of the corner mesh cells. Considering again the corner mesh cell at the injection well $(x, y) = (0, 0)$, it should hold

$$\begin{aligned} u_N = \mathbf{u} \cdot \mathbf{n}_1 = u_1(0, y) = ay^2 + by + c \qquad & \text{on } \{0\} \times [0, h] \\ u_N = \mathbf{u} \cdot \mathbf{n}_2 = u_2(x, 0) = ax^2 + bx + c \qquad & \text{on } [0, h] \times \{0\}. \end{aligned}$$

Assume u_N is zero and has zero derivative at the mesh cell vertex nodes away from the corners, i.e. $u_N(h) = ah^2 + bh + c = 0$ and $u_N^{'}(h) = 2ah + b = 0$, considering the corner node $(x, y) = (0, 0)$. Again the divergence theorem gives

$$\int_{\partial[0,1]^2} u_N(x,y)\,ds = 2\int_0^h ax^2+bx+c\,dx = \frac{2}{3}ah^3+bh^2+2ch = \frac{1}{4} = \int_{[0,1]^2} \frac{1}{4}\delta(x,y)\,dxdy.$$

To determine a, b, c solve

$$\begin{pmatrix} \frac{2}{3}h^3 & h^2 & 2h \\ h^2 & h & 1 \\ 2h & 1 & 0 \end{pmatrix} \cdot \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} \frac{1}{4} \\ 0 \\ 0 \end{pmatrix}$$

$\Rightarrow$

$$-\frac{3}{2h^3}\begin{pmatrix} -1 & * & * \\ 2h & * & * \\ -h^2 & * & * \end{pmatrix} \cdot \begin{pmatrix} \frac{1}{4} \\ 0 \\ 0 \end{pmatrix} = \begin{pmatrix} a \\ b \\ c \end{pmatrix}$$

thus $u_N(x) = \frac{3}{8h^3}x^2 - \frac{3}{4h^2}x + \frac{3}{8h}$. The quadratic distribution of u_N is uniquely defined on the edges (see fig. 5.9 right).

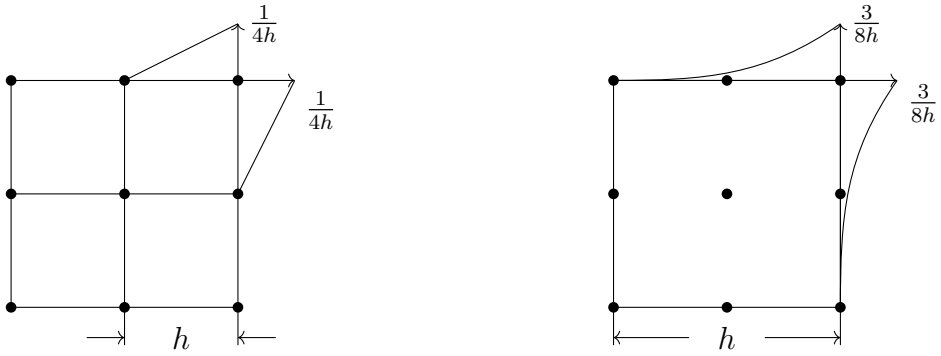

Figure 5.9: Five-spot problem. Distribution of u_N along the corner mesh cell at the production well. The distribution of u_N at the injection well is the same with opposite direction.

As in the previous example the numerical solution was carried out with the standard Galerkin method. The same pairs of finite element spaces on triangular and quadrilateral grids as in the previous example were used.

The results are illustrated in fig. 5.11. The direct solver UMFPACK and PETSc LSC preconditioner were again the fastes methods for all discretizations.

The LSC and Vanka preconditioners were the slowest methods. Vanka preconditioner was faster than LSC preconditioner, except for $\mathcal{RT}_0$ discretizations. On triangular grids the LSC preconditioner did not converge for $\mathcal{RT}_2$ discretization. For the Vanka preconditioner the time spend on the coarsest quadrilateral grid was negligible for $\mathcal{RT}_0$ discretization.

On coarse grids (level 0, 1) PETSc LSC and Euclid preconditioners behaved similarly regarding the computing time. However the computing times of the Euclid preconditioner did increase faster than the computing times of the PETSc LSC preconditioner if the grid was refined. Euclid preconditioner was still more efficient than the Vanka and LSC preconditioners.

In fig. 5.12, it can be seen that the Vanka and LSC preconditioners needed more FGMRES iterations than PETSc LSC and Euclid preconditioners. For PETSc LSC the number of necessary FGMRES iterations was very small and did not increase if the grid was refined. In contrast, these numbers increased considerably for the Euclid preconditioner. Still both methods had similar computing times on coarse grids.

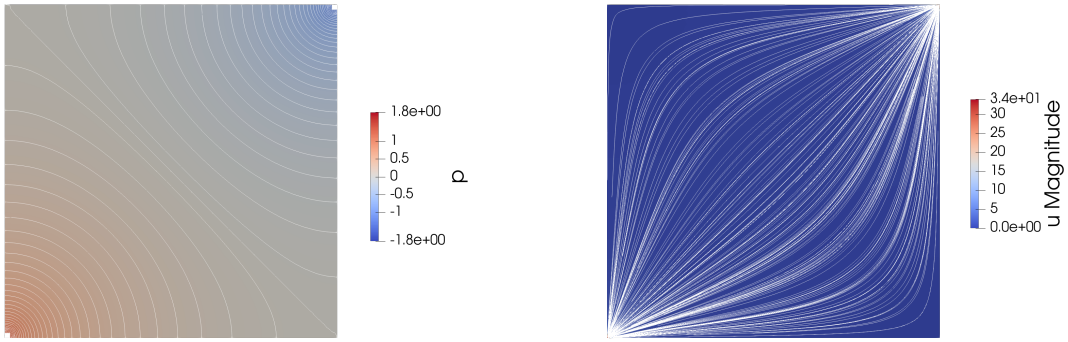

Figure 5.10: Five-spot problem. Isolines (left) and streamlines (right).

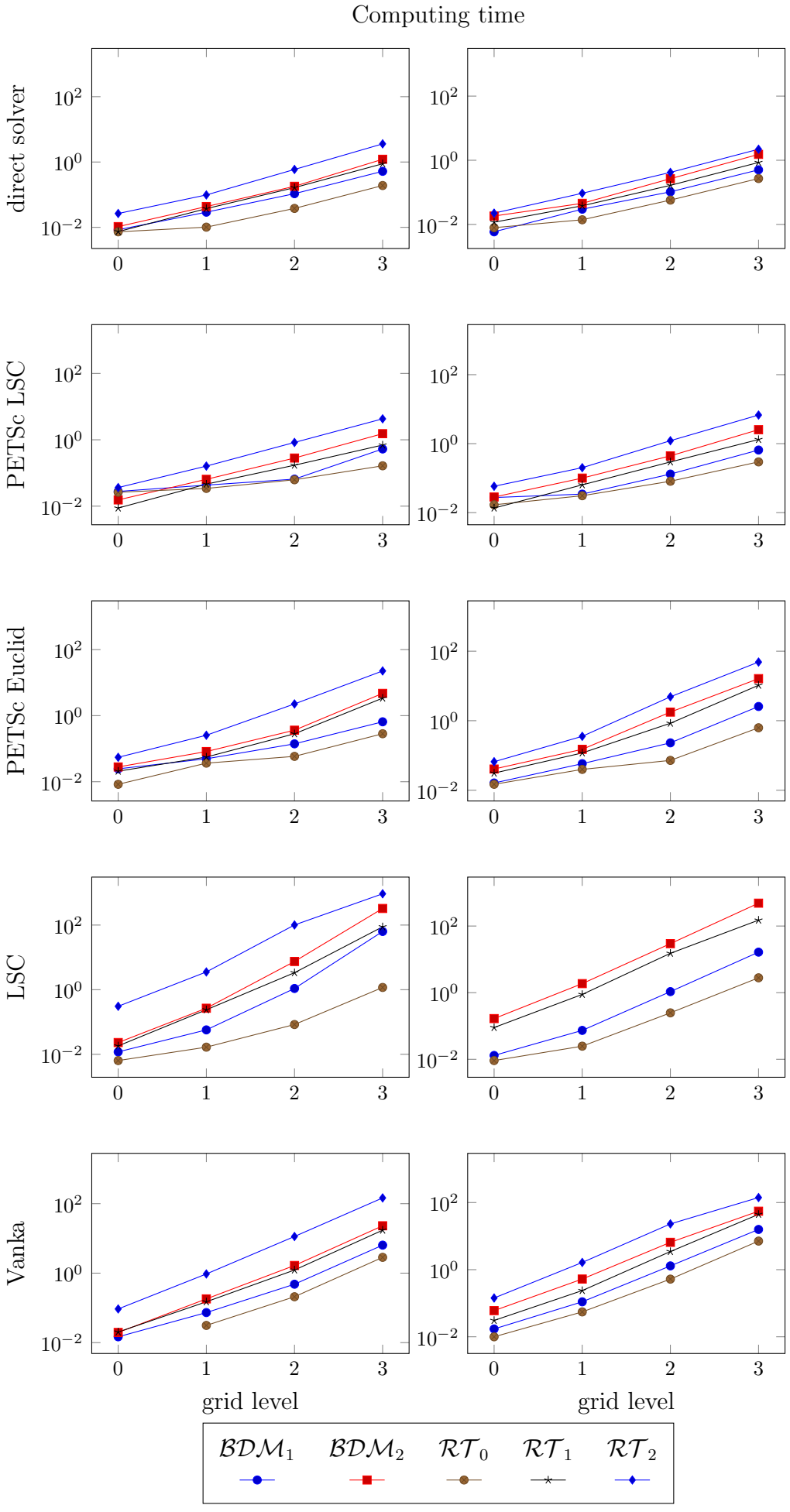

Figure 5.11: Five-spot problem. Computing times on triangular (right column) and quadrilateral (left column) grids.

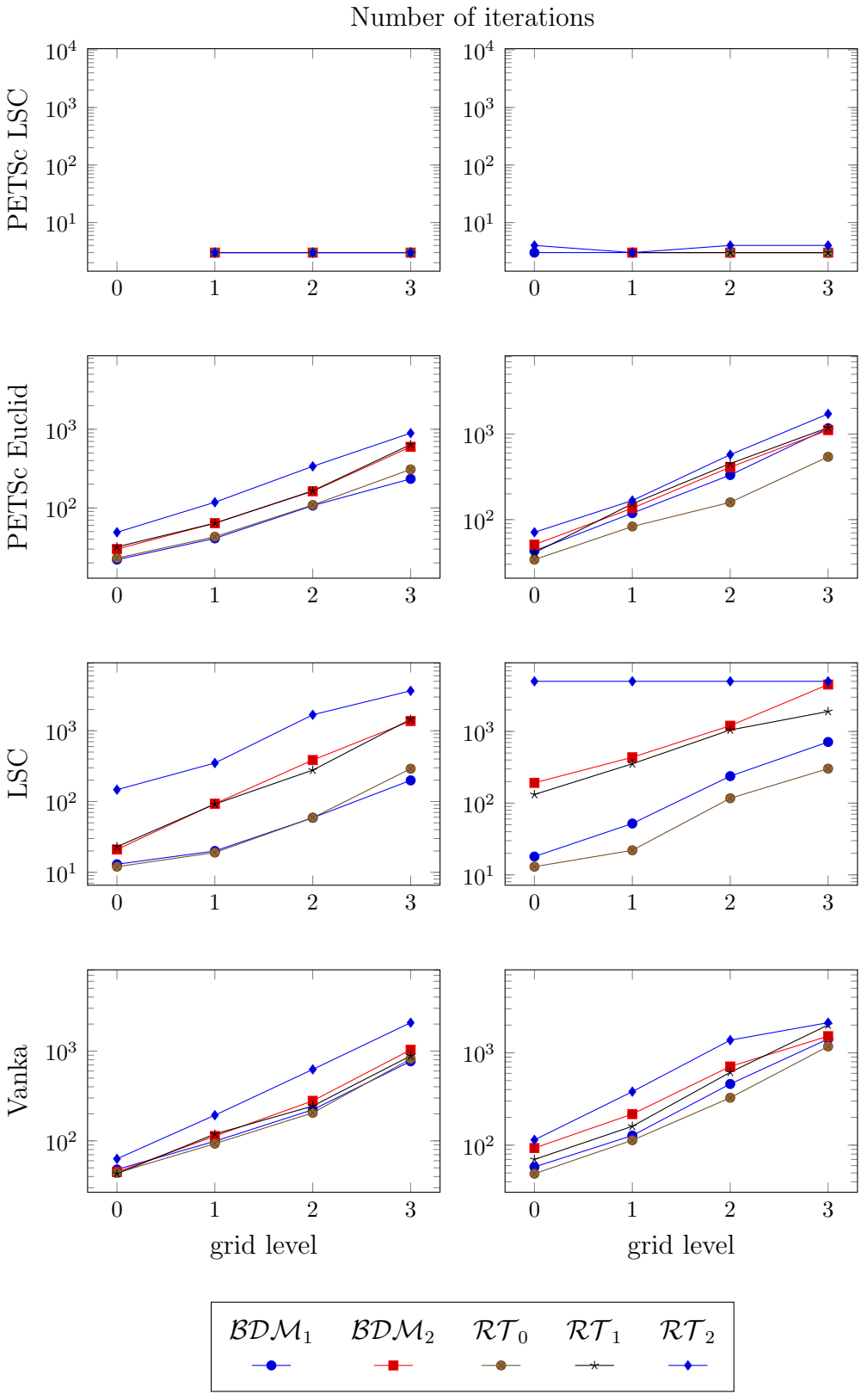

Figure 5.12: Five-spot problem. Number of FGMRES iterations on triangular (right column) and quadrilateral (left column) grids.

5.2.1 Nine-Spot Problem

Consider an extension of the quarter five-spot problem to the three dimensional case. Assume that the cube $\Omega = [0,1]^3$ has prescribed velocity at the lower left-hand corner and the upper right-hand corner and zero normal flow prescribed along the faces of the cube. Assume the divergence of the velocity field, f, consists of Dirac delta functions acting at the lower left-hand and upper right-hand corner, with strength $\frac{1}{8}$ and $-\frac{1}{8}$, respectively.

The fundamental solution of Laplace's equation in 3d is given by

$$\Phi(x,y,z) = \frac{1}{4\pi\sqrt{x^2+y^2+z^2}}.$$

Therefore the pressure satisfying

$$-\Delta p = \frac{1}{8}\,\delta(x,y,z) - \frac{1}{8}\,\delta(x-1,y-1,z-1) \tag{5.2}$$

in $\mathbb{R}^3$ is given by

$$p(x,y,z) = \frac{1}{32\pi}\left(\frac{1}{\sqrt{x^2+y^2+z^2}} - \frac{1}{\sqrt{(x-1)^2+(y-1)^2+(z-1)^2}}\right)$$

and the velocity is obtained with Darcy's law

$$u_1(x,y) = -\frac{\partial}{\partial x}p(x,y) = \frac{1}{32\pi}\left(\frac{x}{\left(x^2+y^2+z^2\right)^{\frac{3}{2}}} - \frac{x-1}{\left((x-1)^2+(y-1)^2+(z-1)^2\right)^{\frac{3}{2}}}\right)$$

$$u_2(x,y) = -\frac{\partial}{\partial y}p(x,y) = \frac{1}{32\pi}\left(\frac{y}{\left(x^2+y^2+z^2\right)^{\frac{3}{2}}} - \frac{y-1}{\left((x-1)^2+(y-1)^2+(z-1)^2\right)^{\frac{3}{2}}}\right).$$

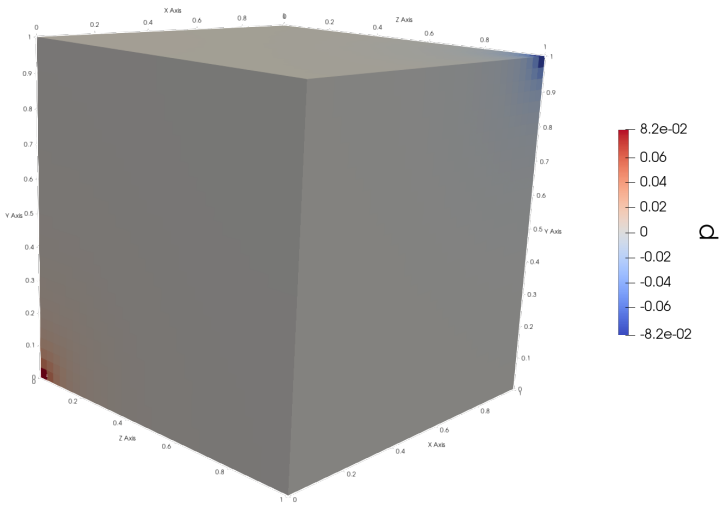

Figure 5.13: Nine-spot problem. $\mathcal{RT}_0$ - pressure solution on finest grid.

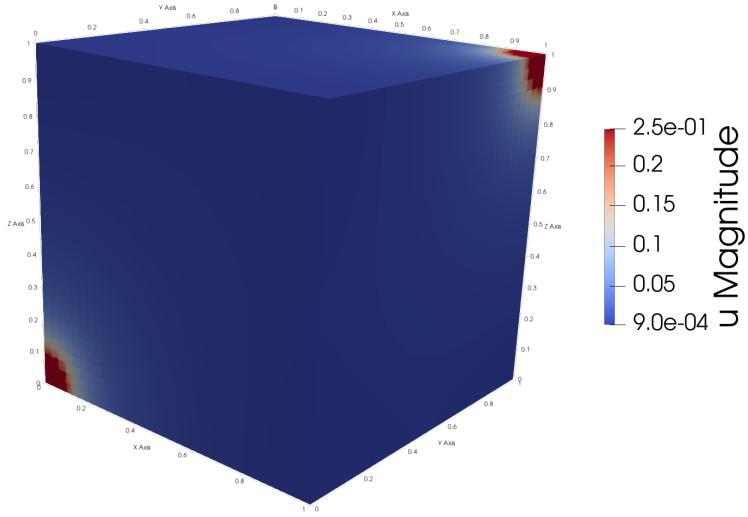

Figure 5.14: Nine-spot problem. $\mathcal{RT}_0$ - velocity solution on finest grid.

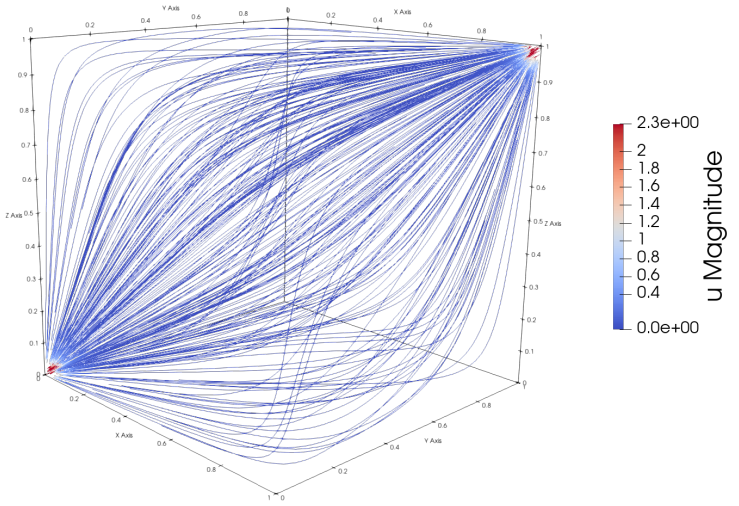

Figure 5.15: Nine-spot problem. Streamfunction.

Consider a decomposition of the domain $\Omega = [0,1]^3$ into uniform hexahedra or tetrahedra. The mesh parameter $h > 0$ is taken to be the edge length of the hexahedra or tetrahedra. Assume a constant distribution of the normal velocity u_N along the external faces of the corner mesh cells $[0,h]^3$ and $[1-h,1]^3$ and perform the simulation of the problem with u_N, setting $f = 0$.

The solution is carried out with the Galerkin finite element method. The velocity finite element spaces $\mathcal{RT}_k$, $k = 0, 1, 2$ and $\mathcal{BDM}_k$, $k = 1, 2$ with corresponding pressure finite element spaces $\mathcal{P}_k^d$ were used on uniform tetrahedal and hexahedral grids.

Table 5.2 and fig. 5.16 illustrate the number of degrees of freedom on hexahedral grids obtained by uniform refinement of the domain.

Table 5.2: Number of cells and number of degrees of freedom on hexahedral grids.

grid level	number of cells	number of degrees of freedom $\mathcal{BDM}_1$	$\mathcal{BDM}_2$	$\mathcal{RT}_0$	$\mathcal{RT}_1$	$\mathcal{RT}_2$
0	64	784	1888	304	2240	7344
1	512	5696	13952	2240	17152	57024
2	4096	43264	107008	17152	134144	449280
3	32768	336896	837632	134144	1060864	3566592

Number of degrees of freedom (# dof) on the used hexahedral grids

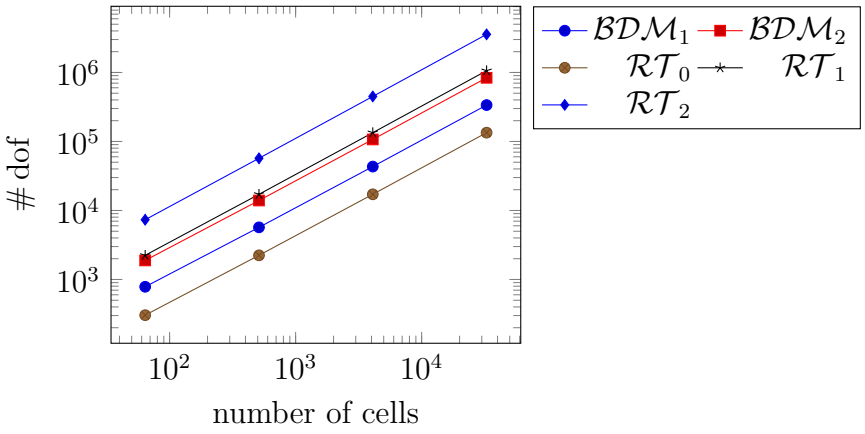

Figure 5.16: Plot of number of degrees of freedom on hexahedral grids

The results are presented in fig. 5.17. In 3d the Euclid preconditioner showed the best performance. The computing times of Vanka and Euclid preconditioners were similar; on fine grids Vanka was slightly slower than Euclid.

The LSC preconditioner had worst results: systems on the finest grids could not be solved and for high order discretizations ($\mathcal{RT}_2$ and $\mathcal{BDM}_2$) the LSC preconditioner did only converge on coarse grids (level 2, 3). However for low order discretizations LSC was efficient on the coarsest grids (level 2) in particular for $\mathcal{RT}_0$ on triangular grids.

The direct solver UMFPACK showed rapid increasement of computing times if grid was refined for all discretizations compared to Vanka and Euclid preconditioners and did

not converge on the finest grids (level 5) for $\mathcal{RT}_2$ discretizations. In addition, solving the systems with the direct solver needed high memory requirements.

On coarse grids (level 2, 3) the results of the PETSc LSC preconditioner were in general similar to the results of Euclid and Vanka preconditioners. The computing times of the PETSc LSC preconditioner did increase faster than the computing times of Euclid and Vanka preconditioners (except for $\mathcal{RT}_0$ on triangular grids) if the grid was refined in particular for $\mathcal{BDM}_1$. On the finest quadrilateral grid (level 5) the PETSc LSC preconditioner did not converge for $\mathcal{RT}_2$ and $\mathcal{BDM}_2$ discretizations, whereas on the finest triangular grid the computing times of the PETSc LSC preconditioner did not increase.

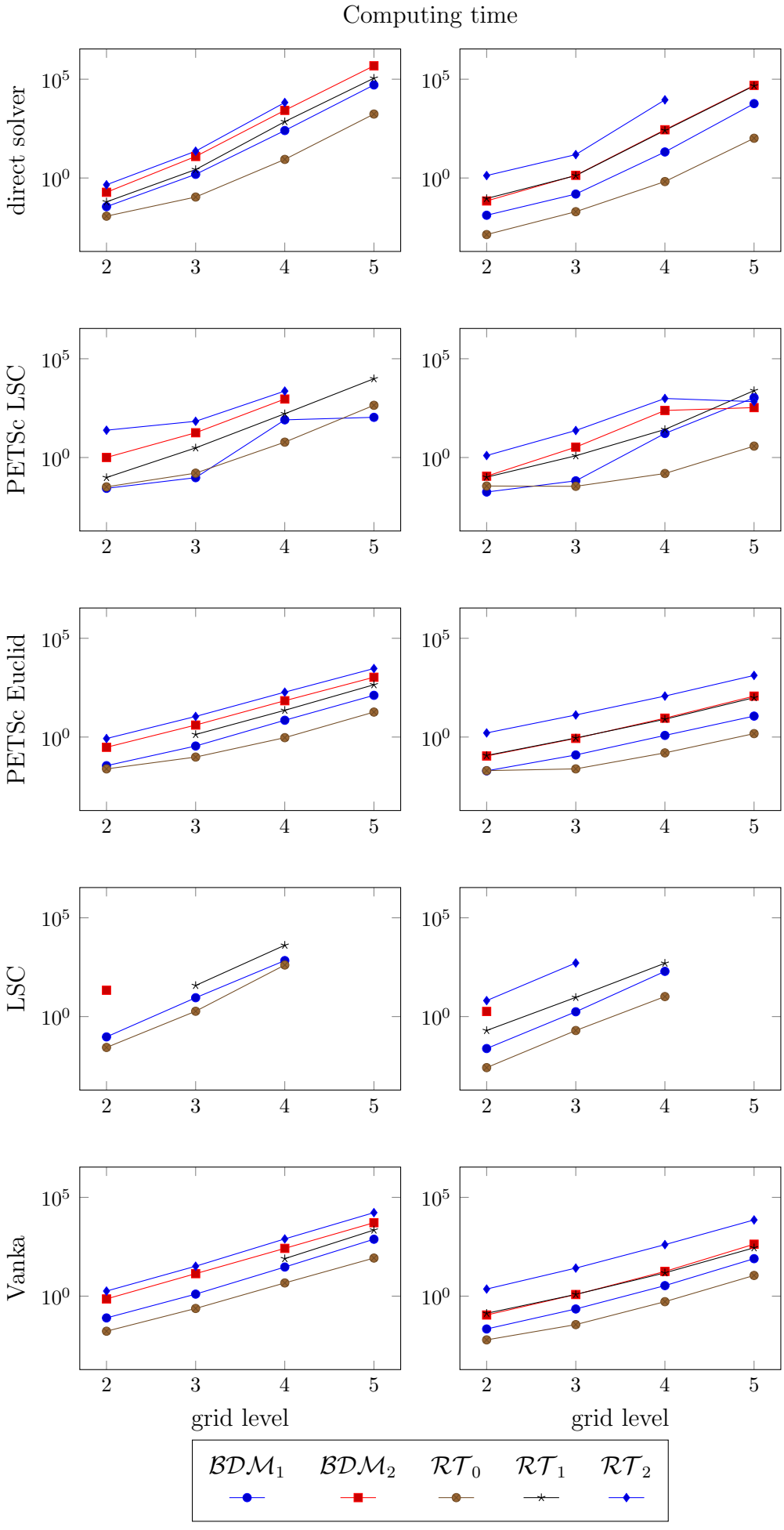

Figure 5.17: Nine-spot problem. Computing times on triangular (right column) and quadrilateral (left column) grids.

5.3 The Checkerboard Domain

This simulation considers cases in which there are abrupt changes in the permeability parameter. We consider the quarter five-spot problem described earlier, now zoned as shown in fig. 5.18 with a piecewise constant permeability tensor.

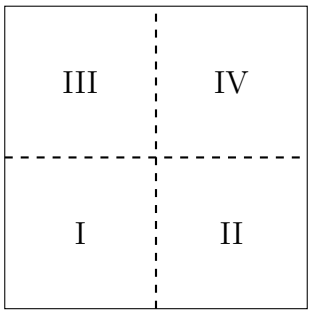

Figure 5.18: Checkerboard domain with sharp change in permeability value.

Consider the case where the permeability is set to $\mathbb{K} = \mathbb{I}$ in zones I and IV and $\mathbb{K} = 100 \cdot \mathbb{I}$ in zones II and III. The numerical solution was carried out with the standard Galerkin method for the velocity finite element spaces $\mathcal{RT}_k$, $k = 0, 1, 2$ and $\mathcal{BDM}_k$, $k = 1, 2$ on uniform triangular and quadrilateral grids with corresponding pressure finite element spaces $\mathcal{P}_k^d$. The results are illustrated in fig. 5.20. The direct solver UMFPACK was the most efficient method. PETSc LSC preconditioner showed inferior efficiency to UMFPACK in particular for higher order discretizations on fine grids, but still superior efficiency to the other solvers. LSC precondtioner did only converge for $\mathcal{RT}_0$ and $\mathcal{BDM}_1$ discretizations on coarse grids (level 0,1), hence showed worst performance. Vanka was only efficient for low order discretizations on coarse grids, the computing times did increase fast if the grid was refined. The systems on the finest grids (level 3) could not be solved with Vanka preconditioner. Euclid behaved similar to PETSc LSC, but was slower or did not converge on the finest grids (level 3). On the finest triangular grid Euclid did not converge for for $\mathcal{RT}_2$. On the finest quadrilateral grid Euclid did only converge for $\mathcal{RT}_0$ disretization.

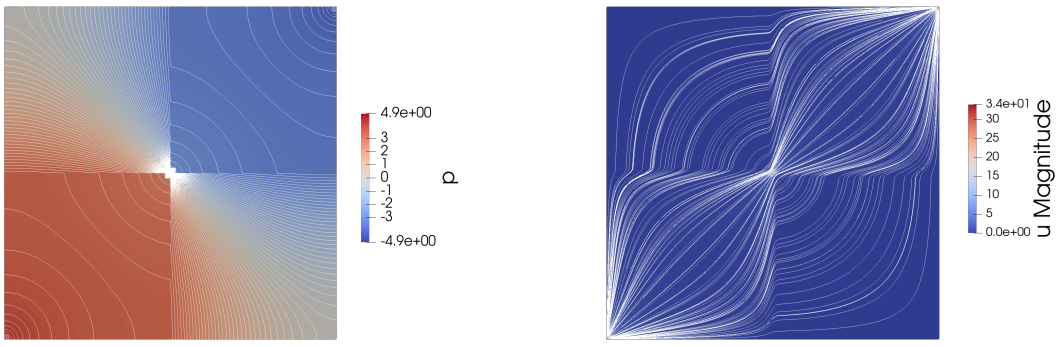

Figure 5.19: Checkerboard problem: Isolines (left) and streamlines (right).

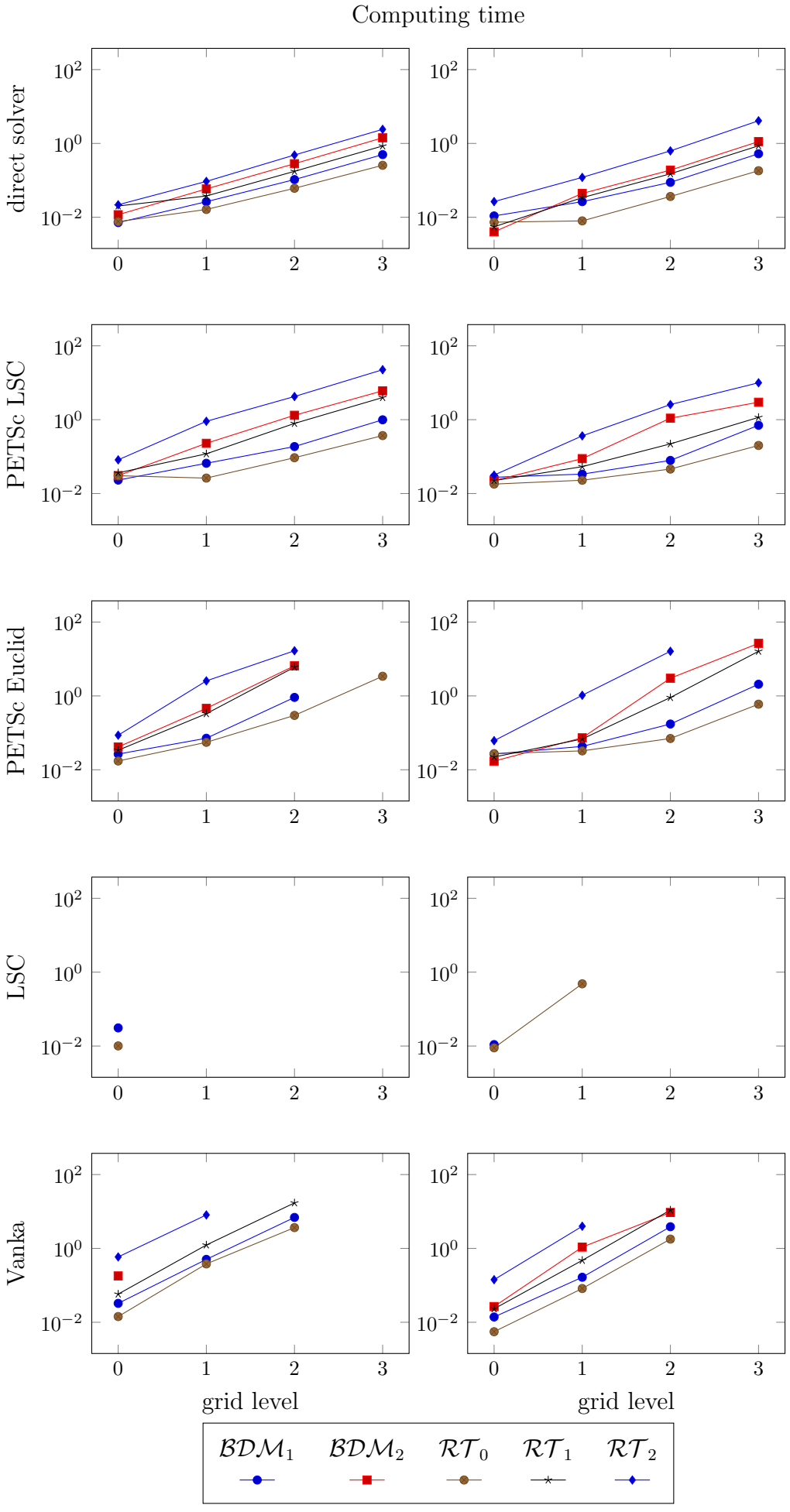

Figure 5.20: Checkerboard domain. Computing times of solvers for discretizations on triangular (right column) and quadrilateral (left column) grids.

5.3.1 Checkerboard Domain in 3D

An extension of the previous problem to the three dimensional case is considered, i.e. the nine-spot problem for a cube zoned as shown in fig. 5.21 where the permeability is set to $\mathbb{K} = \mathbb{I}$ in zones I and VIII and $\mathbb{K} = 100 \cdot \mathbb{I}$ in the remaining zones.

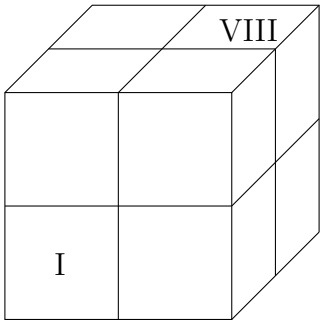

Figure 5.21: Checkerboard domain for a cube.

The numerical solution was carried out with the standard Galerkin method for the velocity finite element spaces $\mathcal{RT}_k$, $k = 0, 1, 2$ and $\mathcal{BDM}_k$, $k = 1, 2$ with corresponding pressure finite element spaces $\mathcal{P}_k^d$ on uniform tetrahedral and hexahedral grids.

The results are illustrated in fig. 5.22: The behavior of the solvers was the same as in the previous example in 3d, the nine-spot problem.

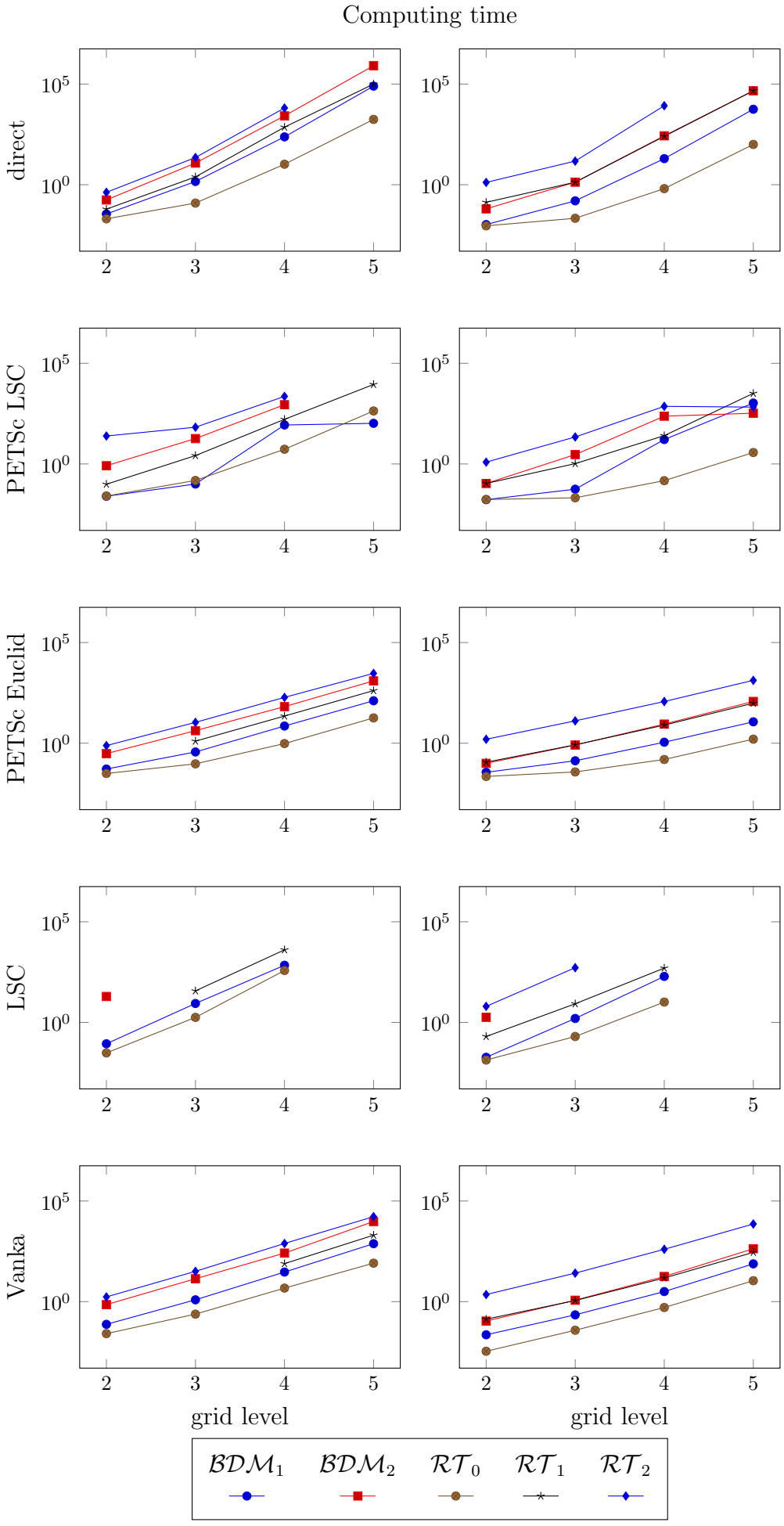

Figure 5.22: Checkerboard domain in 3d. Computing times of solvers for discretizations on triangular (right column) and quadrilateral (left column) grids.

5.4 Summary of Results

In the examples presented in this chapter, we could generally observe

- the direct solver UMFPACK was the fastes method for all examples in 2d. In contrast, the efficiency of the direct solver decrased considerably on finer grids in 3d.
- Considering only FGMRES with LSC and Euclid preconditioners provided by PETSc, one could see that for the considered examples in 2d the LSC preconditioner proved to be faster than Euclid preconditioner regarding the computing times in particular on fine grids. For the studied examples in 3d the Euclid preconditioner had the best results compared to the other solvers.
- FGMRES with Vanka and LSC preconditioners provided by ParMooN were the least efficient methods for the considered examples in 2d. In general, the Vanka preconditioner demonstrated better results than LSC preconditioner, in particular for higher order discretizations. For the examples in 3d Vanka was most efficient together with Euclid.
- In 2d the permeability tensor with abrupt changes (checkerboard domain) did impact the performance of the solvers. All solvers, exept for the direct solver, performed worse in particular on fine grids. That was not the case in 3d.

6 Conclusion and Outlook

This thesis studied different solvers for linear systems in saddle point form that arise in finite element discretizations of the Darcy equations. Firstly the mathematical foundations for the Darcy equations and the corresponding saddle point systems were presented. Inf-sup stable pairs of finite element spaces employed in the numerical studies were introduced. Afterwards the basic properties of the saddle point matrices were briefly considered and techniques of preconditioning, were discussed. In numerical studies for problems in 2d and 3d the performance of LSC and Vanka preconditioners provided by ParMooN, LSC and Euclid preconditioners provided by PETSc and the direct solver UMFPACK were compared. The conclusions of these studies are as follows.

- For problems in 2d the direct solver UMFPACK was the best choice. FGMRES with the LSC preconditioner provided by PETSc was competitive to the direct solver in case of a constant permeability tensor $\mathbb{K} = \mathbb{I}$. The computing times of PETSc LSC and Euclid preconditioners were similar on coarse grids however computing times of Euclid increased faster if the grid was refined. Vanka and LSC preconditioners provided by ParMooN were slow compared to the other solvers, in particular LSC preconditioner for high order discretizations.

- The performance of the studied solvers for problems in 2d depended on the permeability tensor. Examples with $\mathbb{K} = \mathbb{I}$ and a piecewise constant permeability tensor (checkerboard domain) were considered. The mass matrix A in both cases has different properties, which have a different impact on the efficiency of the studied solvers. For a piecewise constant permeability parameter the studied solvers performed worse than for $\mathbb{K} = \mathbb{I}$.

- The simulation of problems in 3d could be performed fastest with FGMRES and the Euclid or Vanka preconditioners. UMFPACK was ineffcient with respect to computation time and memory requirements. The results of the PETSc LSC preconditioner were in general comparable with Euclid and Vanka preconditioners, whereas LSC preconditioner provided by ParMooN could only solve systems for low order discretizations on coarse grids.

Further investigations could consider different permeability parameters such as nondiagonal tensors and check their influence on the efficiency of the solvers. Iterative methods for the solution of the arising subproblems in the Vanka and LSC preconditioners could be investigated. In the further investigations could consider other preconditioning techniques such as multigrid preconditioning and other block preconditioners based on the Schur complement approximation in order to compare their performance as well as time-dependend problems.

List of Figures

List of Tables

Bibliography

[1] N. Ahmed, C. Bartsch, V. John, U. Wilbrandt: *An assessment of some solvers for saddle point problems emerging from the incompressible Navier–Stokes equations*, Computer Methods in Applied Mechanics and Engineering, 331:492 – 513, 2018.

[2] M. Benzi, G.H. Golub, J. Liesen: *Numerical Solution of Saddle Point Problems*, Acta Numerica, Cambridge University Press, Cambridge, 1-137, 2005

[3] D. Boffi, F. Brezzi, M. Fortin: *Mixed Finite Element Methods and Applications*, Springer Series in Computational Mathematics, Vol. 44, Springer 2013

[4] T. A. Davis: *Algorithm 832: UMFPACK V4.3—an unsymmetric-pattern multifrontal method*, ACM Trans. Math. Softw., 30(2):196–199, June 2004. ISSN 0098-3500.

[5] R. G. Durán: *Mixed Finite Element Methods*, In Mixed finite elements, compatibility conditions, and applications. Lectures given at the C.I.M.E. summer school, Cetraro, Italy, June 26–July 1, 2006, pages 1–44. Berlin: Springer; Florenz: Fondazione CIME Roberto Conti, 2008.

[6] H. C. Elman, D. J. Silvester, A. J. Wathen: *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*, Numerical Mathematics and Scientific Computation, second ed., Oxford University Press, Oxford, 2014, p. xiv+479

[7] L. C. Evans : *Partial Differential Equations,* American Mathematical Society, 1998

[8] M. Fortin: *An analysis of the convergence of mixed finite element methods*, R.A.I.R.O. Anal. Numer. 11, 341-354, 1977

[9] V. Girault, P.-A. Raviart: *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, Berlin, 1986

[10] T.J.R. Hughes, A. Masud: *A stabilized mixed finite element method for Darcy flow*, Comput. Methods in Appl. Mech. Engrg., 191/39-40, 4341-4370, 2002.

[11] T.J.R Hughes, A. Masud and J. Wan: *A Stabilized mixed discontinuous Galerkin method for Darcy Flow.* Computer Methods in Applied Mechanics and Engineering, 195:3347–3381, 2006

[12] V. John: *Finite Element Methods for Incompressible Flow Problems*, Springer Series in Computational Mathematics, vol. 51, Springer-Verlag, Berlin, 2016.

[13] V. John: *Numerical Mathematics II: Iterative Methods for Solving Linear System Equations*, Lecture notes, Berlin 2016/17

[14] M. Mamode, *Fundamental solution of the Laplacian on flat tori and boundary value problems for the planar Poisson equation in rectangles*, Boundary Value Problems (2014):221. https://doi.org/10.1186/s13661-014-0221-4, SpringerOpen 2014

[15] T. Rusten, R. Winther: *A preconditioned iterative method for saddlepoint problems*, SIAM J. Matrix Anal. Appl. 13, 887–904, 1992

[16] N. Schönknecht: *On solvers for saddle point problems arising in finite element discretizations of incompressible flow problems*, Masterarbeit, Berlin 2015

[17] R. Falgout, A. Barker, T. Kolev, R. Li, D. Osei-Kuffuor, J. Schroder, P. Vassilevski, L. Wang, U. Meier Yang: *Hypre documentation and Web page https://hypre.readthedocs.io/en/latest/index.html, https://hypre.readthedocs.io/en/latest/solvers-euclid.html, https://computing.llnl.gov/projects/hypre-scalable-linear-solvers-multigrid-methods*

[18] U. Wilbrandt, C. Bartsch, N. Ahmed, N. Alia, F. Anker, L. Blank, A. Caiazzo, S. Ganesan, S. Giere, G. Matthies, R. Meesala, A. Shamim, J. Venkatesan, V. John: *ParMooN—A modernized program package based on mapped finite elements*, Comput. Math. Appl. 74 (1) (2017) 74–88.

[19] S. Balay, S. Abhyankar, M. F. Adams, J. Brown, P. Brune, K. Buschelman, L. Dalcin, V. Eijkhout, W. D. Gropp, D. Kaushik, M. G. Knepley, L. Curfman McInnes, K. Rupp, B. F. Smith, S. Zampini, H. Zhang: *PETSc Web page http://www.mcs.anl.gov/petsc*, 2016.

[20] S. Balay, S. Abhyankar, M. F. Adams, J. Brown, P. Brune, K. Buschelman, L. Dalcin, V. Eijkhout, W. D. Gropp, D. Kaushik, M. G. Knepley, L. Curfman McInnes, K. Rupp, B. F. Smith, S. Zampini, H. Zhang: *PETSc users manual. Technical Report ANL-95/11 - Revision 3.7, Argonne National Laboratory*, 2016.