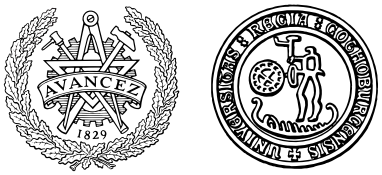
THESIS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

Computational characterization of mixing in flows

ERIK D. SVENSSON

CHALMERS | GÖTEBORG UNIVERSITY



Department of Mathematical Sciences
Chalmers University of Technology and Göteborg University
Göteborg, Sweden 2006

This work was supported by:
The Network in Applied Mathematics;
The Swedish Foundation for Strategic Research;
IMEGO, the Institute of Microelectronics in Gothenburg; and
The Department of Mathematical Sciences, Chalmers University of Technology

Computational characterization of mixing in flows
Erik D. Svensson
ISBN 91-7291-771-7

Doktorsavhandlingar vid Chalmers tekniska högskola
Ny serie nr. 2453
ISSN 0346-718x

Department of Mathematical Sciences
Chalmers University of Technology and Göteborg University
SE-412 96 Göteborg
Sweden
Telephone +46 (0)31 772 1000

Printed in Göteborg, Sweden 2006

COMPUTATIONAL CHARACTERIZATION OF MIXING IN FLOWS

ERIK D. SVENSSON

Department of Mathematical Sciences
Chalmers University of Technology and Göteborg University

Abstract

The major theme of this thesis is mathematical aspects of fluid mixing in the case when diffusion is negligible, which is commonly refered to as 'mixing by stirring' or 'mixing by chaotic advection' in the engineering literature. In this case the mixing is driven by a velocity field and is characterized by the flow generated by the velocity field. We propose a general methodology that can be used to characterize mixing in flows.

In this work we assume that the velocity field is modeled by the incompressible Stokes equations but in principle we can choose to use any other fluid model. We derive pointwise a posteriori error estimates for finite element approximations of the Stokes equations and investigate the flow generated by the velocity field by computing a large number of orbits in the flow. We demonstrate that the computed orbits are close to exact orbits by deriving a shadowing error estimate. Principal to this estimate is that we compute the orbits and the velocity field sufficiently accurately.

On the basis of notions from dynamical systems theory we devise a tractable mixing measure that resolves the mixing process both in space and time. We provide an error estimate for computed mixing measures which relies on the error estimate for the computed orbits.

Finally, we discuss a few additional computational issues. (1) We suggest an optimal search algorithm that given a query point can locate the n-simplex in a finite element triangulation that contains the query point. (2) We analyse and discuss finite element multigrid methods for quadratic finite elements and for adaptively refined triangulations.

Key Words: mixing, hyperbolicity, shadowing, finite elements, flow simulation, a priori error estimates, Stokes equations, point location, multigrid, refinements

2000 Mathematics Subject Classification: 37A25, 37C50, 76M10, 65N15, 65N30, 76D07, 37A25, 37C50, 76M10, 68U05, 65N55, 65N50

Appended papers

This thesis is based on the work contained in the following papers, referred to by Roman numbers in the text.

Paper I. Erik D. Svensson, *Computational characterization of mixing in flows*

Paper II. Erik D. Svensson and Stig Larsson, *Pointwise a posteriori error estimates for the Stokes equations in polyhedral domains*

Paper III. Erik D. Svensson, *Computational characterization of flows with some hyperbolicity*

Paper IV. Erik D. Svensson, *Optimal search in finite element triangulations using binary trees*

Paper V. Erik D. Svensson, *Multigrid for quadratic finite elements*

Paper VI. Erik D. Svensson, *Multigrid methods on adaptively refined triangulations: practical considerations*

Preface

This thesis is an outcome of the collaboration I had a few years ago with the Institute of Microelectronics in Gothenburg (IMEGO), where we computationally investigated mixing in micro fluid systems. In connection to this I like to thank Peter Björkholm, Anatol Krozer and Dag Billger at IMEGO for support, inspiration and motivation. At the time I was enrolled in the post graduate program in industrial mathematics organized by the European Consortium for Mathematics in Industry (ECMI), designed to improve the participants interdisciplinary skills, and to promote and nourish the use of mathematical methods in industry. I like to thank the local ECMI administration Professors Leif Arkeryd, Jöran Bergh, Peter Kumlin, Axel Ruhe and Bernt Wennberg for admission, initial support and encouragements.

Since then I have refined, improved and generalized the work on mixing in micro fluid systems.

I am grateful to my advisor Professor Stig Larsson for teaching me, helping with various technical difficulties and for his over all support but also to my assistant advisor Professor Grigori Rozenblioum. I specially thank Professor Mats G Larson for helpful discussions and particularly for suggesting to use shadowing error analysis and irregular triangulations. I thank Professors Zhongwei Shen and Jürgen Roßmann for helping me with some technical difficulties I had in the preparation of Paper II. I am also much obliged to the financial support from: the Network in Applied Mathematics, The Swedish Foundation for Strategic Research, IMEGO -the Institute of Microelectronics in Gothenburg, and the department of mathematical sciences at Chalmers University of Technology and Göteborg University.

Major parts of this work involves computer implementations of various algorithms, in connection to this I like to thank Johan Jansson for introducing me to Binary Space Partitioning algorithms and David Heintz for introducing me to Ubuntu, the Linux distribution.

I like to thank Jovan Pankovski for printing this thesis.

Finally, I am thankful to the persistent support from my family, friends and Maria.

Göteborg, April 2006
Erik Svensson

COMPUTATIONAL CHARACTERIZATION OF MIXING IN FLOWS

ERIK D. SVENSSON

Contents

1. Introduction

Let Ω be a domain in $\mathbf{R}^n$ for $n = 2, 3$ and let $c_0(x) : \Omega \to \mathbf{R}$ be a positive function. We think of mixing as a relaxation process $c(x,t) : \Omega \to \mathbf{R}$ for $t > 0$ such that $c(x,0) = c_0$ and

$$\|c(x,t) - \overline{c}_0\|_{L^\infty(\Omega)} \to 0 \quad \text{as } t \to \infty, \tag{1.1}$$

where

$$\overline{c}_0 = |\Omega|^{-1} \int_\Omega c_0(x)\, dx.$$

In practice $c(x,t)$ could be the concentration of a chemical compound that we wish to distribute uniformly in Ω, for example, in order to mix

with one or several other miscible compounds in Ω. We model this process by the *convection-diffusion equation* which is in dimensionless form

$$\begin{aligned} \partial_t c + u \cdot \nabla c - \mathrm{Pe}^{-1} \Delta c &= 0 && \text{in } [0,T] \times \Omega, \\ \nabla c \cdot \nu &= 0 && \text{in } [0,T] \times \partial\Omega, \\ c(x,0) &= c_0(x) && \text{in } \Omega, \end{aligned} \tag{1.2}$$

where ν is the outward normal to $\partial\Omega$, u is an incompressible velocity field such that $u \cdot \nu = 0$ on $\partial\Omega$ and

$$\mathrm{Pe} = \frac{UL}{D} \tag{1.3}$$

is the *Péclet number* for characteristic velocity and length scales U and L and the diffusion constant D.

Solving (1.2) it is important to recognize the Péclet number and discriminate between three different regimes.

(1) For $\mathrm{Pe} \ll 1$ the problem is diffusion-dominated.
(2) For $\mathrm{Pe} \sim 1$ the problem involves both convection and diffusion.
(3) For $\mathrm{Pe} \gg 1$ the problem is convection-dominated.

In this thesis we only consider the case when $\mathrm{Pe} \gg 1$, in fact, we assume that Pe is so large that the diffusion is negligible and thus in stead of the convection-diffusion equation (1.2) we consider the *transport equation*

$$\begin{aligned} \partial_t c + u \cdot \nabla c &= 0 && \text{in } [0,T] \times \Omega, \\ \nabla c \cdot \nu &= 0 && \text{in } [0,T] \times \partial\Omega, \\ c(x,0) &= c_0(x) && \text{in } \Omega. \end{aligned} \tag{1.4}$$

It is now apparent to ask what properties the velocity field u must have in order to make $c(x,t)$ mixing. As we shall see, this problem could be formulated within the realm of dynamical systems theory where we also will find a more precise answer to the question.

In the engineering literature mixing in the case $\mathrm{Pe} \gg 1$ is commonly refereed to as *mixing by chaotic advection* and for further references we refer to the survey articles [2, 3, 34] or the book [33].

1.1. **Dynamical systems approach.** We do not solve the transport equation (1.4) because it seems to be too involved, for example, if c_0 is discontinuous and u is sufficiently irregular it will be technically difficult to accurately evolve c_0 in time. Instead we consider the flow $[0,t] \times \Omega \ni (t,x) \mapsto g(t,x) \in \Omega$ generated by a sufficiently smooth velocity field u, describing

the motion of fluid particles in Ω as solutions to the system of ordinary differential equations

$$\partial_t g(t,x) = u(g(t,x)), \quad t>0; \quad g(0,x)=x. \tag{1.5}$$

Now (1.4) and (1.5) are equivalent and $c(t,x) = c_0 \circ g(t,x)$, or formally $c(t,x)$ defines the Koopman operator associated to the semigroup of transformations $\{S_t\}_{t>0}$ corresponding to the flow (1.5) [27, p. 210].

Thus instead of trying to solve (1.4) we compute numerically a limited number of orbits $g_k(t,x_j)$ for $j=1,2,\ldots,J$, $x_j \in \Omega$, and where k denotes the step in a time discretization.

We remark that there is a similar correspondence between the convection-diffusion equation (1.2) and (1.5) with an additional stochastic term on the right hand side of (1.5). Hence for large and moderate Péclet numbers (1.2) will be difficult to solve for the same reasons as (1.4) is and we may instead consider a stochastic flow.

Let μ be a measure that is preserved in the flow, that is, for every open set $A \subseteq \Omega$, $\mu(g(t,A)) = \mu(A)$. Then the flow $g(t,x)$ is called *mixing* if for every open set $A, B \subseteq \Omega$

$$\mu(A \cap g(t,B)) \to \frac{\mu(A)\mu(B)}{\mu(\Omega)} \quad \text{as} \quad t \to \infty, \tag{1.6}$$

see fore example [13, 40]. Interpreting this definition it may be instructive to instead consider the limit $\mu(A\cap g(t,B))/\mu(A) \to \mu(B)/\mu(\Omega)$ as $t\to\infty$ and thus the relative amount of $A \cap g(t,B)$ in A should go to the relative amount of B in Ω measured with μ.

Related to mixing is the decay of correlations between open sets $A, B \subseteq \Omega$ defined by

$$C_t(A,B) = \mu(A\cap g(t,B)) - \frac{\mu(A)\mu(B)}{\mu(\Omega)}, \tag{1.7}$$

see for example [7, 40]. Its asymptotic behavior indicates whether the mapping is mixing and also the rate at which mixing occurs. The decay may be exponential $C_t(A,B) \sim e^{-\alpha t}$ or polynomial $C_t(A,B) \sim t^{-\alpha}$ for some $\alpha > 0$.

1.2. **Convection versus diffusion.** In order to get a feeling for mixing in general we now consider (1.2) for $\mathrm{Pe} \ll 1$ and $\mathrm{Pe} \gg 1$ and estimate a finite time such that $\|c(x,t) - \overline{c}_0\|_\infty$ in (1.1) supposedly is small. We call this the *mixing time*.

For $\mathrm{Pe} \ll 1$ the convection is negligible and $c(x,t)$ will mix in time for any initial data $c_0(x)$ due to diffusion. We consider the system mixed when the mean-square displacement, $\sqrt{Dt}$, of a diffusing non-interacting point mass (a molecule or particle) equals L and thus we obtain the relation

$$t_m = L^2/D. \tag{1.8}$$

For $\mathrm{Pe} \gg 1$ the diffusion is negligible but instead we assume that the velocity field u is mixing and that the flow is dynamically unstable in the sense that small perturbations of size ℓ will grow exponentially as $\ell \exp(\sigma t)$ for some $\sigma > 0$. We choose $\ell = (\tau D)^{1/2}$, where $\tau = L/U$ is the characteristic time for the flow and consider the system mixed when $\ell \exp(\sigma t) = L$ which implies that

$$t_m = 1/(2\sigma) \ln(\mathrm{Pe}). \tag{1.9}$$

where we used $\ell/L = (D/LU)^{1/2} = \mathrm{Pe}^{-1/2}$.

Example 1.1. We consider the mixing of a dilute water/particle dispersion with pure water, that is, let $A \subset \Omega$ contain the water/particle dispersion and $\Omega \setminus A$ contain water at time $t = 0$. We suppose the particles diffuse with diffusion constant given by the Einstein relation

$$D = \frac{k_B T}{6\pi\eta a}, \tag{1.10}$$

where k_B is the Boltzmann constant, T is the temperature, η is the viscosity of water, and a is the particle size. Now for

$$\begin{aligned} k_B &= 1.38 \times 10^{-23}\ [\mathtt{J/K}], \\ T &= 298\ [\mathtt{K}], \\ \eta &= 1.04 \times 10^{-3}\ [\mathtt{Ns/m^2}], \end{aligned}$$

and a in the range $[10^{-10}, 10^{-5}]$ [$\mathtt{m}$] we plot the mixing time by diffusion (1.8) for various L in Figure 1.1, that is, in the absence of convection.

Suppose next there is a velocity field in Ω generating a flow that is mixing, cf. Section 1.1. The velocity field is characterized by the *Reynolds number*

$$\mathrm{Re} = \frac{UL}{\mu} \tag{1.11}$$

where $\mu = \eta/\rho$ is the kinematic viscosity and ρ is the density. Notice that $\mathrm{Pe} = \mathrm{Re}\,\mu/D$ and hence with the Einstein relation (1.10) we rewrite (1.9)

as

$$t_m = 1/(2\sigma)\ln\left(\mathrm{Re}\,\frac{6\pi\eta^2}{k_B T\rho}a\right).$$

Now for

$$\mathrm{Re} = 1,$$
$$\rho = 10^3\ [\mathtt{kg/m^3}],$$

and for the other parameters as above, we plot the mixing time by convection for various σ in Figure 1.1, that is, in the absence of diffusion. We note that the dependence on the Reynolds number is logarithmic and hence the actual choice of Re is not so important. However the choice we made, $\mathrm{Re} = 1$, reflect that there is no turbulence in the velocity field. Examining Figure 1.1 we note that there are regimes where mixing by diffusion is very slow, $t_m > 10^3$ [s], whereas the mixing by convection is faster, $t_m > 1$ [s], this is the regime we particularly is interested in for this work.

1.3. **Mixing in micro fluid systems.** Mixing as outlined in the sections above, and also from the practical point of view, has recently be revived, spurred by the development of *microfluidics*, see the book [23] for a general reference and the review articles [20, 30] on mixing in microfluid systems, where "micro" refers to length scales $L \lesssim 1$ [μm].

Significant for many micro fluid systems is the combination of small Reynolds numbers $\mathrm{Re} \lesssim 10$ (possibly $\ll 10$) and large Péclet numbers $\mathrm{Pe} \gtrsim 100$. As a consequence it is difficult to mix fluids in these systems. For small Reynolds numbers there is no turbulence and we cannot rely on inertial effects for mixing, and for large Péclet numbers convective effects dominate over diffusive effects and mixing by diffusion is a relatively slow process, see the discussion in Section 1.2.

Consider for example stationary flow in channels with characteristic length scale L, the diameter (width) of the channel, and characteristic velocity U, the maximum velocity in the flow along the channel. The objective is to mix two miscible fluids A and B that enter in a Y-junction, Figure 1.2. If L and U are sufficiently small and the channel walls are smooth, then a parabolic velocity profile is maintained along the channel. There will be little mixing and only due to diffusion. This has been demonstrated in several papers, for example, in [21, 22, 26] and we illustrate it in Figure 1.2.

One possible way to enhance the mixing in these situations is to force convection in the cross section by modifying the geometry of the boundary

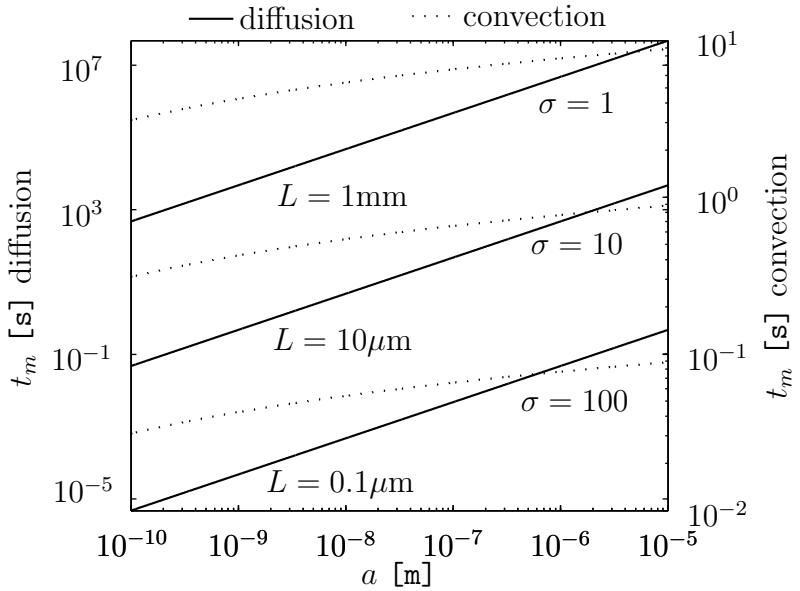


Figure 1.1: Mixing time, t_m, of a dilute water-particle dispersion as a function of particle size a. The **left** y-axis shows the mixing time by diffusion based on the estimate (1.8) for $L = (10^{-7}, 10^{-5}, 10^{-3})$ [m] and where we assumed the diffusion coefficient is given by the Einstein relation (1.10). The **right** y-axis shows the mixing time by convection based on the estimate (1.9) for $\sigma = (1, 10, 100)$ [s^{-1}] and where we assumed $\mathrm{Re} = 1$.

$\partial\Omega$. This was proposed and experimentally studied in [38] in channels like the ones depicted in Figures 2.1 and 2.2, and with velocity profiles as in Figures 2.3 and 2.4, where the velocity field is given by the Stokes equations as described in Section 2.1 below.

2. Outline of the thesis

The goal of this thesis is to computationally characterize mixing in the context outlined above. Ideally, this implies that, for all open sets $A \subset \Omega$, we need to

(1) solve (1.5) for every $x \in A$,
(2) compute the decay of correlations (1.7).

However, in practice we can not expect to solve (1.5) for every $x \in A$ and all open sets $A \subset \Omega$ and compte (1.7) for all open sets $A, B \subset \Omega$.

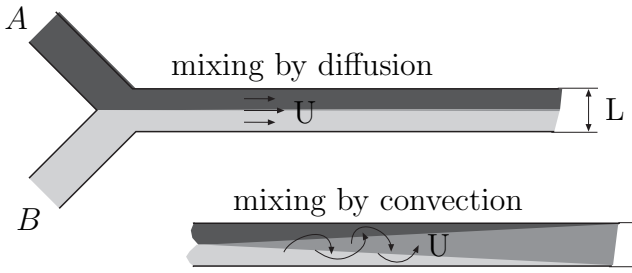


Figure 1.2: Miscible fluids A and B enter the channel in a Y-junction. **(top)** In microfluidics the characteristic length L and the characteristic velocity U, are both typically small. Inertial effects are weak and if the channel walls are smooth the two fluids will flow along next to each other. There will be little mixing and only due to diffusion. **(bottom)** Introducing convection in the channel cross section may enhance mixing.

Therefore we must make some approximations and set up some additional framework.

We assume that u is sufficiently smooth and incompressible, that is,

$$\nabla \cdot u(x) = 0 \quad \forall x \in \Omega.$$

Let ν be the outward normal to $\partial\Omega$, the boundary to Ω. We assume that $\nu(x) \cdot u(x) = 0$ for $x \in \partial\Omega$, that is, there is no flow through $\partial\Omega$. Imposing some additional constraints on Ω and u, we distinguish two types of flows.

(1) The flow $g(t, x)$ is said to be *confined* if Ω is bounded.

Let $\Gamma \subset \Omega$ with $\dim \Gamma = n - 1$ such that $\partial\Gamma \subset \partial\Omega$ and $\Gamma + mr \in \Omega$ for some $r \in \mathbf{R}^n$ and any integer m.

(2) The flow $g(t, x)$ is said to be *space periodic* if $u|_\Gamma = u|_{\Gamma+mr}$ and $\nu_\Gamma(x) \cdot u(x) \leq 0$ for $x \in \Gamma$, where ν_Γ is the normal to Γ

We now describe the procedure in the context of confined flows for which the volume measure $|\cdot|$ will be the appropriate measure.

Instead of considering all open sets $A \subset \Omega$ we consider a partition $\{A_i\}_{i=1}^N$ of Ω and

(1) compute a limited number of orbits $g_k(t, x_j)$ to (1.5) for $x_j \in A_i$ and $j = 1, \ldots, M$,

(2) for all A, B in the partition compute the approximate decay of correlations as explained in **Paper I** and defined by

$$C_{k,t}^M(A,B) = |B|M^{-2}\sum_{i=1}^{M}\det(\nabla g_k(t,x_i))\sum_{j=1}^{M}\chi_A(g_k(t,x_j)) - |A||B| \tag{2.1}$$

where χ_A is the characteristic function for A.

Moreover, in many situations we may not even know u a priori in a closed form but it will rather be defined from a model, for example, a partial differential equation, and we will have to use approximate data u_h for u, where h denotes the space discretization.

Finally, in order to validate the entire approach we provide error bounds for the approximate correlation sequence in the sense that

$$|C_{k,t}^M(A,B) - C_t(A,\widetilde{B})| \leq \varepsilon\big(|\partial A| + |A||\partial B|\big) + R_M, \tag{2.2}$$

where $\widetilde{B}$ a set close to B and R_M is a residual from the approximation of $|A \cap T^t B|$ by quadrature and ε is a parameter depending on estimates of:

(1) the error in the computed velocity field

$$e_u = u_h - u,$$

(2) the error in the computed orbits

$$e = g_k(t,x) - g(t,y),$$

where $x \in B$ is probably not equal to $y \in \widetilde{B}$, but still we may obtain a small e due to a shadowing argument.

The overall methodology described above is more carefully discussed in **Paper I** and the error estimates for e_u and e are outlined in the section below.

2.1. **Stokes flow.** Modeling the velocity field we only consider the Stokes equations, see [18, 37] for general mathematical introductions. Let $\Omega \subset \mathbf{R}^n$, $n = 2, 3$, be a polyhedral domain and consider the Dirichlet Stokes problem in dimensionless form

$$\begin{aligned} -\Delta u + \nabla p &= f \quad \text{in}\,\Omega, \\ \nabla \cdot u &= g \quad \text{in}\,\Omega, \\ u &= 0 \quad \text{on}\,\partial\Omega, \end{aligned} \tag{2.3}$$

where $u = (u_1, \ldots, u_n)$ is the unknown velocity field, p is the unknown pressure, $f = (f_1, \ldots, f_n)$ is an external body force and g is a function describing the compressibility of the flow, for incompressible flows $g = 0$.

2.1.1. *A posteriori error estimates.* Let $\mathcal{T}$ be a regular triangulation of Ω and for $T \in \mathcal{T}$ set $h_T = \operatorname{diam}(T)$ and $h_{\min} = \min_{T\in\mathcal{T}} h_T$. Let (u_h, p_h) be a conforming finite element solution to (2.3) and define the residual in the momentum equation (me) by

$$R_{\text{me}} := f + \Delta u_h - \nabla p_h$$

and the residual in compressibility constraint (cc) by

$$R_{\text{cc}} := g - \nabla \cdot u_h.$$

In **Paper II**, which is inspired by [14, 31], we prove that provided the data f and g are sufficiently smooth there is a constant C such that

$$\|u_h - u\|_{L^\infty(\Omega)} \leq C|\log h_{\min}|^{\alpha_n}\eta + C_1 h_{\min}^{\beta}, \tag{2.4}$$

where $\alpha_2 = 2$, $\alpha_3 = 4/3$ and

$$\eta = \max_{T\in\mathcal{T}}\Big(h_T^2\|R_{\text{me}}\|_{\infty,T} + \frac{1}{2}\|[\partial_\nu u_h]\|_{\infty,\partial T\setminus\partial\Omega} + h_T\|R_{\text{cc}}\|_{\infty,T}\Big),$$

where $[\partial_\nu u_h]$ denotes the jump across ∂T in the normal derivative, $\partial_\nu u_h = \nu \cdot \nabla u_h$, where ν denotes the outward normal to ∂T, and where β could be chosen arbitrary large.

We note that the estimate above is asymptotic in the sense that the constant C is bounded but not explicitly known, and thus the error goes to zero as h_T goes to zero.

2.1.2. *Periodic Stokes flows.* In the actual mixing experiments we consider a model problem inspired by [38] where laminar fluid mixing was experimentally studied in small channels. Let $\Omega \subset \mathbf{R}^3$, be a polyhedral domain with periodic boundaries Γ_A and Γ_B, as in Figures 2.1 and 2.2, and consider the following Stokes problem with periodic boundary conditions in dimensionless form

$$\begin{aligned} -\Delta u + \nabla p &= 0 \quad \text{in } \Omega, \\ \nabla \cdot u &= 0 \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega \setminus (\Gamma_A \cup \Gamma_B), \\ u|_{\Gamma_A} &= u|_{\Gamma_B}, \\ p|_{\Gamma_A} &= p|_{\Gamma_B} + R_p, \end{aligned} \tag{2.5}$$

where R_p is a constant modeling the pressure drop.

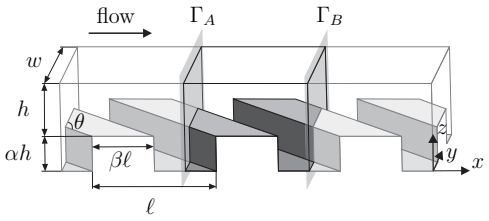


Figure 2.1: Three juxtaposed Ridge Domains. The shaded planes Γ_A and Γ_B are periodic boundaries.

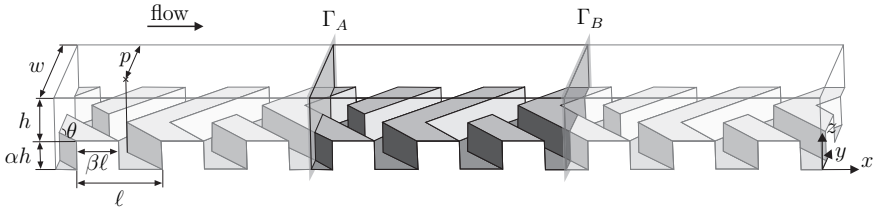


Figure 2.2: Three juxtaposed Herringbone Domains. The shaded planes Γ_A and Γ_B are periodic boundaries.

We refer to the domains in Figures 2.1 and 2.2 as the Ridge Domain and the Herringbone Domain, respectively, where the names are quoted from [38]. Accurate solutions to (2.3) in the two domains are computed by a finite element method, see for example [9, 11, 17] for general texts and [18, 29] for Stokes equations. We use Taylor-Hood P_2P_1 finite elements on fine triangulations and illustrate the solutions in Figure 2.3 and 2.4.

2.2. **Finite time shadowing.** Given a dynamical system (1.5) and a number $\mathrm{Tol} > 0$, we may ask if there in practice is a threshold time T such that we can compute orbits $g_k(t,x)$ with error $\|g_k(t,x) - g(t,x)\| \leq \mathrm{Tol}$ for all $t \in [0,T]$ and a fixed $x \in \Omega$, that is, such that the error is uniformly bounded on $[0,T]$.

In the present case accurate predictions of this kind are inherently difficult since mixing is only obtained if u is sufficiently irregular meaning that the flow generated by u will have to be sufficiently hyperbolic, which

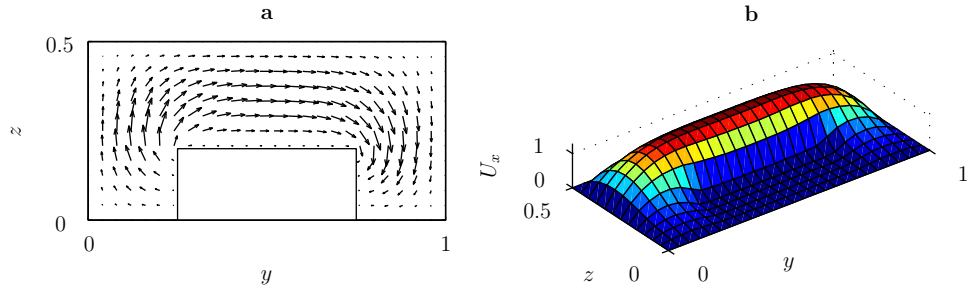


Figure 2.3: Velocity field for (2.3) solved in the Ridge Domain, Figure 2.1, at $x = 0.0$. **(a)** The y and z components of the velocity field. **(b)** The x component of the velocity field.

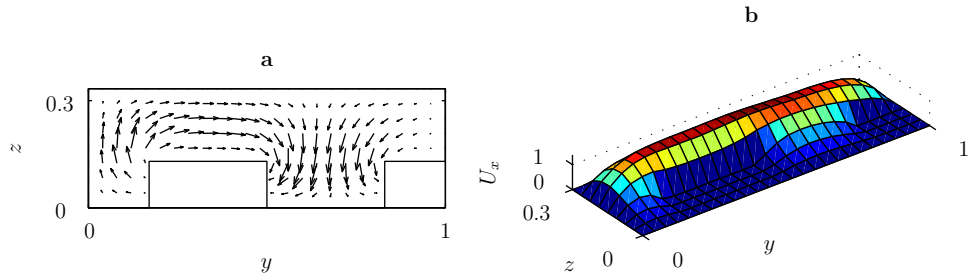


Figure 2.4: Velocity field for (2.3) solved in the Herringbone Domain, Figure 2.2, at $x = 0.0$. **(a)** The y and z components of the velocity field. **(b)** The x component of the velocity field.

loosely speaking involves that the flow has to have enough contractive and expensive directions, see [24, 40] for a more precise statement of hyperbolicity. Such systems are dynamically unstable, sensitive to perturbations, which renders the computation delicate. The error will probably grow at a exponential rate and we will only be able to compute accurate orbits $g_k(t, x)$ for relatively small time intervals.

However, if the system is sufficiently hyperbolic we may argue by shadowing, that is, provided $g_k(t, x)$ is computed accurately enough there is an exact orbit $g(t, y)$ such that $\|g_k(t, x) - g(t, y)\|$ is small for $t \in [0, T]$ and where T is relatively large [10, 12, 39]. Moreover approximating u with u_h

in (1.5) and if e_u is small we may arrive to the similar conclusion, which is proved in **Paper III**.

We solve (1.5) by the finite element method, see for example [16]. Partition $[0,T]$ into intervals $[t_{i-1},t_i]$ for $i=1,2,\dots,N$, and set $k_i=t_{i-1}-t_i$, and let $k=k(t)$ be piecewise constant function defined by $k|_{[t_{i-1},t_i]}=k_i$.

Define the residual to the finite element solution

$$R(g_k) := u_h(g_k) - \partial_t g_k.$$

Set

$$r_N(u_h,\rho) := \rho h_{\min}^{-1-n/p}\|\nabla u_h\|_{L^p(\Omega)} + \max_{T\in\mathcal{T}}\|[\nabla u_h]\|_{L^\infty(\partial T\setminus\partial\Omega)},$$

and

$$r_E(e_u) := h_{\min}^{-n/p}\|\nabla e_u\|_{L^p(\Omega)}. \tag{2.6}$$

for some $1\le p\le\infty$ depending of what kind of estimates we have for u_h, and set

$$\mathcal{B}_\rho := \{e\in C^1([0,T]) : \|e\|_{L^\infty([0,T])}\le\rho\}.$$

Let $S_1(T)$ and $S_2(T)$ be stability factors obtained by solving an appropriate dual problem and let ρ, u_h and g_k be such that

$$\begin{aligned} CS_2(T)r_E(e_u) &\le 1/4,\\ CS_2(T)r_N(u_h,\rho) &\le 1/4, \end{aligned} \tag{2.7}$$

and suppose

$$\begin{aligned} S_1(T)\|k^{q+1}R(g_k)\|_{L^\infty([0,T])} &\le \frac{1}{4}\rho,\\ S_2(T)\|e_u\|_{L^\infty(\Omega)} &\le \frac{1}{4}\rho. \end{aligned}$$

Then the numerical solution $g_k(t,x)$ is shadowed by an exact solution $g(t,y)$ and the error $e(t)=g_k(t,x)-g(t,y)$ is bounded from above for all $t\in[0,T]$ by

$$|e(t)| \le S_1(T)\|k^{q+1}R(g_k)\|_{L^\infty([0,T])} + S_2(T)\|e_f\|_{L^\infty(\Omega)} \le \rho. \tag{2.8}$$

We note that $S_{1,2}(T)$ will be relatively small provided the flow (1.5) is sufficiently hyperbolic.

Example 2.1. As a concrete example we consider the Lorenz system

$$\begin{aligned} \partial_t g &= (\sigma(g_2-g_1), \rho g_1 - g_2 - g_1 g_3, g_1 g_2 - \beta g_3), \quad t>0;\\ g(0) &= (1,0,0); \quad \text{for} \quad (\sigma,\rho,\beta)=(10,28,8/3). \end{aligned}$$

In [28] this problem was solved accurately in the sense that the error $\|g_k(t, g(0)) - g(t, g(0))\|$ is small up to $T = 50$, which is predicted to be the threshold beyond which the error becomes too large to be represented with double precision arithmetics (in the same work the threshold $T = 100$ is predicted for quadruple precision).

This result should be compared to [12] where the same problem is solved accurately up to $T = 9 \times 10^6$ in the sense that $\|g_k(t, g(0)) - u(t, y)\|$ is small for $t \in [0, T]$, that is, very close to the computed orbit $g_k(t, g(0))$ there is an exact orbit $g(t, y)$.

This example obviously suggests that long time error control for problems that are dynamically unstable will fail with the first method, but could possibly be archived with the last method, provided the structure of the problem is sufficiently 'hyperbolic-like'.

2.3. **Search in triangulations.** Working with finite element methods in practice we may face the problem to locate which n-simplex in the triangulation contains a given point. In the present work we need to perform such search when solving (1.5) with finite element data in the right hand side. A simple search will require $O(N)$ operation if N is the number of n-simplices in the triangulation and thus if we for some reason must solve this problem many times we would like do the search more efficiently.

In **Paper IV** we discuss the implementation of a binary search algorithm that will solve the search problem in an optimal way, that is, with $O(N)$ preprocessing time, $O(N \log N)$ storage and $O(\log N)$ search time [25].

2.4. **Multigrid solvers.** Finite element multigrid methods solve linear systems of equations arising from finite element approximations to linear elliptic partial differential equations with the number of operations proportional to the number of unknowns, see [8, 19, 36] for comprehensive introductions. We say that the multigrid method has optimal complexity or scales optimally.

However, it is important to note that this rather general statement is really implicitly assuming that the finite element basis functions are linear and that the triangulations are quasi-uniform. Side-stepping these requirements the convergence rate of the multigrid solver may deteriorate but the accuracy of u_h may be improved. For example, higher degree finite element approximation are appealing for use on problems that are sufficiently regular since the error $u - u_h$ may converge as $O(h^{q+1})$ where q is the degree of the approximation. Moreover, adaptively refined triangulations (violating

the quasi-uniformity) may give better approximations on problems with less regularity.

In **Paper V** we demonstrate that the multigrid method in practice also works well for second degree finite element approximations of problem with both full regularity and less than full regularity. We compare two different finite element approximations, the Lagrange approximation and the hierarchical approximation proposed in [5] and [4]. We use the general theory outlined in [8] to indicate how the point Gauss-Seidel smoother deteriorates as a function of the dimension n of the problem and the degree of the approximation q.

In **Paper VI** we consider the practical aspects implementing the method on adaptively refined triangulations for conforming linear and quadratic finite elements in two and three dimensions.

We choose to use a refinement method that renders the finite element spaces nested and thus the formulation of the multigrid method is straightforward with well defined projection operators on the finite element spaces, in contrast to the situation when the finite elements spaces are non-nested [8, 35]. Moreover, this choice is also motivated by the fact that the refinement algorithm becomes simple compared to the rather involved refinement algorithm proposed in [6], which also renders the finite element spaces non-nested.

The refined triangulations are irregular [15] in the sense that there will be 'hanging' nodes and the construction of conforming finite element spaces is a non-trivial task that in practice requires flexible data structures to be implemented. This and the even more general aspects of *hp*-refinements has already be considered in [1, 15, 32]. We reformulate these results using concepts from modern finite element theory.

3. Concluding remarks

In this thesis we have computationally characterize mixing in flows in the sense that we now in principle quite generally should be able to analyze mixing. Principal to the work is to compute the velocity field u by the finite element method and provide error bound to the computed solution. We note that estimate (2.4) is asymptotic in character, that is, there is an unknown but bounded constant in the right hand side and we can only deduce that the error goes to zero as $h \to 0$. It may be possible to obtain other error estimates with better control on the constant in the right hand

side but then we must probably measure the error in some other weaker norm than the maximum-norm.

Better control on error in the velocity field will impact and improve the other estimates: the shadowing estimate (2.8) and the estimate of the computed mixing measure (2.2). Thus, error control in the velocity field is a key issue.

We remark that ideally we should have solved the Stokes problems (2.5) adaptively with the techniques outlined in Paper V and VI. However we did not manage to completely finish this part of the work.

References

[1] M. Ainsworth and B. Senior, *Aspects of an adaptive hp-finite element method: adaptive strategy, conforming approximation and efficient solvers*, Comput. Methods Appl. Mech. Engrg. **150** (1997), 65–87.

[2] H. Aref, *Stochastic particle motion in laminar flows*, Phys. Fluids A **3** (1991), 1009–1016.

[3] ———, *The development of chaotic advection*, Phys. Fluids **14** (2002), 1315–1325.

[4] O. Axelsson and I. Gustafsson, *Preconditioning and two-level multigrid methods of arbitrary degree of approximation*, Math. Comp. **40** (1983), 219–242.

[5] R. E. Bank, T. F. Dupont, and H. Yserentant, *The hierarchical basis multigrid method*, Numer. Math. **52** (1988), 427–458.

[6] J. Bey, *Tetrahedral grid refinement*, Computing **55** (1995), 355–378.

[7] C. Bonatti, L. J. Díaz, and M. Viana, *Dynamics beyond uniform hyperbolicity*, vol. 102, Springer-Verlag, 2005.

[8] J. H. Bramble and X. Zhang, *The Analysis of Multigrid Methods*, Handbook of Numerical Analysis, Vol. VII, North-Holland, 2000.

[9] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, second ed., Springer-Verlag, 2002.

[10] S-N. Chow and E. S. Van Vleck, *A shadowing lemma approach to global error analysis for initial value ODEs*, SIAM J. Sci. Comput. **15** (1994), 959–976.

[11] P. G. Ciarlet, *Basic error estimates for elliptic problems*, Handbook of Numerical Analysis, Vol. II, North-Holland, 1991.

[12] B. A. Coomes, H. Koçak, and K. J. Palmer, *Rigorous computational shadowing of orbits of ordinary differential equations*, Numer. Math. **69** (1995), 401–421.

[13] I. P. Cornfeld, S. V. Fomin, and Ya. G. Sinaĭ, *Ergodic Theory*, Springer-Verlag, 1982.

[14] E. Dari, R. G. Durán, and C. Padra, *Maximum norm error estimators for three-dimensional elliptic problems*, SIAM J. Numer. Anal. **37** (2000), 683–700.

[15] L. Demkowicz, J. T. Oden, W. Rachowicz, and O. Hardy, *Toward a universal h-p adaptive finite element strategy. I. Constrained approximation and data structure*, Comput. Methods Appl. Mech. Engrg. **77** (1989), 79–112.

[16] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Computational Differential Equations*, Cambridge University Press, 1996.
[17] A. Ern and J. L. Guermond, *Theory and Practice of Finite Elements*, Springer-Verlag, 2004.
[18] V. Girault and P. A. Raviart, *Finite Element Methods for Navier-Stokes Equations*, Springer-Verlag, 1986.
[19] W. Hackbusch, *Multigrid Methods and Applications*, Springer-Verlag, 1985.
[20] V. Hessel, H. Lowe, and F. Schonfeld, *Micromixers-a review on passive and active mixing principles*, Chem. Eng. Sci. **60** (2005), 2479–501.
[21] R. F. Ismagilov, A. D. Stroock, P. J. A. Kenis, H. A. Stone, and G. Whitesides, *Experimental and theoretical scaling laws for transverse diffusive broadening in two-phase laminar flows in microchannels*, Appl. Phys. Lett. **76** (2000), 2376–8.
[22] A. E. Kamholz and P. Yager, *Theoretical analysis of molecular diffusion in pressure-driven laminar flow in microfluidic channels*, Biophys. J. **80** (2001), 155–60.
[23] G. E. Karniadakis and A. Beskok, *Micro Flows*, Springer-Verlag, 2002.
[24] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge University Press, 1995.
[25] D. Kirkpatrick, *Optimal search in planar subdivisions*, SIAM J. Comput. **12** (1983), 28–35.
[26] J. B. Knight, A. Vishwanath, J. P. Brody, and R. H. Austin, *Hydrodynamic focusing on a silicon chip: mixing nanoliters in microseconds*, Phys. Rev. Lett. **80** (1998), 3863–6.
[27] A. Lasota and M. C. Mackey, *Chaos, Fractals, and Noise*, second ed., Springer-Verlag, 1994.
[28] Anders Logg, *Multi-adaptive Galerkin methods for ODEs. II. Implementation and applications*, SIAM J. Sci. Comput. **25** (2004), 1119–1141.
[29] M. Marion and R. Temam, *Navier-Stokes Equations: Theory and Approximation*, Handbook of Numerical Analysis, Vol. VI, North-Holland, 1998.
[30] Nam-Trung N. and Zhigang W., *Micromixers-a review*, J. Micromech. Microeng. **15** (2005), 1–16.
[31] R. H. Nochetto, *Pointwise a posteriori error estimates for elliptic problems on highly graded meshes*, Math. Comp. **64** (1995), 1–22.
[32] J. T. Oden, *A general theory of finite elements. II. Applications*, Int. J. Numer. Methods Eng. **1** (1969), 247–259.
[33] J. M. Ottino, *The Kinematics of Mixing: stretching, chaos, and transport*, Cambridge University Press, 1989.
[34] J.M. Ottino, *The mixing of fluids*, Sci. Am. **260** (1989), 40–49.
[35] L. R. Scott and S. Zhang, *Higher-dimensional nonnested multigrid methods*, Math. Comp. **58** (1992), 457–466.
[36] V. V. Shaĭdurov, *Multigrid Methods for Finite Elements*, Kluwer Academic Publishers Group, 1995.
[37] H. Sohr, *The Navier-Stokes Equations*, Birkhäuser Verlag, 2001.

[38] A. D. Stroock, S. K. W. Dertinger, A. Ajdari, I. Mezic, H. A. Stone, and G. M. Whitesides, *Chaotic mixer for microchannels*, Science **295** (2002), 647–51.

[39] E. S. Van Vleck, *Numerical shadowing using componentwise bounds and a sharper fixed point result*, SIAM J. Sci. Comput. **22** (2000), 787–801.

[40] L-S. Young, *Geometric and ergodic theory of hyperbolic dynamical systems*, Current Developments in Mathematics, Int. Press, Somerville, MA, 1999, pp. 237–278.

Paper I

COMPUTATIONAL CHARACTERIZATION OF MIXING IN INCOMPRESSIBLE FLOWS

ERIK D. SVENSSON

Abstract. We propose a computational methodology for characterizing fluid mixing in incompressible flows. Principal to the methodology is the definition of a mixing measure that will resolve the mixing process both in space and time. We propose a mixing measure based on rigorous notions, *mixing* and *decay of correlation* in the flow, known from dynamical systems theory. We analyse the error when the mixing measure is computed numerically and obtain an upper error bound for the mixing measure that in principle could be used for rigorous computational characterization of the mixing process.

1. Introduction

In order to mixing miscible fluids on a time scale where diffusion is negligible the fluids will have to be displaced by means of a velocity field that is sufficiently irregular. In the engineering literature such process is commonly refereed to *mixing by chaotic advection* and for further references we refer to the survey articles [1, 2, 21] or the book [20]. The problem has recently undergone a revival spurred by the development of microfluidics, see the book [15] for a general reference and the review articles [14, 19] on mixing in micro fluid systems.

We consider mixing as a relaxation process going from an unmixed state to a homogeneous (mixed) state and in order to characterize this relaxation we will need a measure that describes the mixing process in space and time. For example, the mixing could be nonuniform in space and we may be interested in resolving these spatial variations; or a process may be mixing although at a slow rate, too slow to be useful in an engineering application. There seems to be no consensus on what mixing measure to

Date: April 19, 2006.

2000 *Mathematics Subject Classification.* 37A25, 37C50, 76M10.

Key words and phrases. mixing, shadowing, finite elements, flow simulation.

use. The mixing measures suggested in the literature vary and are often heuristical, *cf.* [3, 12, 20].

In dynamical systems theory *mixing* has a precise meaning, see for example [16, 26] or the survey article [25]. Related to mixing in this context is the *decay of correlations* in the flow which we use as a mixing measure in this work. Since the decay of correlations in the flow, *cf.* (1.6) and (1.7), in principle is numerically intractable we propose a computable approximation and analyze the error in this approximation. Moreover, we consider a situation where the velocity field generating the mixing process is not known a priori in closed form but rather given by computed data from some model, *e.g.*, the Stokes equations or possibly the Navier-Stokes equations or some other fluid model. This aspect is also included in the error analysis. The error bound provided for the mixing measure can in principle be used for a rigorous computational characterization of the mixing process so that the total error is controlled and made small.

1.1. **Assumptions.** Let $\Omega \subset \mathbf{R}^d$ for $d = 2, 3$ be an open set containing some fluids. We assume that the fluids are moved by a sufficiently smooth velocity field $f : \Omega \to \mathbf{R}^d$ so that f generates a flow $[0, t] \times \Omega \ni (t, x) \mapsto u(t, x) \in \mathbf{R}^d$. The flow describes the motion of the fluid particles in Ω and is given by the solution to the system of ordinary differential equations

$$\partial_t u(t, x) = f(u(t, x)), \quad t > 0; \quad u(0, x) = x. \tag{1.1}$$

We also assume that f is incompressible in the sense that

$$\nabla \cdot f(x) = 0 \quad \forall x \in \Omega. \tag{1.2}$$

Let ν be the outward unit normal to $\partial\Omega$. We assume that $\nu(x) \cdot f(x) = 0$ for $x \in \partial\Omega$, that is, there is no flow through the boundary $\partial\Omega$. Imposing some additional constraints on Ω and f we distinguish two types of flows.

(1) The flow $u(t, x)$ is said to be *confined* if Ω is bounded.

Let $\Gamma \subset \Omega$ with $\dim \Gamma = d - 1$ such that $\partial\Gamma \subset \partial\Omega$ and $\Gamma + mr \in \Omega$ for some $r \in \mathbf{R}^d$ and any integer m, see Figure 1.1.

(2) The flow $u(t, x)$ is said to be *space-periodic* if $f|_\Gamma = f|_{\Gamma+mr}$ and $\nu_\Gamma(x) \cdot f(x) \leq 0$ for $x \in \Gamma$, where ν_Γ is the unit normal to Γ.

1.2. **Considerations.** We ask to what extent the mixing process can be computationally characterized in the sense that computed predictions are accurate. Suppose we compute a mixing measure, of our choice, that reflects the amount of mixing in Ω generated by f. Then we also would like

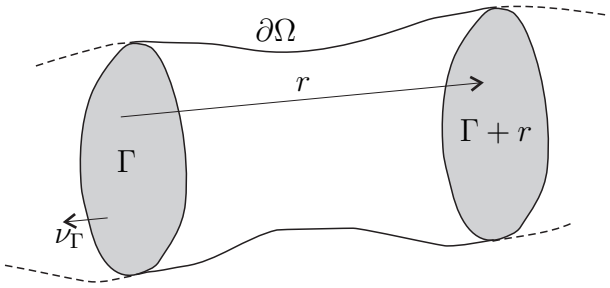


Figure 1.1: Space-periodic domain.

to estimate the error in the computed measure and recursively compute the mixing measure more accurately. Accurate predictions of this kind are inherently difficult since mixing requires that f is sufficiently irregular meaning that the flow generated by f must be hyperbolic-like which loosely involves that the flow has to have enough contractive and expansive directions, see [16, 26] for a more precise statement of hyperbolicity. Such systems are dynamically unstable, sensitive to perturbations, which renders the computation delicate.

In practice we will not know the flow $u(t, x)$ a priori in a closed form and in order to study the properties of the flow we may instead analyze a limited number of numerically computed orbits $u_k(t, x_j)$ for $j = 1, 2, \ldots, J$, where k denotes the time step. Moreover, in many situations we may not even know f a priori in a closed form but it will rather be defined from a model, for example, a partial differential equation, and we will have to use numerically computed data f_h for f where h denotes the space discretization. Now let $u_k(t, x_j)$ be a computed orbit to (1.1) with right hand side $f = f_h$. Then the error

$$e(t, x) := u_k(t, x_j) - u(t, x), \tag{1.3}$$

will depend on the discretization error associated with the numerical method used to compute $u_k(t, x_j)$ and the error in the velocity field $e_f := f_h - f$. Since (1.1) probably is dynamically unstable we will only be able to compute $u_k(t, x_j)$ with small $e(t, x)$ for a rather small time. However, if the system is hyperbolic-like we may argue by shadowing, that is, provided $u_k(t, x_j)$ is computed accurately enough and provided e_f is small enough there is an exact orbit $u(t, y)$ with an other initial value y such that

$\|u_k(t, x_j) - u(t, y)\|$ is small for $t \in [0, T]$ [7, 8, 24]. The overall idea is to use this kind of argument in order to control the error in a computed mixing measure.

1.3. **Notions from dynamical systems theory.** Within the realm of dynamical systems theory mixing has a precise meaning. For a probability space $(X, \mathcal{M}, \mu)$ a measure preserving bijective mapping $T : (X, \mathcal{M}, \mu) \to (X, \mathcal{M}, \mu)$ is called *mixing* if, for discrete time systems

$$\forall\, A, B \in \mathcal{M} \quad \mu(A \cap T^n B) \to \mu(A)\mu(B) \quad \text{as} \quad n \to \infty, \tag{1.4}$$

or for continuous time systems $T^t : (X, \mathcal{M}, \mu) \to (X, \mathcal{M}, \mu)$

$$\forall\, A, B \in \mathcal{M} \quad \mu(A \cap T^t B) \to \mu(A)\mu(B) \quad \text{as} \quad t \to \infty. \tag{1.5}$$

We remark that T is measure preserving if for every $A \in \mathcal{M}$, $\mu(T^{-1}A) \in \mathcal{M}$ and $\mu(T^{-1}(A)) = \mu(A)$, see for example [9, 26].

Mixing is also defined for measure preserving maps that are only surjective but then we must replace $\mu(A \cap T^n B)$ by $\mu(T^{-n}(A) \cap B)$ in the definition, and likewise for the continuous case, [9, 26].

Related to mixing is the *decay of correlations* between the sets $A, B \in \mathcal{M}$ defined by

$$C_n(A, B) = \mu(A \cap T^n B) - \mu(A)\mu(B) \tag{1.6}$$

for discrete time systems, and

$$C_t(A, B) = \mu(A \cap T^t B) - \mu(A)\mu(B) \tag{1.7}$$

for continuous time systems, see for example [4, 26]. The asymptotic behavior of the decay of correlations indicates whether the mapping is mixing or not and we may also have an estimate of the rate of mixing. The decay is exponential if $C_n(A, B) \sim e^{-\alpha n}$, or polynomial if $C_n(A, B) \sim n^{-\alpha}$, for some $\alpha > 0$, and likewise for $C_t(A, B)$.

1.4. **Computability.** In practice some of the notions in Section 1.3 are too general and numerically intractable. We will have to approximate $\mathcal{M}$, which in our case is a Borel σ-algebra on $X = \Omega$. It seems natural to replace $\mathcal{M}$ by a family of partitions $\{\mathcal{U}_h\}_{h>0}$ where $\mathcal{U}_h$ is the class of a finite number of disjoint sets U_i such that $\bigcup_i (U_i) = \Omega$ and where h denotes the size of the largest set in $\mathcal{U}$, that is, $h = \max_i \operatorname{diam}(U_i)$. In principle any type of partition will suffice although if Ω is a polyhedral domain it is convenient to let $\{\mathcal{U}_h\}_{h>0} = \{\mathcal{T}_h\}_{h>0}$ be a family of quasi-uniform triangulations, see for example [11].

The size h will determine the resolution of the approximation and could, for example, be motivated by some physical length scale, *i.e.*, set $h \sim (\tau D)^{1/2}$ which is the mean-square displacement of a diffusing non-interacting point mass, and where τ is a typical diffusion time scale and D is the diffusion constant.

In order to investigate whether a mapping is mixing we may consider the decay of correlations $C_n(A, B)$ (or $C_t(A, B)$) for $A, B \in \mathcal{U}_h$ and for some finite n. We will have to evaluate the measure $\mu(A \cap T^n B)$ (or $\mu(A \cap T^t B)$) which is inherently difficult since the mapping is probably dynamically unstable and even though B may have a simple geometry $T^n B$ will be severely deformed. It is reasonable to assume that we only know $T^n x_j$ for a finite number of $x_j \in B$, $j = 1, \ldots, M$. It then seems viable to evaluate $\mu(A \cap T^n B)$ by a simple Monte Carlo method. We will discuss this kind of implementation in more detail in the sections below where the measure μ is chosen explicitly.

1.4.1. *Monte Carlo integration.* For further reference we now briefly recall the Monte Carlo method, see for example [17]. Consider integrable functions f and g on $\omega \in \mathcal{M}$ such that $g \geq 0$ and

$$\int_\omega g(x)\, dx = 1,$$

and independent random variables $\{x_j\}_{j=1}^M$ that are $g(x)\, dx$ distributed on ω. Then

$$\int_\omega f(x)g(x)\, dx = \frac{1}{M} \sum_{j=1}^{M} f(x_j) + R_M(\sigma), \tag{1.8}$$

where $R_M(\sigma)$ is a residual that must be interpreted statistically in the sense that R_M is normally distributed with standard deviation $\sigma/\sqrt{M}$, where

$$\sigma^2 = \int_\omega f(x)^2\, dx - \left(\int_\omega f(x)\, dx \right)^2$$

is the variance. In practice we may estimate this variance by the empirical variance

$$\hat{\sigma}^2 = \frac{1}{M-1} \sum_{j=1}^{M} (f(x_j) - \bar{f})^2$$

where $\bar{f}$ is the mean of $\{f(x_j)\}_{j=1}^M$.

In the sequel we will use $f = \chi_A$, the characteristic function defined by

$$\chi_A(x) = \begin{cases} 1, & \text{if } x \in A, \\ 0, & \text{if } x \notin A. \end{cases} \tag{1.9}$$

1.5. **Error analysis.** We compute approximate orbits $u_k(t, x_j)$ to (1.1) for $t \in [0, T]$ and $x_j \in B$, $j = 1, \ldots, M$ by a continuous finite element method. This involves partitioning $[0, T]$ into intervals $I_i = [t_{i-1}, t_i]$ with $0 = t_0 < t_1 < \ldots < t_N = T$ and $k_i = t_i - t_{i-1}$. Let $P_q(I_i)$ denote the polynomials of degree less or equal to q on I_i and set

$$V_q([0,T]) := \{v \in C^0([0,T]) : v|_{I_i} \in P_q(I_i) \text{ for } i = 1 \ldots N\}$$

$$W_q([0,T]) := \{v \in C^0(\bigcup_{i=1}^{N}(t_{i-1}, t_i)) : v|_{I_i} \in P_q(I_i) \text{ for } i = 1 \ldots N\}$$

which is the finite element spaces of continuous and discontinuous piecewise polynomials of degree q.

For $q \geq 1$ we now obtain the finite element formulation to (1.1) with $f = f_h$, an approximate velocity field. Find $u_k \in V_q([0,T])^n$ with $u_k(t, 0) = x_j$ such that

$$\int_0^T (\partial_t u_k - f_h(u_k)) \cdot v \, dt = 0 \quad \forall v \in W_{q-1}([0,T])^n. \tag{1.10}$$

This is the continuous Galerkin method of degree q, referred to as the cG(q) method in [10, p. 210].

There are $q+1$ points in the interval I_i, where the piecewise polynomials are evaluated, referred to as local nodes. In the same way there are $N(q+1) - 1$ points in the interval $[0, T]$ referred to as global nodes.

We now assume that $u_k(t, x_j)$ is computed sufficiently accurately and that e_f is sufficiently small. A precise statement of this can be found in [10] for general finite element approximations and particularly in the present situation in [24]. The condition requiring $u_k(t, x_j)$ to be computed sufficiently accurately can be translated to $u_k(t, x_j)$ being a *pseudo orbit* *cf.* [16, 22]. We will use this notion in the sequel and in addition, when it is not explicitly stated, we always assume that e_f is sufficiently small.

If Ω is hyperbolic-like for (1.1) then every pseudo orbit $u_k(t, x_j)$ will be shadowed by an exact orbit $u(t, y_j)$ at least for some finite time $t \in [0, T]$, *cf.* [7, 8, 22, 24]. This implies that $\|u_k(t, x_j) - u(t, y_j)\|$ can be made small for $t \in [0, T]$.

Now for every pseudo orbit $u_k(t, x_j)$ for $t \in [0, T]$ and $x_j \in B$, $j = 1, \ldots, M$, we thus assume that there is a shadow orbit $u(t, y_j)$ such that

$$\|u_k(t, x_j) - u(t, y_j)\| \leq \varepsilon_j, \tag{1.11}$$

for some small number ε_j and we set

$$\varepsilon = \max_j \varepsilon_j. \tag{1.12}$$

By setting $t = 0$ in (1.11) we note that

$$y_j \in \widetilde{B} := \bigcup_{x \in B} \mathcal{B}(x, \varepsilon), \tag{1.13}$$

where $\mathcal{B}(x, \varepsilon)$ is the ball or radius ε about x.

2. Mixing in confined incompressible flows

Let $u(t, x)$ be a confined incompressible flow as defined in Section 1.1. For an open set $A \subseteq \Omega$ we define the measure as the volume of A normalized with the volume of Ω

$$\mu(A) = c_0^{-1} \int_A dx = c_0^{-1}|A|. \tag{2.1}$$

where

$$c_0 = \int_\Omega dx = |\Omega|,$$

is the volume of Ω. Then $(\Omega, \mathcal{M}, \mu)$ is a probability space.

Now set $T^t(\cdot) = u(t, \cdot)$. Since u is a flow and f is incompressible T is bijective and measure preserving, *i.e.*, $|A| = |T^t A|$ for every $A \in \mathcal{M}$ or, in other words, we say that T^t preserves volume. Hence mixing according to (1.5) and the decay of correlations (1.7) is well defined in this case.

2.1. Computational characterization.

Set $T_k^t(x) = u_k(t, x)$ and let $u_k(t, x_j)$ for $t \in [0, T]$ and $x_j \in B \subseteq \Omega$ be pseudo orbits. Approximate $\mathcal{M}$ with a partition $\mathcal{U}_h$ as defined in Section 1.4.

In order to approximately compute (1.7) we let $\{x_j\}_{j=1}^M$ be independent random variables uniformly distributed on B. Now $\{T_k^t(x_j)\}_{j=1}^M$ will be independent random variables $T_k^t(x)\,dx$ distributed on $T_k^t B$ and we compute $|A \cap T_k^t B|$ by the Monte Carlo method. Set $f(x) = \chi_A(x)$ and

$g(x) = |T_k^t B|^{-1}$ in (1.8) to obtain

$$|A \cap T_k^t B| = \int_{T_k^t B} \chi_A(x)\, dx \approx |T_k^t B| M^{-1} \sum_{j=1}^{M} \chi_A(T_k^t x_j).$$

It may seem tempting to set $|T_k^t B| = |B|$ but since T_k^t in general is not measure preserving we cannot do so. Instead we evaluate $|T_k^t B|$ by Monte Carlo integration and by a change of variables we obtain

$$|T_k^t B| = \int_{T_k^t B} dx = \int_B |\det(\nabla T_k^t x)|\, dx \approx |B| M^{-1} \sum_{j=1}^{M} |\det(\nabla T_k^t x_j)|,$$

where we note that $|\det(\nabla T_k^t x)| = 1$ if is T_k^t is measure preserving, *i.e.*, if f_h is incompressible [6, p. 10].

Hence, we define the following approximation to the decay of correlation (1.7) for $A, B \in \mathcal{U}_h$

$$C_{k,t}^M(A, B) = |B| M^{-2} \sum_{i=1}^{M} |\det(\nabla T_k^t x_i)| \sum_{j=1}^{M} \chi_A(T_k^t x_j) - |A||B|. \tag{2.2}$$

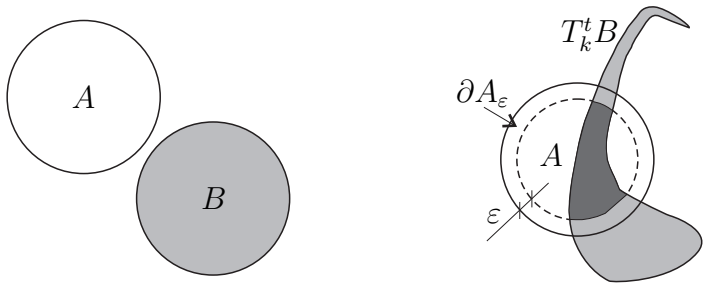


Figure 2.1: (left) $A, B \subset \Omega$ at $t = 0$. **(right)** Intersection $A \cap T_k^t B$ and the ε-shell ∂A_ε inside A.

2.2. **Error analysis.** As outlined in Section 1.5 we assume that to every pseudo orbit $u_k(t, x_j)$ there is an exact orbit $u(t, y_j)$ such that (1.11) is satisfied and that ε in (1.12) is small. Then for $A, B \in \mathcal{U}_h$ and $\widetilde{B}$ as defined in (1.13) we argue that

$$\big| |A \cap T_k^t B| - |A \cap T^t \widetilde{B}| \big| \leq |\partial A_\varepsilon \cap T_k^t B| \leq \varepsilon |\partial A|$$

where ∂A_ε is the ε-shell inside A as in Figure 2.1 and $|\partial A_\varepsilon \cap T_k^t B|$ is the measure of the points $T_k^t x \in T_k^t B$ such that $\operatorname{dist}(T_k^t x, \partial A) \leq \varepsilon$. In the same way

$$\big||B| - |\widetilde{B}|\big| \leq \varepsilon |\partial B|,$$

and hence we may estimate

$$\big||A \cap T_k^t B| - |A||B| - C_t(A, \widetilde{B})\big| \leq \varepsilon\big(|\partial A| + |A||\partial B|\big).$$

Now with the estimate above and for $A, B \in \mathcal{U}_h$ and $\widetilde{B}$ as defined in (1.13) we estimate the error in the approximate correlation sequence (2.2)

$$\big|C_{k,t}^M(A, B) - C_t(A, \widetilde{B})\big| \leq \varepsilon\big(|\partial A| + |A||\partial B|\big) + R, \tag{2.3}$$

where R must be interpreted statistically as explained in Section 1.4.1.

3. Mixing in periodic channel flows

Let $u(t, x)$ be a space periodic incompressible flow as defined in Section 1.1. For an open set $A \subseteq \Gamma$ we define the measure as the flow through A normalized with the flow trough Γ

$$\mu(A) = c_0^{-1} \int_A f \cdot \nu_\Gamma \, dx, \tag{3.1}$$

where

$$c_0 = \int_\Gamma f \cdot \nu_\Gamma \, dx,$$

is the total flow through Γ. Then $(\Gamma, \mathcal{M}, \mu)$ is a probability space.

Now for $u(0, x) \in \Gamma$ let t be such that $u(t, x) \in \Gamma + r$ and let $T : \Gamma \to \Gamma$ be the mapping defined by $T(x) := u(t, x) - r$. We note that since u is a flow and f is incompressible T is bijective and measure preserving. Iterating $T^n(x) = T \circ T^{n-1}(x)$ with $T^0(x) = x$ we obtain the Poincaré map for which mixing according to (1.4) and the decay of correlations (1.6) is well posed.

3.1. **Computational characterization.**

We need to define an approximate measure based on f_h instead of f. For an open set $A \subseteq \Gamma$ set

$$\mu_h(A) = c_{h0}^{-1} \int_A f_h \cdot \nu_\Gamma \, dx, \tag{3.2}$$

where

$$c_{h0} = \int_\Gamma f_h \cdot \nu_\Gamma \, dx.$$

Set $T_k(x) = u_k(t,x) - r$ as in the previous section and let $u_k(t,x_j)$ for $t \in [0,T]$ and $x_j \in B \subseteq \Gamma$ be pseudo orbits. Approximate $\mathcal{M}$ by a partition $\mathcal{U}_h$ of Γ as defined in Section 1.4.

In order to approximately compute (1.6) we let $\{x_j\}_{j=1}^M$ be independent random variables $f_h(x) \cdot \nu_\Gamma \, dx$ distributed on B. Now $\{T_k^n x_j\}_{j=1}^M$ will be independent random variables $f_h(T_k^n x) \cdot \nu_\Gamma \, dx$ distributed on $T_k^n B$ and we compute $\mu_h(A \cap T_k^n B)$ by the Monte Carlo method. Set $f(x) = \chi_A(x)$ and $g(x) = \mu_h(T_k^n B)^{-1}$ in (1.8) and we obtain

$$\mu_h(A \cap T_k^n B) = \int_{T_k^n B} \chi_A(x) f_h \cdot \nu_\Gamma \, dx \approx \mu_h(T_k^n B) M^{-1} \sum_{j=1}^M \chi_A(T_k^n x_j).$$

Since T_k^n in general is not μ_h measure preserving we cannot set $\mu_h(T_k^n B) = \mu_h(B)$. Instead we evaluate $\mu_h(T_k^n B)$ by Monte Carlo integration and by a change of variables we obtain

$$\int_{T_k^n B} f_h \cdot \nu_\Gamma \, dx = \int_B f_h \cdot \nu_\Gamma |\det(\nabla T_k^n x)| \, dx \approx \mu_h(B) M^{-1} \sum_{j=1}^M |\det(\nabla T_k^n x_j)|,$$

where we note that $|\det(\nabla T_k^n x)| = 1$ if f_h is incompressible [6, p. 10].

Hence we define the following approximation to the correlation sequence (1.7) for $A, B \in \mathcal{U}_h$

(3.3)

$$C_{k,n}^M(A,B) = \mu_h(B) M^{-2} \sum_{i=1}^M |\det(\nabla T_k^n x_i)| \sum_{j=1}^M \chi_A(T_k^n x_j) - \mu_h(A)\mu_h(B).$$

3.2. **Error analysis.** As outlined in Section 1.5 we assume that to every pseudo orbit $u_k(t,x_j)$ there is an exact orbit $u(t,y_j)$ such that (1.11) is satisfied and that ε in (1.12) is small. Then for $A, B \in \mathcal{U}_h$ and $\widetilde{B}$ as defined in (1.13) we argue in the same way as we did in Section 2.2 and obtain

$$|\mu(A \cap T_k^n B) - \mu(A)\mu(B) - C_n(A, \widetilde{B})| \leq \varepsilon\big(\mu(\partial A) + \mu(A)\mu(\partial B)\big).$$

In order to make the connection to the approximate correlation function (3.3) we first note that for any $A \in \mathcal{U}_h$

$$|\mu_h(A) - \mu(A)| = \int_A (f_h - f) \cdot \nu_\Gamma \, dx \leq |A| e_f,$$

and thus we arrive at the following estimate. For $A, B \in \mathcal{U}_h$ and $\widetilde{B}$ as defined in (1.13)

$$|C^M_{k,n}(A,B) - C_n(A,B)| \le \varepsilon\big(\mu(\partial A) + \mu(A)\mu(\partial B)\big) + Ce_f + R, \tag{3.4}$$

where the constant $C = C(A,B)$ and where R must be interpreted statistically as explained in Section 1.4.1.

4. Numerical experiments

We only consider two examples of space periodic flows.

Inspired by [23] where laminar fluid mixing was experimentally studied in small channels we set up the following model. Let $\Omega \subset \mathbf{R}^3$, be a polyhedral domain with periodic boundaries Γ_A and Γ_B, see Figures 4.1 and 4.2, and consider the Dirichlet Stokes problem with periodic boundary conditions in dimensionless form

$$\begin{aligned} -\Delta U + \nabla P &= 0 \quad \text{in } \Omega, \\ \nabla \cdot U &= 0 \quad \text{in } \Omega, \\ U &= 0 \quad \text{on } \partial\Omega \setminus (\Gamma_A \cup \Gamma_B), \\ U|_{\Gamma_A} &= U|_{\Gamma_B}, \\ P|_{\Gamma_A} &= P|_{\Gamma_B} + R_P, \end{aligned} \tag{4.1}$$

where $U = (U_1, U_2, U_3)$ is the unknown velocity field, P is the unknown pressure and R_P is a constant modelling the pressure drop.

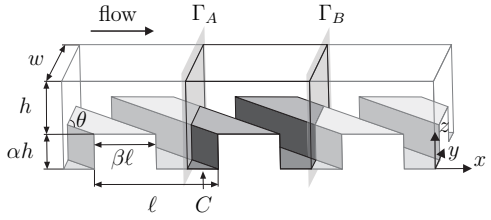


Figure 4.1: Three juxtaposed Ridge Domains. The shaded planes Γ_A and Γ_B are periodic boundaries. We choose the following values for the parameters: $\ell = w = 1$, $h = 0.3$, $\theta = 45°$, $\alpha = 2/3$, $\beta = 0.5$, and the length of the domain is $= 1$.

From [5] and [18] we know that $U \in W^{2,4/3}(\Omega)^3 \cap W^{1,3}_0$ and thus U is continuous although not Lipschitz continuous. There will be singularities

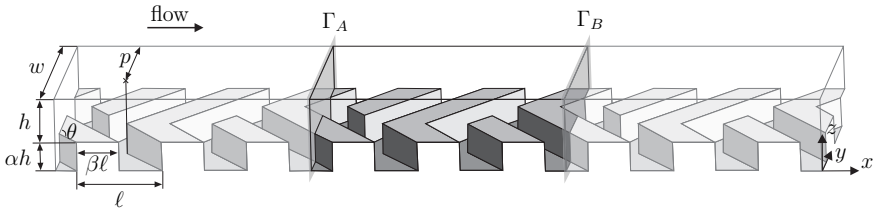


Figure 4.2: Three juxtaposed Herringbone Domains. The shaded planes Γ_A and Γ_B are periodic boundaries. We choose the following values for the parameters: $\ell = 2/3$, $w = 1$, $h = 1/5$, $\theta = 45°$, $\alpha = 2/3$, $\beta = 9/16$, $p = 2/3$, and the length of the domain is $= 14/9$.

in ∇U and P along the edges and vertices of Ω. However, if we let $\Omega' \subset \Omega$ such that $\text{dist}(\Omega', \partial\Omega)$ is not too small, then we may argue that U is Lipschitz continuous in Ω' by an interior estimate as in for example [13, Theorem 4.2, p. 209]. Thus when we compute orbits using $f = U$ (or in practice $f = U_h$) in (1.1) we only consider orbits that are not too close to $\partial\Omega$.

We refer to the domains in Figures 4.1 and 4.2 as the Ridge Domain and the Herringbone Domain respectively, the names are quoted from [23]. Accurate solutions U_h to (4.1) in the two domains are computed by a finite element method, Hood-Taylor P_2P_1 on fine triangulations. We illustrate the solutions in Figure 4.3 and 4.4.

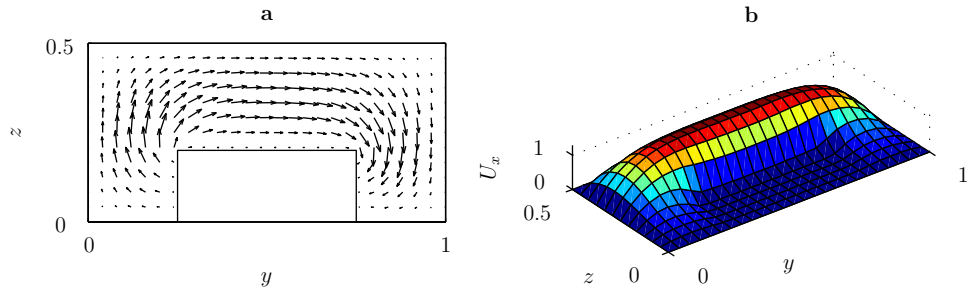


Figure 4.3: Velocity field for (4.1) solved in the Ridge Domain, Figure 4.1, at $x = 0.0$. **(a)** The y and z components of the velocity field. **(b)** The x component of the velocity field.

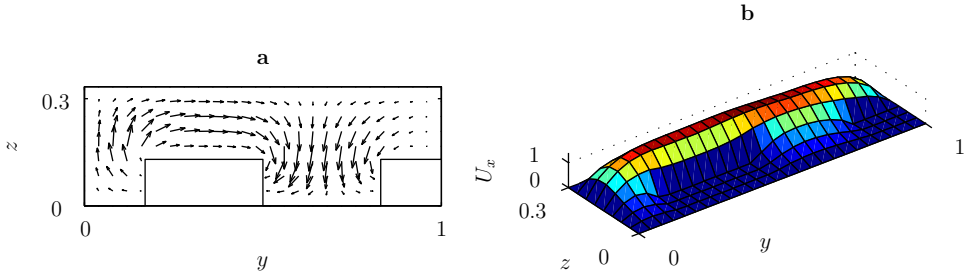


Figure 4.4: Velocity field for (4.1) solved in the Herringbone Domain, Figure 4.2, at $x = 0.0$. **(a)** The y and z components of the velocity field. **(b)** The x component of the velocity field.

4.1. **Decay of correlations and Poincaré sections.** We identify Γ_A and Γ_B in (4.1) with Γ as defined in Section 1.1. Since Γ is a polygon we partition it into a regular triangulation $\mathcal{T}_0$.

In order to examine the approximate decay of correlations (3.3) we choose one $B \in \mathcal{T}_S$ and three $A = A_i \in \mathcal{T}_i$, $i = 0, 1, 2$ such that $A_2 \subset A_1 \subset A_0$ and where $\mathcal{T}_{1,2}$ are defined by uniformly refining $\mathcal{T}_0$ two and four times, thus $\operatorname{diam}(A_0) > \operatorname{diam}(A_1) > \operatorname{diam}(A_2)$. In this case $\mathcal{T}_S$ is part of the triangulation use to solve (4.1) and hence is not commensurate with $\mathcal{T}_0$, this is not important for the conclusions. We illustrate $\mathcal{T}_0$ and A_i and B in Figures 4.5 and 4.6 and the triangles A_i and B are explicitly specified in Tables 4.1 and 4.2.

We may think of this numerical experiment as modelling a mixing process where one fluid flowing through B is supposed to mix with another fluid flowing through $\Gamma \setminus B$. The approximate decay of correlation will reflect the amount of mixing in A_i as a function of the number of iterates n, and $\operatorname{diam}(A_i)$ will reflect the length scale on which we resolve the mixing process, *cf.*, the discussion in Section 1.4.

As outlined in Section 3.1 we let $x_j \in B$ for $j = 1, \ldots, M$ be $U_h(x)\,dx$ distributed random variables for a relatively large number $M = 80964$ for the Ridge Domain and $M = 73445$ for the Herringbone Domain. We compute orbits to (1.1) for these initial points using the simple cG(1) method described in Section 1.5, with $f = U_h$ where U_h now is the computed solution to (4.1) . The time steps k_i for $i = 1, 2, \ldots, N$ are chosen adaptively

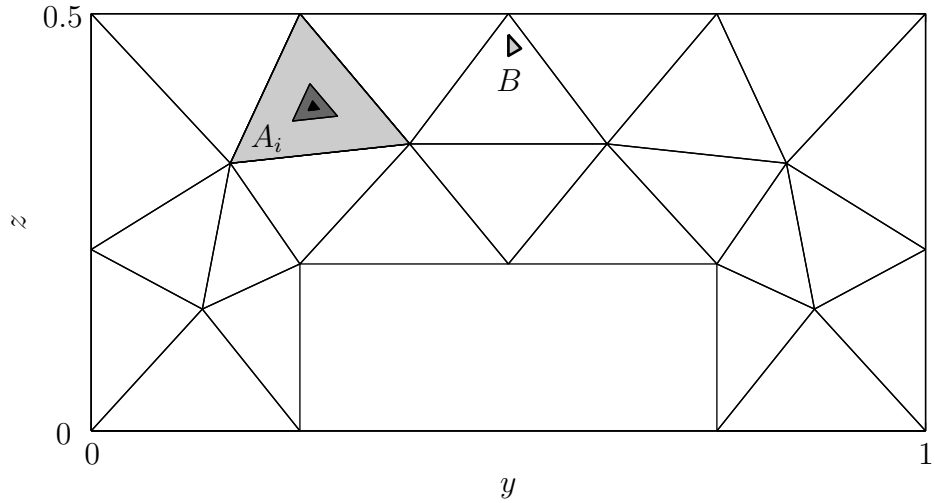


Figure 4.5: Partition of Γ for the Ridge Domain in terms of a triangulation $\mathcal{T}_0$. Shaded triangles illustrate B and A_i, $i = 0, 1, 2$, where $A_{1,2}$ are defined by refining A_0 two and four times, respectively, picking the central triangle.

so that the local residual is less than a relatively small tolerance, for more details see [10].

Table 4.1: A_i and B for the Ridge Domain where a_0, a_1 and a_2 denote the (x, y)-coordinates of vertices.

	a_0	a_1	a_2
B	(0.499997, 0.474672)	(0.499994, 0.449343)	(0.515248, 0.458255)
A_0	(0.381751, 0.344022)	(0.250000, 0.500000)	(0.166667, 0.320676)
A_1	(0.262104, 0.416175)	(0.241271, 0.371343)	(0.295042, 0.377180)
A_2	(0.259922, 0.384010)	(0.273365, 0.385470)	(0.265131, 0.395218)

We remark that a complete characterization of the decay of correlations (1.6) involves examining all combinations of $A, B \in \mathcal{T}_{i,s}$. Such general analysis would be computationally challenging since the amount of work

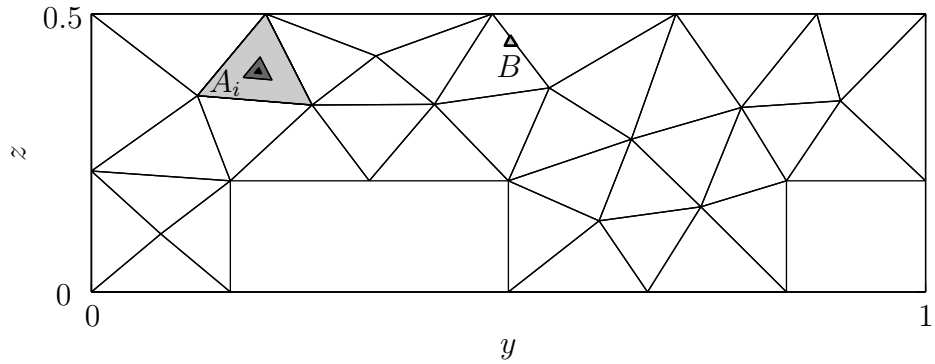


Figure 4.6: Partition of Γ for the Herringbone Domain in terms of a triangulation $\mathcal{T}_0$. Shaded triangles illustrate B and A_i, $i = 0, 1, 2$, where $A_{1,2}$ are defined by refining A_0 two and four times, respectively, picking the central triangle.

Table 4.2: A_i and B for the Herringbone Domain where a_0, a_1 and a_2 denote the (x, y)-coordinates of vertices.

	a_0	a_1	a_2
B	$(0.502007, 0.510363)$	$(0.495657, 0.295325)$	$(0.510363, 0.295325)$
A_0	$(0.209104, 0.333333)$	$(0.127166, 0.235331)$	$(0.264605, 0.224165)$
A_1	$(0.182010, 0.257040)$	$(0.216370, 0.254248)$	$(0.202495, 0.281541)$
A_2	$(0.204311, 0.261769)$	$(0.200842, 0.268592)$	$(0.195721, 0.262467)$

grows quadratically in the number of triangles in $\mathcal{T}_{i,s}$. A general characterization of this kind is beyond the scope of this work.

Instead of the complete characterization of the decay of correlation we plot the Poincaré sections for orbits starting in B see Figures 4.7 and 4.8. This will give qualitative information of the mixing in the entire domain Γ and from such plots we may readily identify regions with either poor or good mixing.

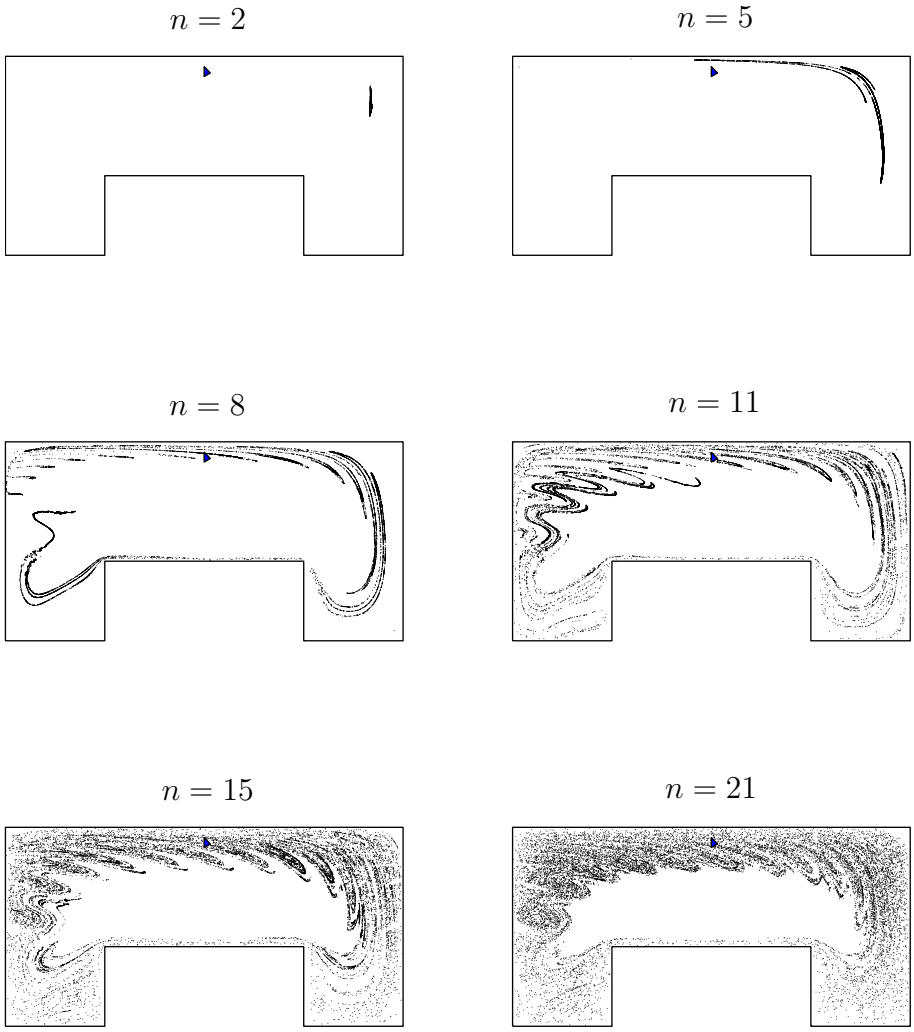


Figure 4.7: Poincaré sections for the flow in the Ridge Domain. 80964 orbits starting in B are included in the data.

Finally, we plot the correlation sequence in Figure 4.9 and normalized the data in the following way,

$$\widehat{C}_{k,n}^{N} = \left| \frac{C_{k,n}^{N}}{\mu(A)\mu(B)} \right|. \tag{4.2}$$

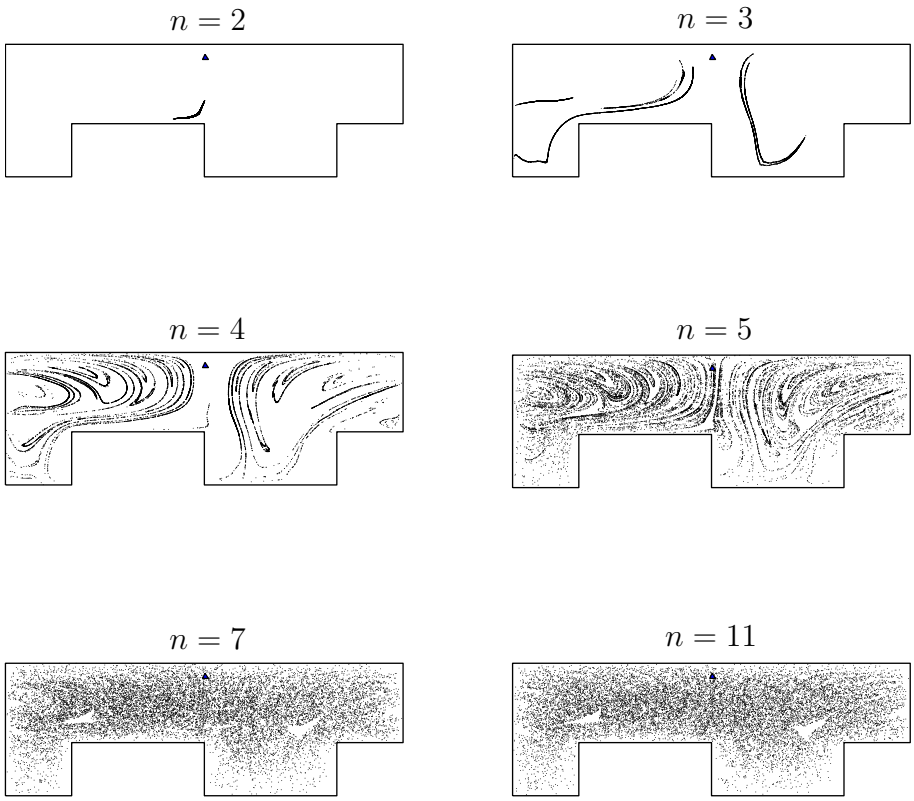


Figure 4.8: Poincaré sections for the flow in the Herringbone Domain. 73445 orbits starting in B are included in the data.

4.2. **Discussion.** We stress that the treatment of the examples in this section are not meant to be exhaustive in characterizing the mixing properties of Ridge Domain and Herringbone Domain. We rather meant to indicate how the proposed mixing measure works in practice. The overall impression from the Poincaré mapping, Figures 4.7 and 4.8, are in qualitative agreement with the experiments in [23]. However the correlation sequences in Figure 4.9 are not obviously interpreted, the simulations must for example be run over larger time intervals in order to see whether the

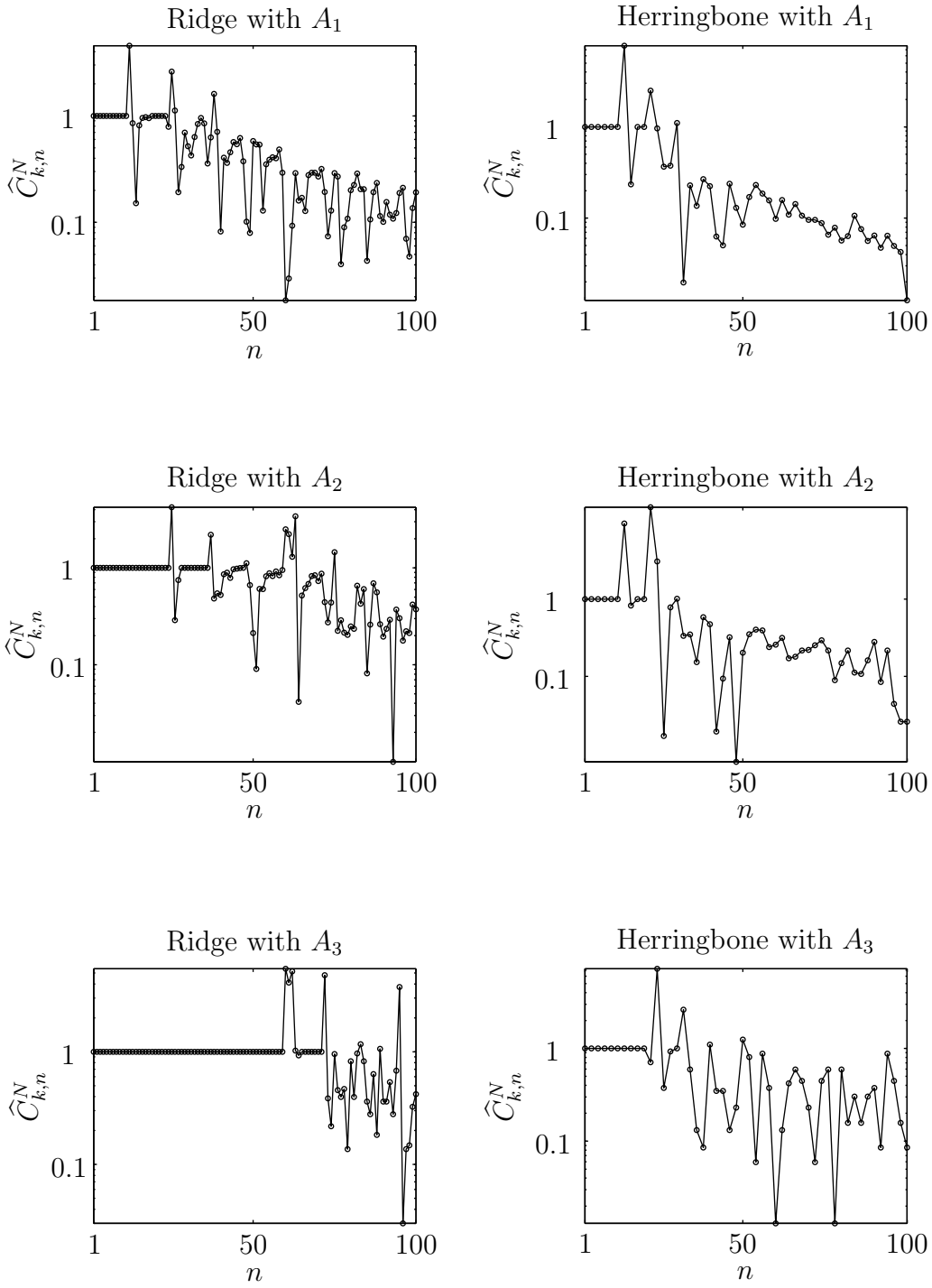


Figure 4.9: The correlation sequence for the Ridge Domain and Herringbone Domain and three differen target sets A_i.

decay rate is exponential. Although when we increase the resolution, decreasing the size of A_i, we can see a clear difference between the Ridge Domain and the Herringbone Domain as to when the decay of correlations start.

5. Conclusions

We have outlined a methodology for computationally characterizing fluid mixing in incompressible flows. This methodology could in principle be used for rigorous computational characterization of fluid mixing in the sense that error in the mixing measure is controlled and made small. We have not attempt to achieve this during the course of this work.

However, we remark that in order to obtain such results we will have to control all kinds of errors: the error in the computed velocity field $e_f = f_h - f$, the error in the computed orbits $u_k(t, x)$, and the error in computed mixing measure. Of these the most difficult to control is the error in the computed velocity field e_f. We note that this type of error control is a vital research field and is rather involved to implement.

References

[1] H. Aref, *Stochastic particle motion in laminar flows*, Phys. Fluids A **3** (1991), 1009–1016.

[2] ———, *The development of chaotic advection*, Phys. Fluids **14** (2002), 1315–1325.

[3] J. Aubin, D.F. Fletcher, and C. Xuereb, *Design of micromixers using cfd modelling*, Chem. Eng. Sci. **60** (2005), 2503–2516.

[4] C. Bonatti, L. J. Díaz, and M. Viana, *Dynamics Beyond Uniform Hyperbolicity*, Springer-Verlag, 2005.

[5] R. M. Brown and Z. Shen, *Estimates for the Stokes operator in Lipschitz domains*, Indiana Univ. Math. J. **44** (1995), 1183–1206.

[6] A. J. Chorin and J. E. Marsden, *A Mathematical Introduction to Fluid Mechanics*, second ed., Springer-Verlag, 1990.

[7] S-N. Chow and E. S. Van Vleck, *A shadowing lemma approach to global error analysis for initial value ODEs*, SIAM J. Sci. Comput. **15** (1994), 959–976.

[8] B. A. Coomes, H. Koçak, and K. J. Palmer, *Rigorous computational shadowing of orbits of ordinary differential equations*, Numer. Math. **69** (1995), 401–421.

[9] I. P. Cornfeld, S. V. Fomin, and Ya. G. Sinaĭ, *Ergodic Theory*, Springer-Verlag, 1982.

[10] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Computational Differential Equations*, Cambridge University Press, 1996.

[11] A. Ern and J-L. Guermond, *Theory and Practice of Finite Elements*, Springer-Verlag, 2004.

[12] M.D. Finn, S.M. Cox, and H.M. Byrne, *Mixing measures for a two-dimensional chaotic stokes flow*, J. Eng. Math. **48** (2004), 129–155.

[13] G. P. Galdi, *An Introduction to the Mathematical Theory of the Navier-Stokes Equations. Vol. I*, Springer-Verlag, 1994.

[14] V. Hessel, H. Lowe, and F. Schonfeld, *Micromixers-a review on passive and active mixing principles*, Chem. Eng. Sci. **60** (2005), 2479–2501.

[15] G. E. Karniadakis and A. Beskok, *Micro Flows*, Springer-Verlag, 2002.

[16] A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*, Cambridge University Press, 1995.

[17] B. Lapeyre, É. Pardoux, and R. Sentis, *Méthodes de Monte-Carlo pour les équations de Transport et de Diffusion*, Springer-Verlag, 1998.

[18] V. G. Maz'ya and J. Rossmann, *Lp estimates of solutions to mixed boundary value problems for the Stokes system in polyhedral domains*, ArXiv Mathematical Physics e-prints (2004).

[19] Nam-Trung N. and Zhigang W., *Micromixers-a review*, J. Micromech. Microeng. **15** (2005), 1–16.

[20] J. M. Ottino, *The Kinematics of Mixing: stretching, chaos, and transport*, Cambridge University Press, 1989.

[21] J.M. Ottino, *The mixing of fluids*, Sci. Am. **260** (1989), 40–49.

[22] K Palmer, *Shadowing in Dynamical Systems*, Kluwer Academic Publishers, 2000.

[23] A.D. Stroock, S.K.W. Dertinger, A. Ajdari, I. Mezic, H.A. Stone, and G.M. Whitesides, *Chaotic mixer for microchannels*, Science **295** (2002), 647–651.
[24] E. D. Svensson, *Computational Characterization of Mixing in Flows*, PhD thesis, Chalmers University of Technology, 2006, Paper III.
[25] L-S. Young, *Developments in chaotic dynamics*, Notices Amer. Math. Soc. **45** (1998), 1318–1328.
[26] ———, *Geometric and ergodic theory of hyperbolic dynamical systems*, Current Developments in Mathematics, Int. Press, Somerville, MA, 1999, pp. 237–278.

Department of Mathematical Sciences, Chalmers University of Technology, SE-412 96 Göteborg, Sweden
E-mail address: erik.svensson@math.chalmers.se

Paper II

POINTWISE A POSTERIORI ERROR ESTIMATES FOR THE STOKES EQUATIONS IN POLYHEDRAL DOMAINS

ERIK D. SVENSSON AND STIG LARSSON

Abstract. We derive pointwise a posteriori residual-based error estimates for finite element solutions to the Stokes equations in polyhedral domains. The estimates relies on the regularity of the of Stokes equations and provide an upper bound for the pointwise error in the velocity field on polyhedral domains. Whereas the estimates provide upper bounds for the pointwise error in the gradient of the velocity field and the pressure only for a restricted class of polyhedral domains, convex polyhedral domains in $\mathbf{R}^2$, and polyhedral domains with angles at edges $< 3\pi/4$ in $\mathbf{R}^3$. In the cause of this study we also derive L^q a posteriori error estimates, generalizing well known L^2 estimates.

1. Introduction

Let $\Omega \subset \mathbf{R}^n$, $n = 2, 3$, be a polyhedral domain and consider the Dirichlet Stokes problem in dimensionless form

$$\begin{aligned} -\Delta u + \nabla p &= f \quad \text{in } \Omega, \\ \nabla \cdot u &= g \quad \text{in } \Omega, \\ u &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{1.1}$$

where $u = (u_1, \dots, u_n)$ is the unknown velocity field, p the unknown pressure, $f = (f_1, \dots, f_n)$ is an external body force and g is a function prescribing the compressibility of the flow, for incompressible flows $g = 0$.

The purpose of this paper is to establish residual-based pointwise a posteriori error estimates for conforming finite element approximations

Date: April 18, 2006.

2000 *Mathematics Subject Classification.* 65N15, 65N30, 76D07.

Key words and phrases. a posteriori, pointwise error estimates, maximum norm, Stokes equations.

(u_h, p_h) to the Stokes problem (1.1). Only requiring that the finite element mesh is regular, allowing adaptively refined meshes, we obtain a number of error estimates.

(1) For polyhedral domains we derive pointwise error estimates for the velocity field
$$\|u_h - u\|_{L^\infty(\Omega)} \leq \mathcal{E}_1(u_h, p_h, f, g, \Omega, \mathcal{T}).$$
(2) For convex polyhedral domains in $\mathbf{R}^2$, and for polyhedral domains in $\mathbf{R}^3$ with angles at edges $< 3\pi/4$ we derive pointwise error estimates for the gradient of the velocity field
$$\|\nabla(u_h - u)\|_{L^\infty(\Omega)} \leq \mathcal{E}_2(u_h, p_h, f, g, \Omega, \mathcal{T})$$
(3) For polyhedral domain as specified in Item 2 above we derive pointwise error estimates for the pressure
$$\|p_h - p\|_{L^\infty(\Omega)} \leq \mathcal{E}_3(u_h, p_h, f, g, \Omega, \mathcal{T}).$$
(4) For polyhedral domains and for $q \in [2n/(n+1), 2n/(n-1)]$ we also derive the following L^q-estimate
$$\|\nabla(u_h - u)\|_{L^q(\Omega)} + \|p_h - p\|_{L^q(\Omega)} \leq \mathcal{E}_4(u_h, p_h, f, g, \Omega, \mathcal{T}).$$

The right hand sides $\mathcal{E}_{1,2,3,4}$ in the estimates above are functions derived from the residuals, depending on the finite element solution, the data, the domain and the triangulation.

The first estimate in Item 1 relies on the fact that, for sufficiently regular data, the velocity field is Hölder continuous in polyhedral domains. Similarly, the pointwise estimates for the gradient of the velocity field, Item 2, and the pressure, Item 3, require continuity. This is generally not obtained in polyhedral domains without imposing extra constraints, convexity for polyhedral domains in $\mathbf{R}^2$ and a minimum inner angle condition, $< 3\pi/4$ at edges, for polyhedral domains in $\mathbf{R}^3$ [13]. We note that estimating the gradient of the velocity field is somewhat more involved since ∇u_h is discontinuous at the $(n-1)$-faces of the triangulation.

The fourth estimate in Item 4 relies on L^q-regularity estimates stated in [3] for Lipschitz domains and also in [13] for polyhedral domains. It is a straightforward generalization of the L^2-based estimates in [19].

The techniques used to prove the pointwise error estimate is inspired by [14], where an a posteriori residual-based pointwise error estimate was derived for Poisson's equation in two dimensions, later this analysis was also done in three dimensions [4]. We remark that the gradient of the

solution was not considered in these studies. The pointwise a priori error analysis for the Stokes problem was worked out in two dimensions for convex domains and quasiuniform triangulations [5], and in three dimensions for polyhedral domains with the similar type of constraints as mentioned above and for quasiuniform triangulations [10].

1.1. **Assumptions and notation.** We only consider functions defined on bounded domains $\omega \subseteq \Omega \subset \mathbf{R}^n$, $n = 2, 3$, with measure denoted by $|\omega|$, and where Ω is associated with the Stokes problem (1.1) and the dual problem (1.4).

Let $\{e_i\}_{i=1}^n$ denote the canonical unit vectors, $e_1 = (1, 0)$ and $e_2 = (0, 1)$ for $n = 2$ and $e_1 = (1, 0, 0)$, $e_2 = (0, 1, 0)$ and $e_3 = (0, 0, 1)$ for $n = 3$.

We denote the i:th partial derivative by

$$D_i := \frac{\partial}{\partial x_i}, \quad i = 1, \ldots, n,$$

and the gradient by

$$\nabla := (D_1, \ldots, D_n),$$

and the matrix of second order derivatives

$$\nabla^2 := (D_i D_j)_{i,j=1}^n.$$

We use standard notation for spaces of smooth functions, for example, $C^m(\omega)$, $C_0^\infty(\omega)$ and $C^{m,\gamma}(\overline{\omega})$, and for Lebesgue and Sobolev spaces, $L^q(\omega) = W^{0,q}(\omega)$, $W^{k,q}(\omega)$ and $W_0^{k,q}(\omega)$, see for example [1]. For $u \in L^q(\omega)$ or $u \in W^{k,q}(\omega)$ we use the following notation for the norm

$$\|u\|_{L^q(\omega)} = \|u\|_{q,\omega} \qquad \text{and} \qquad \|u\|_{W^{q,k}(\omega)} = \|u\|_{q,k,\omega},$$

and likewise for the corresponding seminorms $|u|_{q,k,\omega}$.

When $q = 2$ $L^q(\omega) = L^2(\omega)$ becomes a Hilbert space and we denote the scalar product by

$$(u, v)_\omega := \int_\omega uv \, dx.$$

For $u \in W_0^{1,q}(\omega)$ or for $u \in W^{1,q}(\omega)$ with $\int_{\omega_0} u \, dx = 0$ for some non empty $\omega_0 \subset \omega$, the norm is equivalent to the seminorm, $\|u\|_{1,q,\omega} \approx |u|_{1,q,\omega}$, see for example [18, Lemma 1.1.1–2, pp. 43–44]. We will use this equivalence without further notice throughout this work.

We denote the dual exponent to q by $q' = q/(q-1)$ and the dual space to $W_0^{k,q}(\omega)$ by $W^{-k,q'}(\omega)$ with the dual norm

$$\|u\|_{-k,q',\omega} := \sup_{\varphi \in C_0^\infty(\omega)} \frac{|\langle u, \varphi \rangle|}{\|\varphi\|_{k,q,\omega}}, \tag{1.2}$$

where $\langle \cdot, \cdot \rangle$ denotes the duality pairing.

Generally, for a vector space V we denote its dual space by V' with dual norm

$$\|u\|_{V'} := \sup_{\varphi \in V} \frac{|\langle u, \varphi \rangle|}{\|\varphi\|_V},$$

for example, $W_0^{k,q}(\omega)' := W^{-k,q'}(\omega)$.

When $\omega = \Omega$ we sometimes write L^q instead of $L^q(\Omega)$ and $\|\cdot\|_q$ instead of $\|\cdot\|_{q,\Omega}$ and likewise for Sobolev spaces and their norms and the L^2 scalar product.

We use the quotient space $W^{k,q}/\mathbf{R}$ with the norm

$$\|v\|_{W^{k,q}/\mathbf{R}} := \inf_{c \in \mathbf{R}} \|v + c\|_{k,q}.$$

For vector fields

$$\Omega \ni x \mapsto u(x) = (u_1(x), \dots, u_n(x)) \in \mathbf{R}^n$$

we set

$$\nabla u := (D_i u_j)_{i,j=1}^n,$$
$$\nabla^2 u := (D_i D_j u_k)_{i,j,k=1}^n,$$

and for $u = (u_1, \dots, u_n) \in W^{k,q}(\Omega)^n$ we use the Sobolev (Lebesgue) norm

$$\|u\|_{k,q} := \left(\sum_{i=1}^n \|u_i\|_{k,q}^q \right)^{1/q},$$

and the corresponding seminorms, the maximum norms

$$\|u\|_\infty := \max_i \|u_i\|_\infty,$$
$$\|\nabla u\|_\infty := \max_{i,j} \|D_i u_j\|_\infty,$$

and the scalar product

$$(u, v) = \sum_{i=1}^n (u_i, v_i).$$

We also use the product spaces $\mathcal{W}^{1,q} := W_0^{1,q}(\Omega)^n \times L^q(\Omega)/\mathbf{R}$ with the norm

$$\|(u,p)\|_{\mathcal{W}^{1,q}} := \|u\|_{1,q} + \|p\|_{L^q/\mathbf{R}},$$

and $\mathcal{W}^{2,q} := (W^{2,q}(\Omega)^n \times W^{1,q}(\Omega)) \cap \mathcal{W}^{1,s}$ where $s = nq/(n-q)$, see Theorem 1.3, with the norm

$$\|(u,p)\|_{\mathcal{W}^{2,q}} := \|u\|_{2,q} + \|p\|_{W^{1,q}/\mathbf{R}}.$$

Finally, throughout this work we use C or C_i, $i = 1, 2, \ldots$, to denote various constants, not necessarily with the same value from time to time.

1.2. **Weak formulation.** We follow the standard notation, *cf.* [11, 19], and define the bilinear form

$$\mathcal{L}((u,p),(\phi,\lambda)) := a(u,\phi) + b(\phi,p) - b(u,\lambda),$$

for test functions (ϕ,λ) and where

$$a(u,\phi) := \int_\Omega \sum_{i,j=1}^n \frac{\partial u_i}{\partial x_j}\frac{\partial \phi_i}{\partial x_j}\,dx \qquad \text{and} \qquad b(\phi,p) := -\int_\Omega (\nabla\cdot\phi)p\,dx.$$

For data $f \in W^{-1,q}$ and $g \in L^q$ such that $\int_\Omega g\,dx = 0$ and for $2n/(n+1) < q < 2n/(n-1)$ there is a unique weak solution to (1.1), see Theorem 1.1 for a more precise statement. The weak formulation of (1.1) now reads. Find $(u,p) \in \mathcal{W}^{1,q}(\Omega)$ such that

$$\mathcal{L}((u,p),(\phi,\lambda)) = \langle f,\phi\rangle + (g,\lambda) \quad \forall (\phi,\lambda) \in \mathcal{W}^{1,q'}(\Omega), \tag{1.3}$$

where $\langle\cdot,\cdot\rangle$ denotes the appropriate duality pairing.

The dual problem to (1.1) is

$$\begin{aligned} -\Delta\tilde{u} - \nabla\tilde{p} &= \tilde{f} \quad \text{in } \Omega,\\ -\nabla\cdot\tilde{u} &= \tilde{g} \quad \text{in } \Omega,\\ \tilde{u} &= 0 \quad \text{on } \partial\Omega, \end{aligned} \tag{1.4}$$

where $\tilde{f} \in W^{-1,q'}$ and $\tilde{g} \in L^{q'}$ such that $\int_\Omega \tilde{g}\,dx = 0$ and for $2n/(n+1) < q' < 2n/(n-1)$. The corresponding weak formulation is. Find $(\tilde{u},\tilde{p}) \in \mathcal{W}^{1,q'}(\Omega)$ such that

$$\mathcal{L}((\phi,\lambda),(\tilde{u},\tilde{p})) = \langle\phi,\tilde{f}\rangle + (\lambda,\tilde{g}) \quad \forall (\phi,\lambda) \in \mathcal{W}^{1,q}(\Omega). \tag{1.5}$$

1.3. **Existence and regularity in non-smooth domains.** For any domain $\Omega \subset \mathbf{R}^n$, $n = 2, 3$, and data $f \in W^{-1,2}(\Omega)^n$ and $g \in L^2(\Omega)$ such that $\int_\Omega g \, dx = 0$, it is well known that there exists a unique weak solution $(u, p) \in W_0^{1,2}(\Omega)^n \times L^2(\Omega)/\mathbf{R}$ to (1.1), see for example [18, Chaper 3] and references therein. For sufficiently regular domains and data there are several extensions such that $(u, p) \in W_0^{1,q}(\Omega)^n \times L^q(\Omega)/\mathbf{R}$, see Remark 1.1 below. In Theorem 1.1 we quote one example of such an extension where the Stokes problem is formulated on Lipschitz domains. This is a slight modification of [3, Theorem 2.9] where it was provided with $g = 0$. However the case $g \neq 0$ is readily included.

Theorem 1.1. *Let $\Omega \subset \mathbf{R}^n$, $n = 2, 3$, be a bounded Lipschitz domain. There exist $\varepsilon > 0$ such that if $(3+\varepsilon)/(2+\varepsilon) < q < 3+\varepsilon$ and $f \in W^{-1,q}(\Omega)^n$ and $g \in L^q(\Omega)$ with $\int_\Omega g \, dx = 0$, then there exist a unique weak solution $(u, p) \in W_0^{1,q}(\Omega)^n \times L^q(\Omega)/\mathbf{R}$ to (1.1). Moreover, the solution satisfies the inequality*

$$\|u\|_{1,q} + \|p\|_{L^q/\mathbf{R}} \leq C\big(\|f\|_{-1,q} + \|g\|_q\big), \tag{1.6}$$

for some $C = C(n, q, \Omega)$.

Proof. For $g = 0$ this is [3, Theorem 2.9]. For $g \neq 0$ we use the method of *subtracting the divergence*, see for example [18, Theorem 1.4.1, p. 114], to handle the non-homogenous compressibility constraint.

For Ω and g as stated there exists $v \in W_0^{1,q}(\Omega)^n$ such that

$$\nabla \cdot v = g \quad \text{and} \quad \|v\|_{1,q} \leq C\|g\|_q, \tag{1.7}$$

see, for example, [18, Lemma 2.1.1, p. 68]. Taking $w = u - v$ we see that (1.1) is equivalent to

$$-\Delta w + \nabla p = f + \Delta v, \qquad \nabla \cdot w = 0, \qquad \text{in } \Omega,$$

and $w|_{\partial\Omega} = 0$. Now [3, Theorem 2.9] implies that there exist a unique pair $(w, p) \in W_0^{1,q}(\Omega)^n \times L^q(\Omega)/\mathbf{R}$ satisfying the above equations and the inequality

$$\|w\|_{1,q} + \|p\|_{L^q/\mathbf{R}} \leq C\|f + \Delta v\|_{-1,q},$$

for some $C = C(n, q, \Omega)$.

Thus, $(u, p) \in W_0^{1,q}(\Omega)^n \times L^q(\Omega)/\mathbf{R}$ is a unique solution to (1.1) and the estimate above implies that

$$\|u\|_{1,q} + \|p\|_{L^q/\mathbf{R}} \leq C\big(\|f\|_{-1,q} + \|v\|_{1,q} + \|\Delta v\|_{-1,q}\big).$$

The inequality (1.6) now follows from the estimate in (1.7) and the fact that $\|\Delta v\|_{-1,q} \le \|v\|_{1,q}$. □

Remark 1.1. (1) For $n = 2$ the results of the theorem actually holds with $(4+\varepsilon)/(3+\varepsilon) < q < 4+\varepsilon$. This is provided in the same way as for $n = 3$ [17]. (2) For polyhedral domains a similar theorem was established in [13], in particular, for convex polyhedral domains the result holds with $1 < q < \infty$. (3) For C^1-domains there is a similar theorem again with $1 < q < \infty$, see for example [8].

As a consequence of Theorem 1.1 and Remark 1.1 we obtain the following inf-sup like estimate.

Corollary 1.2. *For q and Ω as in Theorem 1.1 we have*

$$\|(u,p)\|_{\mathcal{W}^{1,q}} \le C \sup_{(\phi,\lambda)\in\mathcal{W}^{1,q'}} \frac{|\mathcal{L}((u,p),(\phi,\lambda))|}{\|(\phi,\lambda)\|_{\mathcal{W}^{1,q'}}} \qquad \forall (u,p)\in \mathcal{W}^{1,q}(\Omega), \tag{1.8}$$

where $C = C(n, q', \Omega)$.

Proof. Let (ϕ_i, λ_i) be the solutions to the following problems

$$\begin{aligned} -\Delta\phi_1 - \nabla\lambda_1 = \tilde{f}, \quad \nabla\cdot\phi_1 = 0, \quad \text{in } \Omega; \qquad \phi_1|_{\partial\Omega} = 0,\\ -\Delta\phi_2 - \nabla\lambda_2 = 0, \quad \nabla\cdot\phi_2 = \tilde{g} - \tilde{g}_0, \quad \text{in } \Omega; \qquad \phi_2|_{\partial\Omega} = 0, \end{aligned}$$

where $\tilde{f} \in W^{-1,q'}(\Omega)^n$ and $\tilde{g} \in L^{q'}(\Omega)$ with the mean $\tilde{g}_0 = |\Omega|^{-1}\int_\Omega \tilde{g}\,dx$.

With Theorem 1.1 applied to the above problems and with (1.5) we get

$$\begin{aligned} &\sup_{(\phi,\lambda)\in\mathcal{W}^{1,q'}} \frac{|\mathcal{L}((u,p),(\phi,\lambda))|}{\|(\phi,\lambda)\|_{\mathcal{W}^{1,q'}}}\\ &\quad\ge \frac{1}{2}\left(\frac{|\mathcal{L}((u,p),(\phi_1,\lambda_1))|}{\|(\phi_1,\lambda_1)\|_{\mathcal{W}^{1,q'}}} + \frac{|\mathcal{L}((u,p),(\phi_2,\lambda_2))|}{\|(\phi_2,\lambda_2)\|_{\mathcal{W}^{1,q'}}}\right)\\ &\quad\ge C\left(\frac{|\langle u,\tilde{f}\rangle|}{\|\tilde{f}\|_{-1,q'}} + \frac{|(p,\tilde{g}-\tilde{g}_0)|}{\|\tilde{g}-\tilde{g}_0\|_{q'}}\right) \end{aligned} \tag{1.9}$$

Since $W^{1,q}$ and L^q are reflexive for $1 < q < \infty$ we get

$$\sup_{\tilde{f}\in W^{-1,q'}(\Omega)^n} \frac{|\langle u,\tilde{f}\rangle|}{\|\tilde{f}\|_{-1,q'}} = \|u\|_{1,q},$$

and since $(p, \tilde{g} - \tilde{g}_0) = (p - p_0, \tilde{g})$, where $p_0 = |\Omega|^{-1} \int_\Omega p \, dx$, we have

$$\sup_{\tilde{g} \in L^{q'}(\Omega)} \frac{|(p, \tilde{g} - \tilde{g}_0)|}{\|\tilde{g} - \tilde{g}_0\|_{q'}} \geq \frac{1}{2} \inf_{c \in \mathbf{R}} \sup_{\tilde{g} \in L^{q'}(\Omega)} \frac{|(p + c, \tilde{g})|}{\|\tilde{g}\|_{q'}} = \frac{1}{2} \|p\|_{L^q/\mathbf{R}},$$

where we also used the estimate $\|\tilde{g} - \tilde{g}_0\|_{q'} \leq 2\|\tilde{g}\|_{q'}$.

Now since (1.9) is valid for any $\tilde{f} \in W^{-1,q'}(\Omega)^n$ and for any $\tilde{g} \in L^{q'}(\Omega)$ we may take the supremum with respect to $\tilde{f}$ and $\tilde{g}$, which together with the last two estimates above completes the proof. □

The next theorem concerns the $W^{2,q}(\Omega)^n \times W^{1,q}(\Omega)$-regularity of the solution to (1.1) in polyhedral domains. The theorem is due to [13], for a review see [12], although it is formulated somewhat differently here.

Theorem 1.3. *Let $\Omega \subset \mathbf{R}^n$, $n = 2, 3$, be a polyhedral domain and let $1 < q \leq 4/3$. Suppose $f \in L^q(\Omega)^n$ and $g \in W^{1,q}(\Omega)$ such that $\int_\Omega g \, dx = 0$. Then there exist a unique weak solution $(u, p) \in W_0^{1,s}(\Omega)^n \times L^s(\Omega)/\mathbf{R}$ to (1.1) for $s = nq/(n-q)$ such that $(u, p) \in W^{2,q}(\Omega)^n \times W^{1,q}(\Omega)$. Moreover, the solution satisfies the inequality*

$$\|u\|_{2,q} + \|p\|_{W^{1,q}/\mathbf{R}} \leq C\big(\|f\|_q + |g|_{1,q}\big), \tag{1.10}$$

for some $C = C(n, q, \Omega)$.

Proof. By virtue of Theorem 1.1 and Remark 1.1 we obtain the existence, since by Sobolev's imbedding theorem we have $L^q \subset W^{-1,s}$ and $W^{1,q} \subset L^s$ for $s = nq/(n-q)$, $1 < q \leq 4/3$ and we readily check that $2 \leq s \leq 4$ for $n = 2$ and $(3 + \varepsilon)/(2 + \varepsilon) < s < 3 + \varepsilon$ for $n = 3$ and any $\varepsilon > 0$.

The regularity $(u, p) \in W^{2,q}(\Omega)^n \times W^{1,q}(\Omega)$ follows from [13, Theorem 5.3] which is also true provided $(u, p) \in W_0^{1,s}(\Omega)^n \times L^s(\Omega)/\mathbf{R}$ [15]. The estimate (1.10) is then as consequence of the open mapping theorem, see for example [6, Corollary 5.11, p. 162]. □

Remark 1.2. (1) For $n = 2$ and if the maximum inner angle in the polyhedral domain is less than $\pi - \delta$ for some $\delta > 0$, then the result can be extended to hold with $1 < q \leq 2 + \varepsilon$ for some $\varepsilon > 0$ [15] and *cf.* [13, §5.5]. (2) For $n = 3$ and if the maximum inner angle at the edges in the polyhedral domain is less than $3\pi/4 - \delta$ for some $\delta > 0$, then the result can be extended to hold with $1 < q \leq 3 + \varepsilon$ for some $\varepsilon > 0$ [15] and *cf.* [13, §5.5]. (3) For C^1-domains there is a similar theorem with $1 < q < \infty$, see for example [8]. In cases (1) and (2) the existence is also true since for convex domains Theorem 1.1 is modified as in Remark 1.1.

We now state a corollary where we assume that we have the higher regularity in Remark 1.2.

Corollary 1.4. *Suppose that the solution $(\tilde{u},\tilde{p})$ to (1.4) with data as in Theorem 1.3 belongs to $W^{2,q'}(\Omega)^n \times W^{1,q'}(\Omega)$ for some $q' > n$. Then the solution (u,p) to (1.1) satisfies*

$$\|u\|_q + \|p\|_{W^{1,q'}(\Omega)'/\mathbf{R}} \leq C\big(\|f\|_{-2,q} + \|g\|_{W^{1,q'}(\Omega)'}\big), \tag{1.11}$$

for some $C = C(n,q',\Omega)$ and where $1/q + 1/q' = 1$ and $W^{1,q'}(\Omega)'/\mathbf{R}$ is the dual space to $W^{1,q'}(\Omega)/\mathbf{R}$.

Proof. We use the same technique as in the proof of Corollary 1.2. With (1.3) we estimate

$$\begin{aligned}\|f\|_{-2,q} + \|g\|_{W^{1,q'}(\Omega)'} &= \sup_{\phi\in C_0^\infty(\Omega)^n} \frac{|\langle f,\phi\rangle|}{\|\phi\|_{2,q'}} + \sup_{\lambda\in W^{1,q'}/\mathbf{R}} \frac{|\langle g,\lambda\rangle|}{\|\lambda\|_{W^{1,q'}/\mathbf{R}}} \\ &\geq \sup_{(\phi,\lambda)\in\mathcal{W}^{2,q'}} \frac{|\mathcal{L}((u,p),(\phi,\lambda))|}{\|(\phi,\lambda)\|_{\mathcal{W}^{2,q'}}}.\end{aligned}$$

Let (ϕ_i,λ_i) be the solutions to the following problems

$$\begin{aligned} -\Delta\phi_1 - \nabla\lambda_1 = \tilde{f}, \quad \nabla\cdot\phi_1 = 0, \quad \text{in } \Omega; \qquad \phi_1|_{\partial\Omega} = 0, \\ -\Delta\phi_2 - \nabla\lambda_2 = 0, \quad \nabla\cdot\phi_2 = \tilde{g} - \tilde{g}_0, \quad \text{in } \Omega; \qquad \phi_2|_{\partial\Omega} = 0, \end{aligned}$$

where $\tilde{f} \in L^{q'}(\Omega)^n$ and $\tilde{g} \in W^{1,q'}(\Omega)$ with the mean $\tilde{g}_0 = |\Omega|^{-1}\int_\Omega \tilde{g}\ dx$. We assumed that $(\phi_i,\lambda_i) \in W^{2,q'}(\Omega)^n \times W^{1,q'}(\Omega)$ and thus we estimate

$$\begin{aligned}&\sup_{(\phi,\lambda)\in\mathcal{W}^{2,q'}} \frac{|\mathcal{L}((u,p),(\phi,\lambda))|}{\|(\phi,\lambda)\|_{\mathcal{W}^{2,q'}}} \\ &\geq \frac{1}{2}\left(\frac{|\mathcal{L}((u,p),(\phi_1,\lambda_1))|}{\|(\phi_1,\lambda_1)\|_{\mathcal{W}^{2,q'}}} + \frac{|\mathcal{L}((u,p),(\phi_2,\lambda_2))|}{\|(\phi_2,\lambda_2)\|_{\mathcal{W}^{2,q'}}}\right) \\ &\geq C\left(\frac{|(u,\tilde{f})|}{\|\tilde{f}\|_{q'}} + \frac{|\langle p,\tilde{g}-\tilde{g}_0\rangle|}{|\tilde{g}|_{1,q'}}\right),\end{aligned} \tag{1.12}$$

for some $C = C(n,q',\Omega)$.

Since L^q and $W^{1,q'}(\Omega)'$ are reflexive for $1 < q < \infty$ we get

$$\sup_{\tilde{f}\in L^{q'}(\Omega)^n} \frac{|\langle u,\tilde{f}\rangle|}{\|\tilde{f}\|_{q'}} = \|u\|_q,$$

and since $(p, \tilde{g} - \tilde{g}_0) = (p - p_0, \tilde{g})$, where $p_0 = |\Omega|^{-1} \int_\Omega p \, dx$, we have

$$\sup_{\tilde{g} \in W^{1,q'}(\Omega)} \frac{|\langle p, \tilde{g} - \tilde{g}_0 \rangle|}{|\tilde{g}|_{1,q'}} \geq \inf_{c \in \mathbf{R}} \sup_{\tilde{g} \in W^{1,q'}(\Omega)} \frac{|\langle p + c, \tilde{g} \rangle|}{\|\tilde{g}\|_{1,q'}} = \|p\|_{W^{1,q'}(\Omega)'/\mathbf{R}}.$$

Now since (1.12) is valid for any $\tilde{f} \in L^{q'}(\Omega)^n$ and any $\tilde{g} \in W^{1,q'}(\Omega)$ we may take the supremum with respect to $\tilde{f}$ and $\tilde{g}$, which together with the last two estimates above completes the proof. □

1.4. **Finite element formulation.** Let $\{\mathcal{T}\}_{h>0}$ denote a family of regular triangulations of Ω and let h_T denote the diameter of an n-simplex $T \in \mathcal{T}$ and set $h_{\min} = \min_{T \in \mathcal{T}_h} h_T$.

We only consider conforming finite element spaces, $X_h \subset W_0^{1,q}(\Omega)^n$ for the velocity and, $M_h/\mathbf{R} \subset L^q(\Omega)/\mathbf{R}$ for the pressure and define the product space $\mathcal{W}_h = X_h \times M_h/\mathbf{R}$. From (1.3) we obtain the finite element formulation. Find $(u_h, p_h) \in \mathcal{W}_h$ such that

$$\mathcal{L}((u_h, p_h), (\phi_h, \lambda_h)) = \langle f, \phi_h \rangle + (g, \lambda_h) \quad \forall (\phi_h, \lambda_h) \in \mathcal{W}_h. \tag{1.13}$$

As usual we also require that $\mathcal{W}_h$ satisfies the inf-sup condition [11], that is,

$$\|(u_h, p_h)\|_{\mathcal{W}^{1,2}} \leq C \sup_{(\phi_h, \lambda_h) \in \mathcal{W}_h} \frac{|\mathcal{L}((u_h, p_h), (\phi_h, \lambda_h))|}{\|(\phi_h, \lambda_h)\|_{\mathcal{W}^{1,2}}}, \tag{1.14}$$

for all $(u_h, p_h) \in \mathcal{W}_h$, which implies that (1.13) is well posed.

We particularly have in mind the family of Taylor-Hood finite elements, see fore example [11], which satisfy the above requirement.

We recall a few standard results from interpolation theory, see for example [16]. Let S_T denote the union of all simplices adjacent to T and let $\mathcal{I}_{X_h}$ and $\mathcal{I}_{M_h}$ denote interpolation operators $\mathcal{I}_{X_h} : W_0^{m,q}(\Omega)^n \to X_h$ and $\mathcal{I}_{M_h} : W^{m-1,q}(\Omega)/\mathbf{R} \to M_h/\mathbf{R}$. For integers $\ell = 0, 1$, $m = 1, \ldots$, and $(\phi, \lambda) \in W^{m,q}(S_T)^n \times W^{m-1,q}(S_T)/\mathbf{R}$, we have

$$\|\nabla^\ell (\phi - \mathcal{I}_{X_h} \phi)\|_{q,T} \leq C h_T^{m-\ell} |\phi|_{m,q,S_T}, \tag{1.15}$$

and

$$\|\lambda - \mathcal{I}_{M_h} \lambda\|_{L^q(T)/\mathbf{R}} \leq C h_T^{m-1} |\lambda|_{W^{m-1,q}(S_T)/\mathbf{R}}. \tag{1.16}$$

On the boundary, ∂T, we use the trace inequality [8, Theorem 3.3, p. 43] and scale it appropriately, *i.e.*, for $w \in W^{1,q}(T)$ we obtain the estimate

$$\|w\|_{q,\partial T} \leq C\big(h_T^{-1/q} \|w\|_{q,T} + h_T^{1-1/q} |w|_{1,q,T}\big),$$

and hence

$$\|\phi - \mathcal{I}_{X_h}\phi\|_{q,\partial T} \leq C h_T^{m-1/q} |\phi|_{m,q,S_T}. \tag{1.17}$$

We also use inverse estimates, see for example [2, Theorem 4.5.3, p. 111]. For any $T \in \mathcal{T}$, let V be a finite dimensional subspace of $W^{k,q}(T) \cap W^{m,s}(T)$, where $1 \leq q \leq \infty$ and $1 \leq s \leq \infty$ and $0 \leq m \leq k$. Then there exist a constant C such that for all $v \in V$

$$\|v\|_{k,q,T} \leq C h_T^{m-k+n/q-n/s} \|v\|_{m,s,T}. \tag{1.18}$$

2. Error analysis

We consider the error in the finite element solution to (1.13),

$$e_u := u_h - u \quad \text{and} \quad e_p := p_h - p,$$

and note that $(e_u, e_p) \in \mathcal{W}^{1,q}$, since the finite elements are conforming.

Define the residual in the momentum equation (me) by

$$R_{\text{me}} := f + \Delta u_h - \nabla p_h \in W^{-1,q}(\Omega)^n, \tag{2.1}$$

and the residual in the compressibility constraint (cc) by

$$R_{\text{cc}} := g - \nabla \cdot u_h \in L^q(\Omega), \tag{2.2}$$

where we note that $\int_\Omega R_{\text{cc}}\, dx = 0$.

In weak form the residual becomes

$$\mathcal{R}((u_h, p_h), (\phi, \lambda)) := \langle f, \phi \rangle + (g, \lambda) - \mathcal{L}((u_h, p_h), (\phi, \lambda)), \tag{2.3}$$

for all $(\phi, \lambda) \in \mathcal{W}^{1,q'}$.

From (1.3) we obtain the identity

$$\mathcal{L}((e_u, e_p), (\phi, \lambda)) = \mathcal{R}((u_h, p_h), (\phi, \lambda)) \quad \forall (\phi, \lambda) \in \mathcal{W}^{1,q'} \tag{2.4}$$

and from (1.13) and it follows

$$\mathcal{R}((u_h, p_h), (\phi_h, \lambda_h)) = 0 \quad \forall (\phi_h, \lambda_h) \in \mathcal{W}_h, \tag{2.5}$$

which is the classical Galerkin orthogonality.

Inspired by [7, Lemma 3.1] we now provide the following lemma.

Lemma 2.1. *For $q \in [1, \infty]$, and $m = 1, 2$, there is a constant C such that*

$$|\mathcal{R}((u_h, p_h), (\phi, \lambda))| \leq C \eta_{m,q} \big(|\phi|_{m,q'} + |\lambda|_{W^{m-1,q'}/\mathbf{R}} \big),$$

for all $(\phi, \lambda) \in \mathcal{W}^{m,q'}$ *where*

$$\eta_{m,q} = \begin{cases} \left(\sum_{T\in\mathcal{T}} \eta_{m,q,T}^q\right)^{1/q} & \text{for } q \in [1,\infty), \\ \max_{T\in\mathcal{T}} \eta_{m,\infty,T} & \text{for } q = \infty, \end{cases}$$

with

$$\eta_{m,q,T} = h_T^m \|R_{\mathrm{me}}\|_{q,T} + \frac{1}{2} h_T^{m-1/q'} \|[\partial_\nu u_h]\|_{q,\partial T\setminus\partial\Omega} + h_T^{m-1} \|R_{\mathrm{cc}}\|_{q,T}.$$

Here $[\partial_\nu u_h]$ *denotes the jump across* ∂T *in the normal derivative,* $\partial_\nu u_h = \nu \cdot \nabla u_h$*, where* ν *denotes the outward normal to* ∂T*.*

Proof. By (2.5) and by integration by parts

$$\begin{aligned} \mathcal{R}((u_h,p_h),(\phi,\lambda)) &= \mathcal{R}((u_h,p_h),(\phi-\mathcal{I}_{X_h}\phi,\lambda-\mathcal{I}_{M_h}\lambda)) \\ &= \sum_{T\in\mathcal{T}} \big((f+\Delta u_h-\nabla p_h, \phi-\mathcal{I}_{X_h}\phi)_T \\ &\qquad + \frac{1}{2}([\partial_\nu u_h], \phi-\mathcal{I}_{X_h}\phi)_{\partial T\setminus\partial\Omega} \\ &\qquad + (g-\nabla\cdot u_h, \lambda-\mathcal{I}_{M_h}\lambda)_T\big). \end{aligned}$$

Since $\int_\Omega (g - \nabla\cdot u_h)\,dx = 0$, we have

$$(g-\nabla\cdot u_h, \lambda-\mathcal{I}_{M_h}\lambda)_T = \inf_{c\in\mathbf{R}} (g-\nabla\cdot u_h, \lambda-\mathcal{I}_{M_h}\lambda + c)_T$$

and hence by Hölder's inequality,

$$\begin{aligned} &|\mathcal{R}((u_h,p_h),(\phi,\lambda))| \\ &\quad\le \sum_{T\in\mathcal{T}} \big(\|f+\Delta u_h-\nabla p_h\|_{q,T}\|\phi-\mathcal{I}_{X_h}\phi\|_{q',T} \\ &\qquad + \frac{1}{2}\|[\partial_\nu u_h]\|_{q,\partial T\setminus\partial\Omega}\|\phi-\mathcal{I}_{X_h}\phi\|_{q',\partial T\setminus\partial\Omega} \\ &\qquad + \|g-\nabla\cdot u_h\|_{L^q(T)}\|\lambda-\mathcal{I}_{M_h}\lambda\|_{L^{q'}(T)/\mathbf{R}}\big). \end{aligned} \tag{2.6}$$

Thus, with the interpolation estimates (1.15)–(1.17) in (2.6) we get

$$
\begin{aligned}
&|\mathcal{R}((u_h,p_h),(\phi,\lambda))| \\
&\quad\le C\sum_{T\in\mathcal{T}}\Big(h_T^m\big(\|f+\Delta u_h-\nabla p_h\|_{q,S_T} \\
&\qquad+\frac{1}{2}h_T^{1/q'}\,\|[\partial_n u_h]\|_{q,\partial T\setminus\partial\Omega}|\phi|_{m,q',S_T} \\
&\qquad+h_T^{m-1}\|g-\nabla\cdot u_h\|_{L^q(T)}|\lambda|_{W^{m-1,q'}(S_T)/\mathbf{R}}\Big).
\end{aligned}
\tag{2.7}
$$

Finally, we conclude the proof by using Hölder's inequality for sums and the notation in (2.1) and (2.2). □

Let $(\tilde{u},\tilde{p})$ be the solution to the dual problem (1.5). By choosing $(\phi,\lambda)=(\tilde{u},\tilde{p})$ in (2.4) we get

$$\mathcal{L}((e_u,e_p),(\tilde{u},\tilde{p}))=\mathcal{R}((u_h,p_h),(\tilde{u},\tilde{p})),$$

and by choosing $(\phi,\lambda)=(e_u,e_p)$ in (1.5) we obtain

$$\mathcal{L}((e_u,e_p),(\tilde{u},\tilde{p}))=\langle e_u,\tilde{f}\rangle+(e_p,\tilde{g}).$$

Thus

$$\langle e_u,\tilde{f}\rangle+(e_p,\tilde{g})=\mathcal{R}((u_h,p_h),(\tilde{u},\tilde{p})). \tag{2.8}$$

In order to proceed in the error analysis we need to choose the data in the dual problem in a certain way. Let $\delta=\delta_{x_0,\rho/2}\in C_0^\infty(\Omega)$ be a regularization of the Dirac distribution at $x_0\in\Omega$, that is, let

$$\operatorname{supp}(\delta)\subset\mathcal{B}(x_0;\rho/2),\qquad \int_{\mathbf{R}^n}\delta\,dx=1,\qquad 0\le\delta\le C\rho^{-n}, \tag{2.9}$$

where $\mathcal{B}(x_0;\rho/2)$ denotes the ball with center in x_0 and radius $\rho/2$ chosen such that

$$\rho\le h_{\min}^{\sigma}, \tag{2.10}$$

where $\sigma>0$ will be specified in the proofs of Lemmas 2.2–2.4 below. For $q\in[1,\infty]$ it follows that

$$|\delta|_{k,q}\le C\rho^{-n(1-1/q)-k}. \tag{2.11}$$

In the remainder of this section we state and prove three lemmas providing estimates of the following kind

$$\|e_u\|_\infty \lesssim |(e_{u_i}, \delta_{x_0,\rho/2})|,$$
$$\|\nabla e_u\|_\infty \lesssim |(e_u, D_i\delta_{x_0,\rho/2}e_j)|,$$
$$\|e_p\|_\infty \lesssim |(e_p, \delta_{x_0,\rho/2})|,$$

where e_{u_i} denotes the i:th component of e_u and where e_j is the j:th unit vector. We stress that x_0 may be different in the there estimates. With these estimates we will be able to make a connection to the estimate in Lemma 2.1, which in turn is crucial for the final pointwise error analysis.

In order to obtain these estimates we will have to assume that e_u and e_p are continuous. This will be the case for e_u provided the data is sufficient regular due to Theorem 1.1, whereas for e_p we also have to impose further constraints on the domain Ω, see Remark 1.2. We note that ∇e_u is not continuous since ∇u_h is discontinuous. However, with the same assumptions as for e_p we derive an estimate that includes jump terms of the same type as in the right hand side of the estimate in Lemma 2.1.

Lemma 2.2. *Let $\Omega \subset \mathbf{R}^n$, $n = 2, 3$, be a polyhedral domain and let $x_0 \in \Omega$ and i be such that $\|e_u\|_\infty = |e_{u_i}(x_0)|$. Then for data to* (1.1) *as in Theorem 1.1 and for some $q > n$ there is a constant C such that*

$$\|e_u\|_\infty \le |(e_{u_i}, \delta)| + Ch_{\min}^\beta\big(\|f\|_{-1,q} + \|g\|_q\big),$$

where $\delta = \delta_{x_0,\rho/2}$ is the regularized Dirac distribution (2.9) and β may be chosen arbitrarily large.

We note that the lemma is meaningful since due to Theorem 1.1 and Remark 1.1 there is $q > n$ such that $e_u \in W_0^{1,q}(\Omega)^n$.

Proof. By Sobolev's imbedding theorem, see [1, p. 98], $W_0^{1,q}(\Omega)^n \subset C^{0,\gamma}(\overline{\Omega})^n$ for some γ such that $0 < \gamma \le 1 - n/q$. Consequently, by the mean value theorem there is $x_1 \in \mathcal{B}(x_0, \rho/2) \cap \overline{\Omega}$ such that $(e_{u_i}, \delta) = e_{u_i}(x_1)$ and thus

$$\|e_u\|_\infty \le |(e_{u_i}, \delta)| + |e_{u_i}(x_0) - e_{u_i}(x_1)|.$$

We estimate the last term in the right hand side above. By Sobolev's inequality

$$|e_{u_i}(x_0) - e_{u_i}(x_1)| \le C\rho^\gamma \|e_{u_i}\|_{C^{0,\gamma}(\mathcal{B}(x_0,\rho/2)\cap\overline{\Omega})} \le C\rho^\gamma \|e_u\|_{1,q}.$$

By the triangle inequality,

$$\|e_u\|_{1,q} \le \|u\|_{1,q} + \|u_h\|_{1,q},$$

and by Theorem 1.1,

$$\|u\|_{1,q} \le C\big(\|f\|_{-1,q} + \|g\|_q\big),$$

and by the inverse estimate (1.18) and the inf-sup condition (1.14),

$$\|u_h\|_{1,q} \le Ch_{\min}^{n(1/q-1/2)}\|u_h\|_{1,2} \le Ch_{\min}^{n(1/q-1/2)}\big(\|f\|_{-1,q} + \|g\|_q\big).$$

Thus, with (2.10) we obtain

$$|e_{u_i}(x_0) - e_{u_i}(x_1)| \le Ch_{\min}^{\beta}\big(\|f\|_{-1,q} + \|g\|_q\big),$$

where $\beta = \gamma\sigma + n(1/q - 1/2)$ may be chosen arbitrarily large by taking σ large. □

Lemma 2.3. *Let $\Omega \subset \mathbf{R}^n$, $n = 2, 3$, be a polyhedral domain such that the solution to (1.1) with data as in Theorem 1.3 is continuous in the sense that $(u, p) \in \mathcal{W}^{2,q}$, for $q > n$. Let $x_0 \in \Omega$, i and j be such that $\|\nabla e_u\|_\infty = |D_i e_{u_j}(x_0)|$. Then there are constants $C_{1,2}$ such that*

$$\begin{aligned}\|\nabla e_u\|_\infty \le |(e_u, D_i\delta e_j)| + C_1 h_{\min}^{\beta}\big(\|f\|_q + |g|_{1,q}\big)\\ + C_2 \max_{T\in\mathcal{T}}\|[\partial_\nu u_h]\|_{\infty,\partial T\setminus\partial\Omega},\end{aligned}$$

where $\delta = \delta_{x_0,\rho/2}$ is the regularized Dirac distribution (2.9), β may be chosen arbitrarily large, and $[\partial_\nu u_h]$ is the jump as described in Lemma 2.1.

We note that the lemma is meaningful since with additional constraints on the domain Ω as in Remark 1.2 there is $q > n$ such that $u \in W^{2,q}(\Omega)^n$ so that $u \in W^{1,\infty}(\Omega)^n$. Note also that ∇u_h is discontinuous across ∂T for $T \in \mathcal{T}$ which need to be taken into account proving Lemma 2.3. However, ∇u_h is continuous in the interior of each $T \in \mathcal{T}$.

Proof. The idea of the proof is the same as for Lemma 2.2. Let

$$\mathcal{B}_{\mathcal{T}} = \bigcup\{T \in \mathcal{T} : T \cap \mathcal{B}(x_0, \rho/2) \neq \emptyset\},$$

where we for simplicity assume that $\mathcal{B}_{\mathcal{T}}$ is convex and note that $\operatorname{card}(\mathcal{B}_{\mathcal{T}}) \le C$ due to the regularity in the triangulation.

By the mean value theorem there are $x_T \in \mathcal{B}(x_0, \rho/2) \cap T$ for $T \in \mathcal{B}_{\mathcal{T}}$ such that

$$(D_i e_{u_j}, \delta) = \sum_{T\in\mathcal{B}_{\mathcal{T}}} (D_i e_{u_j}, \delta)_{\mathcal{B}(x_0,\rho/2)\cap T} = \sum_{T\in\mathcal{B}_{\mathcal{T}}} D_i e_{u_j}(x_T) \int_{\mathcal{B}(x_0,\rho/2)\cap T} \delta\, dx,$$

where $\int_{\mathcal{B}(x_0,\rho/2)\cap T} \delta \, dx < 1$ and thus

$$\|\nabla e_u\|_\infty \leq |(e_u, D_i\delta e_j)| + \sum_{T\in\mathcal{B}_\mathcal{T}} |D_i e_{u_j}(x_0) - D_i e_{u_j}(x_T)|, \tag{2.12}$$

since by integration by parts $(D_i e_{u_j}, \delta) = -(e_u, D_i\delta e_j)$.

We estimate the terms in sum above. For $T \in \mathcal{B}_\mathcal{T}$ consider the line from x_0 to x_T and for $T_\ell \in \mathcal{B}_\mathcal{T}$ suppose this line intersect $m+1$ n-simplices T_ℓ and m boundaries ∂T_ℓ at points x_ℓ for $\ell = 1, \dots, m$. Note that m is bounded from above since $\operatorname{card}(\mathcal{B}_\mathcal{T}) \leq C$. Let x_ℓ^- and x_ℓ^+ be the limits at x_ℓ going from x_0 and x_T respectively. Set $x_0^+ = x_0$ and $x_{m+1}^- = x_T$. We estimate

$$\begin{aligned} |D_i e_{u_j}(x_0) - D_i e_{u_j}(x_T)| \leq &\sum_{\ell=0}^{m} |D_i e_{u_j}(x_\ell^+) - D_i e_{u_j}(x_{\ell+1}^-)| \\ &+ \sum_{\ell=1}^{m} |D_i e_{u_j}(x_\ell^-) - D_i e_{u_j}(x_\ell^+)|. \end{aligned} \tag{2.13}$$

For each term in the first sum above we may now proceed as in the proof of Lemma 2.2. By Sobolev's and the triangle inequality we get

$$\begin{aligned} |D_i e_{u_j}(x_\ell^+) - D_i e_{u_j}(x_{\ell+1}^-)| \leq& C\rho^\gamma \|D_i e_{u_j}\|_{C^{0,\gamma}(\mathcal{B}(x_0,\rho/2)\cap T_\ell)} \\ \leq& C\rho^\gamma \|e_u\|_{2,q,T_\ell} \\ \leq& C\rho^\gamma \big(\|u\|_{2,q} + \|u_h\|_{2,q,T_\ell}\big). \end{aligned}$$

By Theorem 1.3 we have

$$\|u\|_{2,q} \leq C\big(\|f\|_q + |g|_{1,q}\big),$$

and by the inverse estimate (1.18) and the inf-sup condition (1.14)

$$\|u_h\|_{2,q,T_\ell} \leq C h_{T_\ell}^{-1+n(1/q-1/2)} \|u_h\|_{1,2,T_\ell} \leq C h_{\min}^{-1+n(1/q-1/2)} \big(\|f\|_{-1,q} + \|g\|_q\big),$$

since $q > n$.

Thus, with (2.10) and for $T_\ell \in \mathcal{B}_\mathcal{T}$ we obtain the uniform estimate

$$|D_j e_{u_i}(x_\ell^+) - D_j e_{u_i}(x_\ell^-)| \leq C h_{\min}^{\beta} \big(\|f\|_q + |g|_{1,q}\big), \tag{2.14}$$

where $\beta = \gamma\sigma - 1 + n(1/q - 1/2)$ may be chosen arbitrarily large by taking σ large.

As for the terms in the second sum in (2.13) and for $T_\ell \in \mathcal{B}_\mathcal{T}$ we use the following uniform estimate

$$|D_j e_{u_i}(x_\ell^-) - D_j e_{u_i}(x_\ell^+)| \leq \max_{T\in\mathcal{T}} \|[\partial_\nu u_h]\|_{\infty,\partial T\setminus\partial\Omega}. \tag{2.15}$$

Finally, (2.13) – (2.15) in (2.12) concludes the proof. □

Lemma 2.4. *Let $\Omega \subset \mathbf{R}^n$, $n = 2, 3$, be a polyhedral domain such that the solution to (1.1) with data as in Theorem 1.3 is continuous in the sense that $(u, p) \in \mathcal{W}^{2,q}$, for some $q > n$. Let e_p be such that $\int_\Omega e_p \, dx = 0$ and let $x_0 \in \Omega$ be such that $\|e_p\|_\infty = |e_p(x_0)|$. Then there is a constant C such that*

$$\|e_p\|_\infty \leq |(e_p, \delta)| + Ch_{\min}^\beta \big(\|f\|_q + |g|_{1,q}\big),$$

where $\delta = \delta_{x_0,\rho/2}$ is the regularized Dirac distribution (2.9) and β may be chosen arbitrarily large.

We note that the lemma is meaningful since with additional constraints on the domain Ω as in Remark 1.2 there is $q > n$ such that $e_p \in W^{1,q}(\Omega)$ so that $e_p \in L^\infty(\Omega)$.

Proof. The idea of the proof is the same as for Lemma 2.2. By assumption $p \in W^{1,q}(\Omega)$ for $q > n$ and hence it follows by Sobolev's imbedding theorem that e_p is continuous. Consequently, by the mean value theorem there is $x_1 \in \mathcal{B}(x_0, \rho/2) \cap \overline{\Omega}$ such that $(e_p, \delta) = e_p(x_1)$ and thus

$$\|e_p\|_\infty \leq |(e_p, \delta)| + |e_p(x_0) - e_p(x_1)|.$$

We estimate the last term above. By Sobolev's inequality

$$|e_p(x_0) - e_p(x_1)| \leq C\rho^\gamma \|e_p\|_{C^{0,\gamma}(\mathcal{B}(x_0,\rho/2)\cap\overline{\Omega})} \leq C\rho^\gamma \|e_p\|_{1,q}.$$

By the triangle inequality

$$\|e_p\|_{1,q} \leq \|p\|_{1,q} + \|p_h\|_{1,q},$$

and Theorem 1.3

$$\|p\|_{1,q} \leq C\big(\|f\|_q + |g|_{1,q}\big),$$

and by the inverse estimate and the inf-sup condition (1.14)

$$\|p_h\|_{1,q} \leq Ch_{\min}^{-1+n(1/q-1/2)} \|p_h\|_2 \leq Ch_{\min}^{-1+n(1/q-1/2)} \big(\|f\|_{-1,q} + \|g\|_q\big).$$

Thus with (2.10) we obtain

$$|e_p(x_0) - e_p(x_1)| \leq Ch_{\min}^\beta \big(\|f\|_q + |g|_{1,q}\big),$$

where $\beta = \gamma\sigma - 1 + n(1/q - 1/2)$ may be chosen arbitrarily large by taking σ large. □

3. A priori estimates of the dual solution

We consider the dual problem (1.4) for specific choices of data so that we may estimate the scaling of the constants in (1.6) and (1.10) as $q \downarrow 1$. For (1.6) we will consider $(\tilde{f}, \tilde{g}) = (D_i\delta e_j, 0)$ or $(\tilde{f}, \tilde{g}) = (0, \delta - |\Omega|^{-1})$ and for (1.10) we will consider $(\tilde{f}, \tilde{g}) = (\delta e_i, 0)$, where δ is the regularized Dirac distribution (2.11). We proceed as in [14, Theorem 3.1] and [4, Lemma 2.2]. The analysis relies on the explicit knowledge of how the constant in Sobolev's inequality scales as $q \downarrow 1$, which can be estimated by using the the best constant in the Sobolev inequality, where the dependence on the dimension n and the exponent q appear explicitly. We quote Sobolev's inequality from [9, Theorem 7.10, p. 155]. Let ω be a bounded domain in $\mathbf{R}^n$, $n = 2, 3$. Then there is a constant C such that for any $v \in W_0^{1,s}(\omega)^d$, $d = 1, \dots, n$, and for $1 \le s < n$

$$\|v\|_{ns/(n-s),\omega} \le C|v|_{1,s,\omega}, \tag{3.1}$$

where $C = C(n, s)$ scales like

$$C \le \gamma \left(n\frac{s-1}{n-s} \right)^{1-1/s}, \tag{3.2}$$

and where $\gamma = \gamma(n, s) < \infty$ as $s \uparrow n$.

In the analysis below we will find it useful to have (3.1) and (3.2) formulated somewhat differently. By rearranging the exponents in (3.1) and estimating the constant (3.2) accordingly we conclude that, for any $v \in W_0^{1,nr/(n+r)}(\omega)^d$ and for $n/(n-1) \le r < \infty$,

$$\|v\|_{r,\omega} \le Cr^{1-1/n}|v|_{1,nr/(n+r),\omega}. \tag{3.3}$$

The following lemma is a consequence of (3.3).

Lemma 3.1. *Let $\omega \subset \mathbf{R}^n$, $n = 2, 3$, be a bounded domain. Then there is a constant C such that, if $v \in L^q(\omega)^d$, $d = 1, \dots, n$,*

$$\|\nabla^{k-1}v\|_{-k,\tilde{q},\omega} \le C(q-1)^{-1+1/n}\|v\|_{q,\omega}, \tag{3.4}$$

for $\tilde{q} = nq/(n-q)$ and $1 < q \le n$.

Proof. By integration by parts and with Hölder's inequality in the definition of the dual norm (1.2) we estimate

$$\begin{aligned} \|\nabla^{k-1} v\|_{-k,\tilde{q},\omega} &= \sup_{\varphi \in C_0^\infty(\omega)^d} \frac{|\langle v, \nabla^{k-1}\varphi\rangle|}{\|\varphi\|_{k,\tilde{q}',\omega}} \\ &\leq \|v\|_{q,\omega} \sup_{\varphi \in C_0^\infty(\omega)^n} \frac{|\varphi|_{k-1,q',\omega}}{\|\varphi\|_{k,\tilde{q}',\omega}}. \end{aligned} \tag{3.5}$$

Since $1 < q \leq n$ implies $n/(n-1) \leq q' < \infty$, we may use Sobolev's inequality (3.3) to estimate,

$$|\varphi|_{k-1,q',\omega} \leq C q'^{1-1/n} |\varphi|_{k,\tilde{q}',\omega}, \tag{3.6}$$

because $nq'/(n+q') = \tilde{q}'$. Thus, inserting (3.6) in (3.5) concludes the proof. □

As in [14, 4] we introduce a dyadic partition of Ω. Let $d_j = 2^j \rho$ for $j \in \mathbf{N}$ and $d_{-1} = 0$. Define the partition of Ω,

$$A_j = \{x \in \Omega : d_{j-1} \leq |x - x_0| \leq d_j\}, \tag{3.7}$$

and the supersets to A_j,

$$B_j = \{x \in \Omega : 2^{-1} d_{j-1} \leq |x - x_0| \leq 2 d_j\}. \tag{3.8}$$

From this definition we get the simple estimate

$$|B_j| \leq C d_j^n = C 2^{jn} \rho^n. \tag{3.9}$$

Moreover, let $\eta_j \in C_0^\infty(B_j)$ be a mollifier such that, $\eta_j = 1$ in a neighborhood of A_j and such that for $s \in [1, \infty]$,

$$|\eta_j|_{k,s,B_j} \leq C d_j^{n/s-k}. \tag{3.10}$$

Generalizing the last estimate in [14, Proof of Theorem 3.1] we get. For $a > 1$ and as $q \downarrow 1$ we have,

$$\sum_{j=0}^{\infty} 2^{-ja(1-1/q)} = \frac{1}{1 - 2^{-a(1-1/q)}} \leq \frac{C}{q-1}. \tag{3.11}$$

Finally, we recall the following two generalizations of Hölder's inequality. Let $1 \leq q \leq \infty$, $q \leq r \leq \infty$ and $q \leq s \leq \infty$ such that

$$\frac{1}{q} = \frac{1}{r} + \frac{1}{s}$$

and let $u \in L^r(\omega)$ and $v \in L^s(\omega)$. Then $uv \in L^q(\omega)$ and

$$\|uv\|_{q,\omega} \leq \|u\|_{r,\omega}\|v\|_{s,\omega}. \tag{3.12}$$

In the second generalization we estimate the duality pairing. For a vector space V let $u \in V'$ and $v \in V$. Then

$$|\langle u, v\rangle| \leq \|u\|_{V'}\|v\|_V. \tag{3.13}$$

In particular, when $u \in W^{-k,q}(\omega)$ and $v \in W_0^{k,q'}(\omega)$ we get

$$|\langle u, v\rangle| \leq \|u\|_{-k,q,\omega}\|v\|_{k,q',\omega}. \tag{3.14}$$

3.1. $\mathcal{W}^{1,q}$**-estimates as** $q \downarrow 1$**.** In the following theorem we assume that we have the higher regularity in Remark 1.2.

Theorem 3.2. *Let $\Omega \subset \mathbf{R}^n$, $n = 2, 3$, be a polyhedral domain such that the solution to* (1.4) *with data as in Theorem 1.3 is continuous in the sense that $(\tilde{u}, \tilde{p}) \in \mathcal{W}^{2,q}$ for some $q > n$. Then for $1 < q < 2$ there is a constant C such that the solution $(\tilde{u}, \tilde{p})$ to (1.4) with $(\tilde{f}, \tilde{g}) = (D_i\delta e_j, 0)$ or $(\tilde{f}, \tilde{g}) = (0, \delta - |\Omega|^{-1})$ satisfies the inequality*

$$\|\tilde{u}\|_{1,q} + \|\tilde{p}\|_{L^q/\mathbf{R}} \leq C(q-1)^{-2+1/n}\rho^{-n(1-1/q)}.$$

Proof. Let A_j, B_j and η_j be as in (3.7)–(3.10). Choose a fixed value $\tilde{q} = n/(n-1)$. Let $\bar{p} = \tilde{p} + c$ for a fixed $c \in \mathbf{R}$. By Hölder's inequality

$$\begin{aligned}
\|\tilde{u}\|_{1,q} + \|\tilde{p}\|_{L^q/\mathbf{R}} &\leq \sum_{j=0}^{\infty} \big(\|\tilde{u}\|_{1,q,A_j} + \|\bar{p}\|_{q,A_j}\big) \\
&\leq \sum_{j=0}^{\infty} \big(\|\eta_j\tilde{u}\|_{1,q,B_j} + \|\eta_j\bar{p}\|_{q,B_j}\big) \\
&\leq \sum_{j=0}^{\infty} |B_j|^{1/q-1/\tilde{q}} \big(\|\eta_j\tilde{u}\|_{1,\tilde{q},B_j} + \|\eta_j\bar{p}\|_{\tilde{q},B_j}\big).
\end{aligned} \tag{3.15}$$

Notice that $\eta_j\tilde{u}$ and $\eta_j\bar{p}$ satisfy (1.4) in Ω with right hand side $\tilde{f} = \tilde{f}_j = \Delta(\eta_j\tilde{u}) + \nabla(\eta_j\bar{p})$ and $\tilde{g} = \tilde{g}_j = \nabla \cdot (\eta_j\tilde{u})$, where $\tilde{f}_j$ and $\tilde{g}_j$ vanish outside

B_j. Hence, for each term in (3.15) we can apply Theorem 1.1,

$$
\begin{aligned}
\|\eta_j\tilde{u}\|_{1,\tilde{q},B_j} + \|\eta_j\bar{p}\|_{\tilde{q},B_j} &= \|\eta_j\tilde{u}\|_{1,\tilde{q},\Omega} + \|\eta_j\bar{p}\|_{\tilde{q},\Omega} \\
&\le C\big(\|\Delta(\eta_j\tilde{u}) + \nabla(\eta_j\bar{p})\|_{-1,\tilde{q},B_j} + \|\nabla\cdot(\eta_j\tilde{u})\|_{\tilde{q},B_j}\big) \\
&\le C\big(\|\eta_j(\Delta\tilde{u}+\nabla\bar{p}) + 2\nabla\eta_j\cdot\nabla\tilde{u} + \Delta\eta_j\tilde{u} + \nabla\eta_j\bar{p}\|_{-1,\tilde{q},B_j} \\
&\quad + \|\nabla\eta_j\cdot\tilde{u} + \eta_j\nabla\cdot\tilde{u}\|_{\tilde{q},B_j}\big) \\
&\le C\big(\|\eta_j\tilde{f}\|_{-1,\tilde{q},B_j} + \|\eta_j\tilde{g}\|_{\tilde{q},B_j} + \|\nabla\eta_j\bar{p}\|_{-1,\tilde{q},B_j} \\
&\quad + \|\nabla\eta_j\cdot\tilde{u}\|_{\tilde{q},B_j} + \|2\nabla\eta_j\cdot\nabla\tilde{u} + \Delta\eta_j\tilde{u}\|_{-1,\tilde{q},B_j}\big),
\end{aligned}
\tag{3.16}
$$

where $C = C(n,\tilde{q},\Omega)$.

We estimate the right hand side of (3.16) in a few steps. By integration by parts

$$
\|2\nabla\eta_j\cdot\nabla\tilde{u} + \Delta\eta_j\tilde{u}\|_{-1,\tilde{q},B_j} \le \|\nabla\eta_j\cdot\nabla\tilde{u}\|_{-1,\tilde{q},B_j} + \sup_{\varphi\in C_0^\infty(B_j)^n}\frac{(\nabla\eta_j,\tilde{u}\cdot\nabla\varphi)_{B_j}}{\|\varphi\|_{1,\tilde{q}',B_j}}.
$$

Since $(\nabla\eta_j\bar{p},\varphi) \le \|\bar{p}\|_{W^{1,n}(B_j)'}|\nabla\eta_j\cdot\varphi|_{1,n,B_j}$, notice that the dual exponent to $\tilde{q}$ is $\tilde{q}' = n$,

$$
\|\nabla\eta_j\tilde{p}\|_{-1,\tilde{q},B_j} \le \|\bar{p}\|_{W^{1,n}(B_j)'}\sup_{\varphi\in C_0^\infty(B_j)^n}\frac{|\nabla\eta_j\cdot\varphi|_{1,n,B_j}}{\|\varphi\|_{1,n,B_j}},
$$

and since $(\nabla\eta_j\cdot\nabla\tilde{u},\varphi) = -(\tilde{u},\nabla(\nabla\eta_j\cdot\varphi))$,

$$
\|\nabla\eta_j\cdot\nabla\tilde{u}\|_{-1,\tilde{q},B_j} \le \|\tilde{u}\|_{\tilde{q},B_j}\sup_{\varphi\in C_0^\infty(B_j)^n}\frac{|\nabla\eta_j\cdot\varphi|_{1,n,B_j}}{\|\varphi\|_{1,n,B_j}}.
$$

Now by Hölder's inequality

$$
|\nabla\eta_j\cdot\varphi|_{1,n,B_j} \le |\eta_j|_{1,\infty,B_j}|\varphi|_{1,n,B_j} + \|\nabla^2\eta_j\varphi\|_{n,B_j},
$$

and moreover by (3.12) with s such that $1/n = 1/s + 1/q'$, (3.3), and Hölder's inequality

$$
\begin{aligned}
\|\nabla^2\eta_j\varphi\|_{n,B_j} &\le |\eta_j|_{2,s,B_j}\|\varphi\|_{q',B_j} \\
&\le C(q')^{-1+1/n}|\eta_j|_{2,s,B_j}|\varphi|_{1,nq'/(n+q'),B_j} \\
&\le C|B_j|^{1-1/q}(q-1)^{-1+1/n}|\eta_j|_{2,s,B_j}|\varphi|_{1,n,B_j}.
\end{aligned}
\tag{3.17}
$$

Finally, by Hölder's inequality

$$
\|\nabla\eta_j\cdot\tilde{u}\|_{\tilde{q},B_j} + \sup_{\varphi\in C_0^\infty(B_j)^n}\frac{(\nabla\eta_j,\tilde{u}\cdot\nabla\varphi)}{\|\varphi\|_{1,\tilde{q}',B_j}} \le 2|\eta_j|_{1,\infty,B_j}\|\tilde{u}\|_{\tilde{q},B_j}.
$$

Thus, with the above estimates in (3.16) we obtain

$$
\begin{aligned}
\|\eta_j\tilde{u}\|_{1,\tilde{q},B_j} + \|\eta_j\bar{p}\|_{L^{\tilde{q}}(B_j)/\mathbf{R}} &\le C_I\|\eta_j\tilde{f}\|_{-1,\tilde{q},B_j} + C_{II}\|\eta_j\tilde{g}\|_{\tilde{q},B_j} \\
&\quad + C_{III}\big(|\eta_j|_{1,\infty,B_j} + |B_j|^{1-1/q}(q-1)^{-1+1/n}|\eta_j|_{2,s,B_j}\big) \\
&\quad \times \big(\|\tilde{u}\|_{\tilde{q},B_j} + \|\bar{p}\|_{W^{1,n}(B_j)'}\big) \\
&= I_j + II_j + III_j.
\end{aligned} \tag{3.18}
$$

With (3.18) we now estimate (3.15) in three steps. Recall (3.9) that will repeatedly be used in the estimates below.

I. For data $\tilde{f} = D_i\delta e_\ell$ and by integration by parts we obtain by the same argument as in (3.17) and with the same exponents

$$
\begin{aligned}
\|\eta_j D_i\delta e_\ell\|_{-1,\tilde{q},B_j} &\le C\|\delta\|_{\tilde{q},B_j} \sup_{\varphi\in C_0^\infty(B_j)^n} \frac{|\eta_j\varphi|_{1,n,B_j}}{\|\varphi\|_{1,n,B_j}} \\
&\le C\|\delta\|_{\tilde{q},B_j}\big(\|\eta_j\|_{\infty,B_j} + |B_j|^{1-1/q}(q-1)^{-1+1/n}|\eta_j|_{1,s,B_j}\big).
\end{aligned}
$$

Since $\operatorname{supp}(\delta)\cap B_j = \emptyset$ for $j \ge 1$ and with (2.11) and for ρ sufficiently small

$$
\begin{aligned}
\sum_{j=0}^{\infty} |B_j|^{1/q-1/\tilde{q}} I_j &\le C\rho^{n(1/q-1/\tilde{q})}(q-1)^{-1+1/n}\|\delta\|_{\tilde{q}} \\
&\le C\rho^{-n(1-1/q)}(q-1)^{-1+1/n},
\end{aligned} \tag{3.19}
$$

where we used $n/q - n/\tilde{q} - n(1-1/\tilde{q}) + n(1-1/q) + n/s - 1 = -n(1-1/q)$.

II. For data $\tilde{g} = \delta - |\Omega|^{-1}$ and since $\operatorname{supp}(\delta)\cap B_j = \emptyset$ for $j \ge 1$ and with (3.1) and (2.11)

$$
\sum_{j=0}^{\infty} |B_j|^{1/q-1/\tilde{q}} II_j \le C\rho^{n/q-n/\tilde{q}}\|\nabla\delta\|_1 \le C\rho^{-n(1-1/q)}, \tag{3.20}
$$

where we used $n/q - n/\tilde{q} - 1 = -n(1-1/q)$.

III. By Hölder's inequality and since $q < 2$

$$
\begin{aligned}
|B_j|^{1/q-1/\tilde{q}} III_j &\le C d_j^{n/q-n/\tilde{q}}\big(|\eta_j|_{1,\infty,B_j} + d_j^{n(1-1/q)}(q-1)^{-1+1/n}|\eta_j|_{2,s,B_j}\big) \\
&\quad \times \big(\|\tilde{u}\|_{\tilde{q},B_j} + \|\bar{p}\|_{W^{1,n}(B_j)'}\big) \\
&\le C d_j^{-n(1-1/q)}\big(1 + (q-1)^{-1+1/n}\big)\big(\|\tilde{u}\|_{\tilde{q},B_j} + \|\bar{p}\|_{W^{1,n}(B_j)'}\big) \\
&\le C d_j^{-n(1-1/q)}(q-1)^{-1+1/n}\big(\|\tilde{u}\|_{\tilde{q},B_j} + \|\bar{p}\|_{W^{1,n}(B_j)'}\big),
\end{aligned}
$$

where we used $n/q - n/\tilde{q} - 1 = -n(1-1/q)$ and $n/q - n/\tilde{q} + n - n/q + n/s - 2 = -n(1-1/q)$.

Adding all the terms and by Hölder's inequality in the sum with exponent $\tilde{q}$, with conjugate exponent $\tilde{q}' = n$, estimating the geometric sum as in (3.11) and by Corollary 1.4

$$\begin{aligned}\sum_{j=0}^{\infty} |B_j|^{1/q-1/\tilde{q}} III_j &\le C(q-1)^{-1+1/n} \left(\sum_{j=0}^{\infty} d_j^{-n^2(1-1/q)} \right)^{1/n} \\ &\quad \times \left(\sum_{j=0}^{\infty} \big(\|\tilde{u}\|_{\tilde{q},B_j} + \|\bar{p}\|_{W^{1,n}(B_j)'} \big)^{\tilde{q}} \right)^{1/\tilde{q}} \\ &\le C\rho^{-n(1-1/q)}(q-1)^{-1} \big(\|\tilde{u}\|_{\tilde{q}} + \|\tilde{p}\|_{W^{1,n}(\Omega)'/\mathbf{R}} \big) \\ &\le C\rho^{-n(1-1/q)}(q-1)^{-1} \big(\|\tilde{f}\|_{-2,\tilde{q}} + \|\tilde{g}\|_{W^{1,n}(\Omega)'} \big),\end{aligned} \tag{3.21}$$

since $\bar{p} = p + c$ for arbitrary $c \in \mathbf{R}$ we may take the infimum over all c.

For $\tilde{f} = D_i \delta e_j$ and since $\|D_i \delta e_j\|_{-2,\tilde{q}} \le C \|D_i \delta e_j\|_{-2,nq/(n-q)}$ we obtain by Lemma 3.1,

$$\|D_i \delta e_j\|_{-2,\tilde{q}} \le C(q-1)^{-1+1/n} \|\delta\|_q \le C\rho^{-n(1-1/n)}(q-1)^{-1+1/q}, \tag{3.22}$$

For $\tilde{g} = \delta - |\Omega|^{-1}$ we note that $(\delta - |\Omega|^{-1}, \varphi) = (\delta, \varphi - \varphi_0)$ where $\varphi_0 = |\Omega|^{-1} \int_\Omega \varphi \, dx$. Using Sobolev's inequality as in the proof of Lemma 3.1

$$\|\delta - |\Omega|^{-1}\|_{W^{1,n}(\Omega)'} \le C\rho^{-n(1-1/n)}(q-1)^{-1+1/n}. \tag{3.23}$$

Collecting the results in (3.19)–(3.23) concludes the proof. □

Corollary 3.3. *Let $\Omega \subset \mathbf{R}^n$, $n = 2, 3$, be a polyhedral domain such that the solution to (1.4) with data as in Theorem 1.3 is continuous in the sense that $(\tilde{u}, \tilde{p}) \in \mathcal{W}^{2,q}$ for some $q > n$. Then there is a constant C such that the solution, $(\tilde{u}, \tilde{p})$ to (1.4) with $(\tilde{f}, \tilde{g}) = (D_i \delta e_j, 0)$ or $(\tilde{f}, \tilde{g}) = (0, \delta - |\Omega|^{-1})$ satisfies the inequality,*

$$\|\tilde{u}\|_{1,1} + \|\tilde{p}\|_{L^1/\mathbf{R}} \le C|\log \rho|^{2-1/n}.$$

Proof. By Hölder's inequality,

$$\|\tilde{u}\|_{1,1} + \|\tilde{p}\|_{L^1/\mathbf{R}} \le |\Omega|^{1/q'} \big(\|\tilde{u}\|_{1,q} + \|\tilde{p}\|_{L^q/\mathbf{R}} \big).$$

Thus, with Theorem 3.2, taking $q - 1 = 1/|\log \rho|$, we finish the proof. □

3.2. $\mathcal{W}^{2,q}$-estimates as $q \downarrow 1$.

Theorem 3.4. *Let $\Omega \subset \mathbf{R}^n$, $n = 2, 3$, be a polyhedral domain. Then for $q \in (1, 4/3]$ there is a constant C such that the solution $(\tilde{u}, \tilde{p})$ to (1.4) with $(\tilde{f}, \tilde{g}) = (\delta e_i, 0)$ satisfies the inequality*

$$\|\tilde{u}\|_{2,q} + \|\tilde{p}\|_{W^{1,q}/\mathbf{R}} \leq C(q-1)^{-\alpha_n} \rho^{-2(n+1)(1-1/q)}$$

where $\alpha_2 = 2$, $\alpha_3 = 4/3$.

Proof. We proceed as in the proof of Theorem 1.1. Let A_j, B_j and η_j be as in (3.7)–(3.10). Let $\bar{p} = \tilde{p} + c$ for a fixed $c \in \mathbf{R}$. Choose a fixed value $q_0 \in (1, 4/3]$. Then for $1 < q < q_0$ by Hölder's inequality

$$\begin{aligned}
\|\tilde{u}\|_{2,q} + \|\tilde{p}\|_{W^{1,q}/\mathbf{R}} &\leq \sum_{j=0}^{\infty} \big(\|\tilde{u}\|_{2,q,A_j} + \|\bar{p}\|_{1,q,A_j} \big) \\
&\leq \sum_{j=0}^{\infty} \big(\|\eta_j \tilde{u}\|_{2,q,B_j} + \|\eta_j \bar{p}\|_{1,q,B_j} \big) \qquad (3.24) \\
&\leq \sum_{j=0}^{\infty} |B_j|^{1/q-1/q_0} \big(\|\eta_j \tilde{u}\|_{2,q_0,B_j} + \|\eta_j \bar{p}\|_{1,q_0,B_j} \big).
\end{aligned}$$

We note that $\eta_j \tilde{u}$ and $\eta_j \bar{p}$ satisfy (1.4) in Ω with $\tilde{f} = \tilde{f}_j = \Delta(\eta_j \tilde{u}) + \nabla(\eta_j \bar{p})$ and $\tilde{g} = \tilde{g}_j = \nabla \cdot (\eta_j \tilde{u})$, where $\tilde{f}_j$ and $\tilde{g}_j$ vanish outside B_j for each j. Hence, for each term in (3.24) we can apply Theorem 1.3,

$$\begin{aligned}
\|\eta_j \tilde{u}\|_{2,q_0,B_j} + \|\eta_j \bar{p}\|_{1,q_0,B_j} &= \|\eta_j \tilde{u}\|_{2,q_0,\Omega} + \|\eta_j \bar{p}\|_{1,q_0,\Omega} \\
&\leq C \big(\|\Delta(\eta_j \tilde{u}) + \nabla(\eta_j \bar{p})\|_{q_0,B_j} + |\nabla \cdot (\eta_j \tilde{u})|_{1,q_0,B_j} \big) \\
&\leq C \big(\|\eta_j(\Delta \tilde{u} + \nabla \bar{p}) + 2\nabla \eta_j \cdot \nabla \tilde{u} + \Delta \eta_j \tilde{u} + \nabla \eta_j \bar{p}\|_{q_0,B_j} \\
&\quad + |\nabla \eta_j \cdot \tilde{u} + \eta_j \nabla \cdot \tilde{u}|_{1,q_0,B_j} \big) \qquad (3.25) \\
&\leq C_I \|\eta_j \delta e_i\|_{q_0,B_j} + C_{II} \|\nabla^2 \eta_j \tilde{u}\|_{q_0,B_j} \\
&\quad + C_{III} \big(\|\nabla \eta_j \cdot \nabla \tilde{u}\|_{q_0,B_j} + \|\nabla \eta_j \bar{p}\|_{q_0,B_j} \big) \\
&= I_j + II_j + III_j,
\end{aligned}$$

where $C = C(n, q_0, \Omega)$ and with $-\Delta \tilde{u} - \nabla \bar{p} = \delta e_i$ and $\nabla \cdot \tilde{u} = 0$, and where we also used $|\nabla \eta_j \cdot \tilde{u}|_{1,q_0,B_j} \leq \|\nabla^2 \eta_j \tilde{u}\|_{q_0,B_j} + \|\nabla \eta_j \cdot \nabla \tilde{u}\|_{q_0,B_j}$.

With (3.25) we now estimate (3.24) in three steps. Recall (3.9) that will repeatedly be used in the estimates below.

I. Since $\operatorname{supp}(\delta) \cap B_j = \emptyset$ for $j \geq 1$ and with (2.11)

$$\sum_{j=0}^{\infty} |B_j|^{1/q-1/q_0} I_j \leq C\rho^{n/q-n/q_0} \|\delta\|_{q_0} \leq C\rho^{-n(1-1/q)}. \tag{3.26}$$

II. By Hölder's inequality with exponent $\tilde{q} = q/(q-2/n)$ and s such that $1/q_0 = 1/s + 1/\tilde{q}$ and with (3.10)

$$|B_j|^{1/q-1/q_0} II_j \leq C d_j^{n/q-n/q_0} |\eta_j|_{2,s,B_j} \|\tilde{u}\|_{\tilde{q},B_j} \leq C d_j^{-(n+2)(1-1/q)} \|\tilde{u}\|_{\tilde{q},B_j},$$

where we used $n/q - n/q_0 + n/s - 2 = -(n+2)(1-1/q)$.

Adding all the terms and by Hölder's inequality in the sum with exponent $\tilde{q}$, with conjugate exponent $\tilde{q}' = nq/2$, and estimating the geometric sum as in (3.11)

$$\begin{aligned} &\sum_{j=0}^{\infty} |B_j|^{1/q-1/q_0} II_j \\ &\quad \leq C \left(\sum_{j=0}^{\infty} d_j^{-(n+2)(1-1/q)nq/2} \right)^{2/nq} \left(\sum_{j=0}^{\infty} \|\tilde{u}\|_{\tilde{q},B_j}^{\tilde{q}} \right)^{1/\tilde{q}} \\ &\quad \leq C\rho^{-(n+2)(1-1/q)} (q-1)^{-2/nq} \|\tilde{u}\|_{\tilde{q}}. \end{aligned} \tag{3.27}$$

With (3.3), Hölder's inequality ($n\tilde{q}/(n+\tilde{q}) \leq nq/(n-q)$), Theorem 1.1, Lemma 3.1 and (2.11)

$$\begin{aligned} \|\tilde{u}\|_{\tilde{q}} \leq& C\tilde{q}^{1-1/n} \|\tilde{u}\|_{1,n\tilde{q}/(n+\tilde{q})} \\ \leq& C\tilde{q}^{1-1/n} \|\delta\|_{-1,nq/(n-q)} \\ \leq& C\tilde{q}^{1-1/n} (1-q)^{-1+1/n} \|\delta\|_q \\ \leq& C\rho^{-n(1-1/q)} (q-2/n)^{-1+1/n} (1-q)^{-1+1/n}, \end{aligned} \tag{3.28}$$

where we remark that

$$2n/(n+1) \leq nq/(n-q) \leq 2n/(n-1),$$

for $n = 2, 3$ and $1 < q < 4/3$ and thus we may use Theorem 1.1.

Collecting the estimates in (3.27) and (3.28) we obtain

$$\sum_{j=0}^{\infty} |B_j|^{1/q-1/q_0} II_j \leq C\rho^{-2(n+1)(1-1/q)} (q-2/n)^{-1+1/n} (q-1)^{-1-1/n}. \tag{3.29}$$

III. By Hölder's inequality with exponent $\tilde{q} = n/(n-1)$ and s such that $1/q_0 = 1/s + 1/\tilde{q}$

$$\begin{aligned} |B_j|^{1/q-1/q_0} III_j &\le C d_j^{n/q-n/q_0} |\eta_j|_{1,s,B_j} \big(\|\tilde{u}\|_{1,\tilde{q},B_j} + \|\bar{p}\|_{\tilde{q},B_j} \big) \\ &\le C d_j^{-n(1-1/q)} \big(\|\tilde{u}\|_{1,\tilde{q},B_j} + \|\bar{p}\|_{\tilde{q},B_j} \big), \end{aligned}$$

where we used $n/q - n/q_0 + n/s - 1 = -n(1-1/q)$.

Adding all the terms and by Hölder's inequality in the sum with exponent $\tilde{q}$, with conjugate exponent $\tilde{q}' = n$, and estimating the geometric sum as in (3.11)

$$\begin{aligned} &\sum_{j=0}^{\infty} |B_j|^{1/q-1/q_0} III_j \\ &\le C \left(\sum_{j=0}^{\infty} d_j^{-n^2(1-1/q)} \right)^{1/n} \left(\sum_{j=0}^{\infty} \big(\|\tilde{u}\|_{1,\tilde{q},B_j} + \|\bar{p}\|_{L^{\tilde{q}}(B_j)} \big)^{\tilde{q}} \right)^{1/\tilde{q}} \\ &\le C \rho^{-n(1-1/q)} (q-1)^{-1/n} \big(\|\tilde{u}\|_{1,\tilde{q}} + \|\tilde{p}\|_{L^{\tilde{q}}/\mathbf{R}} \big), \end{aligned} \tag{3.30}$$

since $\bar{p} = p + c$ for arbitrary $c \in \mathbf{R}$ we may take the infimum of all c.

With Theorem 1.1, Hölder's inequality ($\tilde{q} \le nq/(n-q)$), Lemma 3.1 and (2.11)

$$\begin{aligned} \|\tilde{u}\|_{1,\tilde{q}} + \|\tilde{p}\|_{L^{\tilde{q}}/\mathbf{R}} &\le C \|\delta\|_{-1,\tilde{q}} \\ &\le C (1-q)^{-1+1/n} \|\delta\|_q \\ &\le C \rho^{-n(1-1/q)} (1-q)^{-1+1/n}, \end{aligned} \tag{3.31}$$

where Theorem 1.1 is applicable in analogy to the remark at (3.28).

Collecting the estimates in (3.30) and (3.31) we obtain

$$\sum_{j=0}^{\infty} |B_j|^{1/q-1/q_0} III_j \le C \rho^{-2n(1-1/q)} (q-1)^{-1}. \tag{3.32}$$

Finally adding (3.26), (3.29) and (3.32) concludes the proof. □

Corollary 3.5. *Let $\Omega \subset \mathbf{R}^n$, $n = 2, 3$, be a polyhedral domain. Then there is a constant C such that the solution, $(\tilde{u}, \tilde{p})$ to (1.4) with $\tilde{f} = \delta e_i$ and $\tilde{g} = 0$ satisfies the inequality,*

$$\|\tilde{u}\|_{2,1} + \|\tilde{p}\|_{W^{1,1}/\mathbf{R}} \le C |\log \rho|^{\alpha_n},$$

with α_n as in Theorem 3.4.

Proof. See the proof of Corollary 3.3. □

4. Main results

We now make a precise statement of the main results and begin with the pointwise error estimate of the velocity field.

Theorem 4.1. *Let $\Omega \subset \mathbf{R}^n$, $n = 2, 3$, be a polyhedral domain. Suppose the data to (1.1) is as in Theorem 1.1 for some $q > n$. Then the error e_u in the finite element solution to (1.13) satisfies*

$$\|e_u\|_\infty \leq C|\log h_{\min}|^{\alpha_n}\eta_{2,\infty} + C_1 h_{\min}^\beta,$$

where $\alpha_2 = 2$, $\alpha_3 = 4/3$ and with $\eta_{2,\infty}$ as in Lemma 2.1 and where β can be chosen arbitrarily large.

Proof. Let $x_0 \in \Omega$ and i be such that $\|e_u\|_{L^\infty} = |e_{u_i}(x_0)|$ and let $(\tilde{u}, \tilde{p})$ be the solution to (1.4) with data $\tilde{f} = \delta e_i$ and $\tilde{g} = 0$. With Lemma 2.2, the identity (2.8), Lemma 2.1 with $q = \infty$, and Corollary 3.5, we obtain

$$\begin{aligned}
\|e_u\|_\infty &\leq (e_u, \delta e_i) + C_1 h_{\min}^\beta \big(\|f\|_{-1,q} + \|g\|_q\big) \\
&\leq |\mathcal{R}((u_h, p_h), (\tilde{u}, \tilde{p}))| + C_1 h_{\min}^\beta \\
&\leq C\eta_{2,\infty}\big(\|\tilde{u}\|_{2,1} + \|\tilde{p}\|_{W^{1,1}/\mathbf{R}}\big) + C_1 h_{\min}^\beta \\
&\leq C|\log \rho|^{\alpha_n}\eta_{2,\infty} + C_1 h_{\min}^\beta.
\end{aligned}$$

Choosing $\rho = h_{\min}^\sigma$ for σ sufficiently large such that β becomes large as in Lemma 2.2 concludes the proof. □

For the gradient of the velocity field and the pressure we only obtain pointwise error estimates on a restricted class of polyhedral domains, namely convex domains when $n = 2$ and under an inner angle condition when $n = 3$, see Remark 1.2.

Theorem 4.2. *Let $\Omega \subset \mathbf{R}^n$, $n = 2, 3$, be a polyhedral domain such that the solution to (1.1) with data as in Theorem 1.3 is continuous in the sense that $(u, p) \in \mathcal{W}^{2,q}$ for some $q > n$. Then the error ∇e_u in the finite element solution to (1.13) satisfies*

$$\|\nabla e_u\|_\infty \leq C|\log h_{\min}|^{2-1/n}\eta_{1,\infty} + C_1 h_{\min}^\beta,$$

with $\eta_{1,\infty}$ as in Lemma 2.1 and where β can be chosen arbitrarily large.

Proof. Let $x_0 \in \Omega$, i and j be such that $\|\nabla e_u\|_\infty = |D_i e_{u_j}(x_0)|$ and let $(\tilde{u}, \tilde{p})$ be the solution to (1.4) with data $\tilde{f} = D_i \delta e_j$ and $\tilde{g} = 0$. With Lemma 2.3, the identity (2.8), Lemma 2.1 with $q = \infty$, and Corollary 3.3, we obtain

$$\begin{aligned}
\|\nabla e_u\|_\infty &\leq (e_u, D_i \delta e_j) + C_1 h_{\min}^\beta \big(\|f\|_q + |g|_{1,q}\big) + C_2 \max_{T \in \mathcal{T}} \|[\partial_\nu u_h]\|_{\infty, \partial T \setminus \partial\Omega} \\
&\leq |\mathcal{R}((u_h, p_h), (\tilde{u}, \tilde{p}))| + C_1 h_{\min}^\beta + C_2 \max_{T \in \mathcal{T}} \|[\partial_\nu u_h]\|_{\infty, \partial T \setminus \partial\Omega} \\
&\leq C \eta_{1,\infty} \big(\|\tilde{u}\|_{1,1} + \|\tilde{p}\|_{L^1/\mathbf{R}}\big) + C_1 h_{\min}^\beta + C_2 \max_{T \in \mathcal{T}} \|[\partial_\nu u_h]\|_{\infty, \partial T \setminus \partial\Omega} \\
&\leq C |\log \rho|^{2-1/n} \eta_{1,\infty} + C_1 h_{\min}^\beta.
\end{aligned}$$

Note that the jump term $[\partial_\nu u_h]$ from Lemma 2.3 is incorporated into the error estimator $\eta_{1,\infty}$ in Lemma 2.1.

Choosing $\rho = h_{\min}^\sigma$ for σ sufficiently large such that β becomes large as in Lemma 2.3 concludes the proof. □

Theorem 4.3. *Let $\Omega \subset \mathbf{R}^n$, $n = 2, 3$, be a polyhedral domain such that the solution to (1.1) with data as in Theorem 1.3 is continuous in the sense that $(u, p) \in \mathcal{W}^{2,q}$ for some $q > n$. Then the error e_p in the finite element solution to (1.13) satisfies*

$$\|e_p\|_\infty \leq C |\log h_{\min}|^{2-1/n} \eta_{1,\infty} + C_1 h_{\min}^\beta,$$

with $\eta_{1,\infty}$ as in Lemma 2.1 and where β can be chosen arbitrarily large.

Proof. Let $x_0 \in \Omega$ be such that $|e_p(x_0)| = \|e_p\|_{L^\infty}$ and let $(\tilde{u}, \tilde{p})$ be the solution to (1.4) with data $\tilde{f} = 0$ and $\tilde{g} = \delta - |\Omega|^{-1}$. With Lemma 2.4, the identity (2.8) and choosing e_p such that $\int_\Omega e_p \, dx = 0$, Lemma 2.1 with $q = \infty$, and Corollary 3.3, we obtain

$$\begin{aligned}
\|e_p\|_\infty &\leq (e_p, \delta) + C_1 h_{\min}^\beta \big(\|f\|_q + |g|_{1,q}\big) \\
&\leq |\mathcal{R}((u_h, p_h), (\tilde{u}, \tilde{p}))| + C_1 h_{\min}^\beta \\
&\leq C \eta_{1,\infty} \big(\|\tilde{u}\|_{1,1} + \|\tilde{p}\|_{L^1/\mathbf{R}}\big) + C_1 h_{\min}^\beta \\
&\leq C |\log \rho|^{2-1/n} \eta_{1,\infty} + C_1 h_{\min}^\beta.
\end{aligned}$$

Choosing $\rho = h_{\min}^\sigma$ for σ sufficiently large, such that β becomes large as in Lemma 2.4 concludes the proof. □

Finally we obtain L^q-estimates of the velocity gradient and the pressure.

Theorem 4.4. *Let $\Omega \subset \mathbf{R}^n$, $n = 2, 3$, be a polyhedral domain. Suppose the data to (1.1) is as in Theorem 1.1 for some $2n/(n+1) \le q \le 2n/(n-1)$. Then the error (e_u, e_p) in the finite element solution to (1.13) satisfies*

$$\|e_u\|_{1,q} + \|e_p\|_{L^q/\mathbf{R}} \le C\eta_{1,q},$$

where $\eta_{1,q}$ is as in lemma 2.1.

Proof. With Corollary 1.2, the identity (2.4), and Lemma 2.1 we get

$$\begin{aligned}
\|(e_u, e_p)\|_{\mathcal{W}^q} &\le C \sup_{(\phi,\lambda)\in\mathcal{W}^{q'}} \frac{|\mathcal{L}((e_u, e_p), (\phi, \lambda))|}{\|(\phi, \lambda)\|_{\mathcal{W}^{q'}}} \\
&= C \sup_{(\phi,\lambda)\in\mathcal{W}^{q'}} \frac{|\mathcal{R}((u_h, u_h), (\phi, \lambda))|}{\|(\phi, \lambda)\|_{\mathcal{W}^{q'}}} \\
&\le C\eta_{1,q} \sup_{(\phi,\lambda)\in\mathcal{W}^{q'}} \frac{\|\phi\|_{1,q'} + \|\lambda\|_{L^{q'}/\mathbf{R}}}{\|(\phi, \lambda)\|_{\mathcal{W}^{q'}}} \\
&\le C\eta_{1,q}.
\end{aligned}$$

□

References

[1] R. A. Adams, *Sobolev Spaces*, Academic Press, 1975.

[2] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, second ed., Springer-Verlag, 2002.

[3] R. M. Brown and Z. Shen, *Estimates for the Stokes operator in Lipschitz domains*, Indiana Univ. Math. J. **44** (1995), 1183–1206.

[4] E. Dari, R. G. Durán, and C. Padra, *Maximum norm error estimators for three-dimensional elliptic problems*, SIAM J. Numer. Anal. **37** (2000), 683–700.

[5] R. G. Durán and R. H. Nochetto, *Weighted inf-sup condition and pointwise error estimates for the Stokes problem*, Math. Comp. **54** (1990), 63–79.

[6] G. B. Folland, *Real Analysis*, second ed., John Wiley & Sons Inc., 1999.

[7] D. A. French, S. Larsson, and R. H. Nochetto, *Pointwise a posteriori error analysis for an adaptive penalty finite element method for the obstacle problem*, Comput. Methods Appl. Math. **1** (2001), 18–38.

[8] G. P. Galdi, *An Introduction to the Mathematical Theory of the Navier-Stokes Equations. Vol. I*, Springer-Verlag, 1994.

[9] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, Springer-Verlag, 1977.

[10] V. Girault, R. H. Nochetto, and R. Scott, *Maximum-norm stability of the finite element Stokes projection*, J. Math. Pures Appl. (9) **84** (2005), 279–330.

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

[12] V. A. Kozlov, V. G. Maz′ya, and J. Rossmann, *Spectral Problems Associated with Corner Singularities of Solutions to Elliptic Equations*, American Mathematical Society, 2001.
[13] V. G. Maz'ya and J. Rossmann, *Lp estimates of solutions to mixed boundary value problems for the Stokes system in polyhedral domains*, ArXiv Mathematical Physics e-prints (2004).
[14] R H. Nochetto, *Pointwise a posteriori error estimates for elliptic problems on highly graded meshes*, Math. Comp. **64** (1995), 1–22.
[15] J. Rossmann, *Private communication*, 2006.
[16] L. R. Scott and S. Zhang, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp. **54** (1990), 483–493.
[17] Z. Shen, *Private communication*, 2005.
[18] H. Sohr, *The Navier-Stokes Equations*, Birkhäuser Verlag, 2001.
[19] R. Verfürth, *A posteriori error estimators for the Stokes equations*, Numer. Math. **55** (1989), 309–325.

Department of Mathematical Sciences, Chalmers University of Technology, SE-412 96 Göteborg, Sweden

E-mail address: erik.svensson@chalmers.se, stig@chalmers.se

Paper III

COMPUTATIONAL CHARACTERIZATION OF FLOWS WITH SOME HYPERBOLICITY

ERIK D. SVENSSON

Abstract. Studying flows in general we do not know if the flow is hyperbolic in a strict sense. Instead we vaguely assume that the flow is dominated by contractions and expansions and say that flow have some hyperbolicity. We compare a posteriori and shadowing error estimates for computed orbits in flows with some hyperbolicity. Principal to the estimates are the stability factors which we estimate in two examples for orbits generated by velocity fields modelled by the Stokes equations and computed by a finite element method.

1. Introduction

We consider domains $\Omega \subseteq \mathbf{R}^3$ and Lipschitz continuous vector fields $\Omega \ni x \mapsto f(x) \in \mathbf{R}^3$ so that the dynamical system

$$\partial_t u(t,x) = f(u(t,x)), \quad t > 0; \quad u(0,x) = x, \tag{1.1}$$

defines a flow $(t,x) \mapsto u(t,x) \in \Omega$ describing the motion of a fluid particle starting at x and moving in the velocity field f.

Generally we can not find a closed expression for the flow and in order to study the properties of the flow we may instead analyse a limited number of numerically computed of orbits $u_k(t,x_i)$ for $i = 1,2,\ldots,I$, where k refers to the time discretization. For a reliable analysis we will have to control the error

$$e(t,x) := u_k(t,x_i) - u(t,x), \tag{1.2}$$

and make is small. From now on we consider a fixed x and set $e(t) = e(t,x)$. We are lead to the following classic question. Given a dynamical system (1.1) and a number $\mathrm{Tol} > 0$, is there a threshold time T so that $\|e(t)\| \leq \mathrm{Tol}$ for all $t \in [0,T]$, i.e., so the error is uniformly bounded on $[0,T]$?

Date: April 19, 2006.

2000 *Mathematics Subject Classification.* 37A25, 37C50, 76M10.

Key words and phrases. shadowing, finite elements, flow simulation.

For dynamical systems that are dynamically unstable, that is, sensitive to perturbations, we anticipate that the error will grow, possibly at an exponential rate, and we will only expect to be able to compute $u_k(t, x_i)$ with a small error for small T. However, if Ω is uniformly hyperbolic for (1.1) and $u_k(t, x_i)$ is computed with sufficient accuracy there is a shadow orbit $u(t, y)$ such that $\|u_k(t, x_i) - u(t, y)\| < \text{Tol}$ for arbitrary t [16].

However, in practice we probably do not know if Ω is uniformly hyperbolic for (1.1) and also this requirement seems to be too strong and mainly of theoretical interest. If we instead alleviate on the uniform hyperbolicity and require Ω to have some hyperbolicity meaning that the flow is dominated by contractions and expansions in a less strict sense we may still obtain shadowing results similar to the aforementioned. In this case we will expect the shadowing to hold for finite but large t, see for example [5, 11, 12, 19] and the book [16].

As a concrete example we consider the Lorenz system

$$\begin{aligned} \partial_t u &= (\sigma(u_2 - u_1), \rho u_1 - u_2 - u_1 u_3, u_1 u_2 - \beta u_3), \quad t > 0; \\ u(0) &= (1, 0, 0); \quad \text{for} \quad (\sigma, \rho, \beta) = (10, 28, 8/3). \end{aligned}$$

In [14] this problem was solved accurately, in the sense that $\|e(T)\|$ is small, up to $T = 50$ which is predicted to be the threshold beyond which $\|e(T)\|$ becomes too large to be represented with double precision arithmetics (from the same work $T = 100$ for quadruple precision is predicted).

This result should be compared to [5] where the same problem is solved accurately up to $T = 9 \times 10^6$ in the sense that $\|u_k(t, u(0)) - u(t, y)\|$ is small for $t \in [0, T]$, that is, very close to the computed orbit $u_k(t, u(0))$ there is an exact orbit $u(t, y)$.

This example obviously suggest that long time error control for problems that are dynamically unstable will fail with the first method but could possibly be archived with the last method, provided the structure of the problem is sufficiently hyperbolic-like.

1.1. **About this work.** In this work we consider the case when the vector field f is not given in closed form but rather defined by a model, e.g., a partial differential equation, and approximated by computed numerical data f_h, where h refers to the space discretization. We solve (1.1) numerically with f_h as right hand side and estimate the error (1.2), where we now also have to take the error in the velocity field

$$e_f := f_h - f \tag{1.3}$$

into account.

We assume that u_k and f_h are finite element approximations obtained by solving appropriate finite element problems, which depend on the choice of finite element method and the type of model defining f.

Provided e_f is small enough and that we solve u_k accurately enough we have the following a posterior error estimate, see for example [6],

$$\sup_{t\in[0,T]} \|e(t)\| \leq S(T)\mathcal{E}(f_h, f, x), \tag{1.4}$$

where $S(T)$ is a stability factor and $\mathcal{E}(f_h, f, x)$ is a function depending on the data and made small as e_f is made small and is u_k solved more accurately. The dependence on initial data for a particular problem will be reflected in the stability factor and for dynamically unstable problems this factor may grow exponentially in T, rendering the estimate useless after some rather small time.

If we in addition to the requirements on e_f and u_k made for the estimate above also require that the flow (1.1) is sufficient hyperbolic then we have the following shadowing error estimate, see for example [5],

$$\sup_{t\in[0,T]} \|u_k(t, x) - u(t, y)\| \leq \widetilde{S}(T)\mathcal{E}_1(f_h, f, x), \tag{1.5}$$

where $u(t, y)$ is an exact solution to (1.1) with different initial data, $\widetilde{S}(T)$ a stability factor and $\mathcal{E}(f_h, f, x)$ is the same function as in the a posteriori estimate above. Depending on the contractive and expansive directions in the flow the stability factor may be subject to a mild growth over time and the estimate will be valid for a rather large time.

In the present work we derive the finite time shadowing error estimate (1.5). The overall idea is from [5] but now expressed using a finite element framework. This work also differs in the way we estimate $\widetilde{S}(T)$ and that we use numeric data f_h in the right hand side to (1.1). We also remark that the overall framework in this paper has been inspired by [13] where shadowing was considered in a more abstract setting, for parabolic partial differential equations.

Finally, we describe a numerical experiment where we obtain f_h as the solution to a Stokes flow and with this data we compute and compare $S(T)$ and $\widetilde{S}(T)$ in (1.4) and (1.5). The experiment is inspired by the experimental work [18] on a micro fluid mixing devise.

2. Notation and preliminaries

For real valued functions $u, v \in \mathbf{R}^3$ we denote their scalar product by $u \cdot v = u_1 v_1 + u_2 v_2 + u_3 v_3$.

For matrixes A and linear operators L we denote their transpose and adjoint by A^* and L^*. We let I denote the identity matrix or identity operator.

We will use $\|\cdot\|$ to denote the appropriate matrix and vector norms.

We only consider bounded domains $\omega \subseteq \Omega \subset \mathbf{R}^3$ with measure denoted by $|\omega|$, and where Ω is associated with the flow (1.1).

We will denote piecewise smooth functions by C^m and use standard notation for Sobolev spaces $W^{k,q}(\omega)$ and $W_0^{k,q}(\omega)$.

For vector fields

$$\Omega \ni x \mapsto f(x) = (f_1(x), \dots, f_n(x)) \in \mathbf{R}^n$$

we set

$$\nabla u := (D_i u_j)_{i,j=1}^n,$$

where

$$D_i := \frac{\partial}{\partial x_i} \quad i = 1, \dots, n,$$

denote the i:th partial derivative.

Finally, throughout this work we will use C or C_i, $i = 1, 2, \dots$, to denote various constants, not necessarily taking the same value from time to time.

2.1. **Hyperbolic sets.** A compact set $\omega \subset \Omega$ is said, see for example [17, p. 8], to be *uniformly hyperbolic* for the flow $u(t, x)$ if there is a continuous decomposition

$$\mathbf{R}^3 = E^0(x) \oplus E^s(x) \oplus E^u(x) \quad \forall x \in \omega, \tag{2.1}$$

and constants $c > 0$ and $0 < \lambda < 1 < \mu$ such that for each $x \in \omega$,

(1) $E^0(x)$ is the one-dimensional subspace generated by $f(x)$;
(2) $\nabla u(t, x) E^s(x) = E^s(u(t, x))$ and $\nabla u(t, x) E^u(x) = E^u(u(t, x))$;
(3) $\|\nabla u(t, x)\xi\| \leq c\lambda^t \|\xi\|$ for all $\xi \in E^s(x)$ and $t \geq 0$;
(4) $\|(\nabla u(t, x))^{-1}\xi\| \leq c\mu^{-t}\|\xi\|$ for all $\xi \in E^u(x)$ and $t \geq 0$.

We remark that $u(t, x)$ is called an *Anosov flow* if Ω is uniformly hyperbolic for $u(t, x)$.

The requirements in this definition are rather strong and there are many examples of dynamical systems with non-trivial and interesting properties

that do not meet these requirements [20]. We therefore relax this requirement and instead vaguely think of the flow as being dominated by contractive and expansive direction in a less strict sense. An example of one such relaxation is the notion of *nonuniform hyperbolicity*, where loosely speaking, "for every" in the definition is replace by "for almost all", see for example [1, 20]. We will not discuss and specify this in more detail. Instead we consider an example that to some extent motivates the reasoning.

Example 2.1. Suppose f is incompressible, that is,

$$\nabla \cdot f(x) = 0 \quad \forall x \in \Omega.$$

Then for every open set $A \subset \Omega$ the volume of A is preserved in the flow (1.1), that is,

$$|A| = |u(t, A)| \quad \text{for } t > 0$$

see for example [4, p. 10].

Now for a small ball $B(x, \varepsilon)$ at $x \in \Omega$ and with radios ε and for small t we consider the deformation of the ball in the flow, $B(x, \varepsilon) \to u(t, B(x, \varepsilon))$. Since the volume of the ball is preserved and since the flow leaves the ball unchanged in the direction of the flow we are left with two possibilities (1) the ball is contracted and expanded in some directions such that the volume is unchanged and (2) the ball is unchanged.

Consequently, it seems reasonable to assume that in large parts of Ω there is a splitting (2.1). If we assume that the case where the ball is not deformed only happen in isolated points then the splitting (2.1) will exist for almost all $x \in \omega$ but now E^s and E^u must not be continuous.

2.2. **Finite element approximation.** The finite element formulation of (1.1) is derived from the following variational formulation of (1.1). Find $u \in C^1([0, T])^3$ with $u(0, x) = x$ such that

$$\int_0^T \big(\partial_t u - f(u)\big) \cdot v \, dt = 0 \quad \forall v \in C^1([0, T])^3. \tag{2.2}$$

As the functions u and v are replaced by piecewise polynomials we obtain the Galerkin finite element approximation.

For simplicity we only consider continuous finite elements although this work is readily generalized to discontinuous finite elements. Partition $[0, T]$ into intervals $I_i = [t_{i-1}, t_i]$ for $i = 1, 2, \ldots, N$ such that $0 = t_0 < t_1 < \ldots < t_N = T$ and set $k_i = t_{i-1} - t_i$. Let $P_q(I_i)$ denote the polynomials of degree

less or equal to q on I_i and set

$$V_q([0,T]) := \{v \in C^0([0,T]) : v|_{I_i} \in P_q(I_i), \text{ for } i = 1 \dots N\}$$
$$W_q([0,T]) := \{v \in C^0(\bigcup_{i=1}^{N}(t_{i-1}, t_i)) : v|_{I_i} \in P_q(I_i), \text{ for } i = 1 \dots N\}$$

which is the finite element spaces of continuous and discontinuous piecewise polynomials of degree q.

For $q \geq 1$ and from (2.2) we obtain the finite element formulation. Find $u_k \in V_q([0,T])^d$ with $u_k(0, x_i) = x_i$ such that

$$\int_0^T (\partial_t u_k - f(u_k)) \cdot v \, dt = 0 \quad \forall v \in W_{q-1}([0,T])^3, \tag{2.3}$$

where we note that now $v \in W_{q-1}([0,T])^3$. This is the continuous Galerkin method of order q, referred to as the cG(q) method in [7, p. 210].

There are $q+1$ points in the interval I_i, where the piecewise polynomials are evaluated, referred to as local nodes. In the same way there are $N(q+1)-1$ points in the interval $[0,T]$ referred to as global nodes.

We recall the following interpolation estimate, see for example [7, Theorem 5.1, p. 79]. For a smooth function v on I_i let $\mathcal{I}_{i,q} v \in P_q(I_i)$ interpolate v at the local nodes. Then $\mathcal{I}_{i,q} v$ satisfies

$$\|\mathcal{I}_{i,q} v - v\|_{L^\infty(I_i)} \leq C \|k_i^{q+1} D^{q+1} v\|_{L^\infty(I_i)}.$$

In the same way, for a smooth function v on $[0,T]$, let $\mathcal{I}_q v \in W_q([0,T])$ interpolate v at the global nodes. For a global estimate we let $k = k(t)$ denote the piecewise constant function so that $k|_{I_i} = k_i$. Then

$$\|\mathcal{I}_q v - v\|_{L^\infty([0,T])} \leq C \|k^{q+1} D^{q+1} v\|_{L^\infty([0,T])}. \tag{2.4}$$

Finally we recall the following inverse estimate. Let $\mathcal{T}$ be a finite element triangulation of Ω and set $h_T = \operatorname{diam}(T)$ for all $T \in \mathcal{T}$. For any $T \in \mathcal{T}$, let V be a finite-dimensional subspace of $W^{k,q}(T) \cap W^{m,s}(T)$, where $1 \leq q \leq \infty$, $1 \leq s \leq \infty$ and $0 \leq m \leq k$. Then there exists a constant C such that for all $v \in V$

$$\|v\|_{W^{k,q}(T)} \leq C h_T^{m-k+n/q-n/s} \|v\|_{W^{m,s}(T)}, \tag{2.5}$$

see for example [2, Theorem 4.5.3, p. 111].

2.3. **Linearization.** Let $\bar{u} \in C^0([0,T])^3$ and rewrite (1.1) by linearization around $\bar{u}$

$$\partial_t u(t,x) + A(t)u = F(t,u), \tag{2.6}$$

where we define the linear part of f,

$$A(t) := -\nabla f(\bar{u}(t)) \tag{2.7}$$

and the nonlinear part,

$$F(t,u) := f(u) + A(t)u. \tag{2.8}$$

Let $L(t,s)$ for $0 \le s \le t \le T$ be the solution operator to the linearized homogeneous problem

$$\partial_t u + A(t)u = 0, \quad t > s; \quad u(s,x) = x. \tag{2.9}$$

Thus, $u(t,x) = L(t,s)x$ is the solution of (2.9). We note that $L(t,s)$ satisfies the following properties: $L(s,s) = I$ and $L(t,r)L(r,s) = L(t,s)$ for $0 \le s \le r \le t \le T$. Consequently we may regard $L(t,s)$ as the inverse to $L(s,t)$.

For $t \in [0,T]$ we consider the following weak formulation of (2.9). Find $u \in C^1([s,t])^3$ with $u(s,x) = x$ such that

$$\int_s^t (\partial_\tau u + A(\tau)u) \cdot v \, d\tau = 0 \quad \forall v \in C^1([s,t])^3. \tag{2.10}$$

We also introduce the dual problem to (2.10). Find $\varphi \in C^1([s,t])^3$ with $\varphi(t,\psi) = \psi$ such that

$$\int_s^t \phi \cdot (-\partial_\tau \varphi + A^*(\tau)\varphi) \, d\tau = 0 \quad \forall \phi \in C^1([s,t])^3, \tag{2.11}$$

which is the weak formulation of the following problem,

$$-\partial_s \varphi + A^*(s)\varphi = 0, \quad s < t; \quad \varphi(t,\psi) = \psi. \tag{2.12}$$

Let $K(s,t)$ denote the solution operator to (2.12), that is, $\varphi(s) = K(s,t)\psi$. Note that $K(s,t) = L^*(t,s)$ since, by integration by parts in (2.11),

$$\int_s^t (\partial_\tau \phi + A(\tau)\phi) \cdot \varphi \, d\tau = \phi(s^+) \cdot \varphi(s^+) - \phi(t^-) \cdot \varphi(t^-),$$

and thus, with $\phi = u$ in the above identity and $v = \varphi$ in (2.10) we get

$$0 = \int_s^t (\partial_\tau u + A(\tau)u) \cdot \varphi \, d\tau = x \cdot K(s,t)\psi - L(t,s)x \cdot \psi.$$

Finally, we consider (1.1) with $f(x)$ replaced by $f_h(x)$ and linearize around $\bar{u} \in C^0([0,T])^3$:

$$\partial_t u(t,x) + A_h(t)u = F_h(t,u), \tag{2.13}$$

where we define the linear part of f_h,

$$A_h(t) := -\nabla f_h(\bar{u}) \tag{2.14}$$

and the nonlinear part,

$$F_h(t,u) := f_h(u) + A_h(t)u. \tag{2.15}$$

Let $L_h(t,s)$ for $0 \le s \le t \le T$ be the solution operator the the linearized homogeneous problem

$$\partial_t u + A_h(t)u = 0, \quad t > s; \quad u(s,x) = x. \tag{2.16}$$

Thus, $u(t,x) = L_h(t,s)x$ is the solution to (2.16). In analogy to (2.9) there is a weak form and a dual problem to (2.16) with solution operator $L_h^*(t,s)$.

2.4. **Exponential dichotomies.** If Ω is uniformly hyperbolic for $u(t,x)$ then the following definition is meaningful, cf. [13].

Definition 2.1. The solution operator $L(t,s)$ is said to have an exponential dichotomy in the interval $[0,T]$ if there are projections $P(t)$, $t \in [0,T]$ and constants $M \ge 1$, $\beta > 0$ such that, for $0 \le s \le t \le T$,

(1) $L(t,s)P(s) = P(t)L(t,s)$;
(2) $\|L(t,s)P(s)\| \le Me^{-\beta(t-s)}$;
(3) $\|L(s,t)(I-P(t))\| \le Me^{-\beta(t-s)}$.

The range $\mathcal{R}(P(t))$ is called the stable subspace and the complementary space $\mathcal{R}(I-P(t)) = \mathcal{N}(P(t))$ (the null space of $P(t)$) is called the unstable subspace.

If $L(t,s)$ has an exponential dichotomy on the interval $[0,T]$ then for sufficiently smooth f the following boundary value problem is well posed,

$$\begin{aligned} &\partial_t \varphi + A(t)\varphi = f(t), \quad t \in (0,T), \\ &P(0)\varphi(0) = \varphi_0, \quad (I-P(T))\varphi(T) = \varphi_T, \end{aligned} \tag{2.17}$$

where $\varphi_0 \in \mathcal{R}(P(0))$ and $\varphi_T \in \mathcal{R}(I-P(T))$.

The solution is given by

$$\varphi(t) = G(t,0)\varphi_0 - G(t,T)\varphi_T + \int_0^T G(t,s)f(s)\,ds, \tag{2.18}$$

where $G(t,s)$ is the operator

$$G(t,s) = \begin{cases} L(t,s)P(s), & 0 \le s \le t, \\ -L(t,s)(I-P(s)), & t < s \le T. \end{cases} \tag{2.19}$$

This is readily verified by the following calculations. By Duhamel's principle on the interval $(0,t)$

$$\varphi(t) = L(t,0)\varphi_0 + \int_0^t L(t,s)f(s)\,ds,$$

and by Property 1 in Definition 2.1,

$$P(t)\varphi(t) = L(t,0)P(0)\varphi_0 + \int_0^t L(t,s)P(s)f(s)\,ds.$$

In the same way on the interval (t,T)

$$(I-P(T))\varphi(T) = L(T,t)(I-P(t))\varphi(t) + \int_t^T L(T,s)(I-P(s))f(s)\,ds.$$

By applying the operator $L(t,T)$ and rearranging the terms,

$$(I-P(t))\varphi(t) = L(t,T)(I-P(T))\varphi(T) - \int_t^T L(t,s)(I-P(s))f(s)\,ds,$$

since $L(t,T)L(T,s) = L(t,s)$ for $s \le T \le t$. The above result now follows by considering

$$\varphi(t) = P(t)\varphi(t) + (I-P(t))\varphi(t).$$

We also see that the solution satisfies the estimate

$$\sup_{t\in[0,T]} \|\varphi(t)\| \le M\big(\|\varphi_0\| + \|\varphi_T\| + 2\beta^{-1} \sup_{t\in[0,T]} \|f(t)\|\big), \tag{2.20}$$

which follows from Property 2 and 3 in Definition 2.1 and the estimates

$$\|\varphi(t)\| \le \|G(t,0)\|\,\|\varphi_0\| + \|G(t,T)\|\,\|\varphi_T\| + \sup_{t\in[0,T]} |f(t)| \int_0^T |G(t,s)|\,ds,$$

and

$$\int_0^T |G(t,s)|\,ds \le \int_0^T e^{-\beta|t-s|}\,ds \le \frac{2M}{\beta}.$$

Note that with $f(t) = -\psi\delta(t-\tau)$ for some $\psi \in \mathbf{R}^n$ and $\tau \in [0,T]$, where δ is the Dirac distribution, we obtain the estimate

$$\sup_{t\in[0,T]} \|\varphi(t)\| \le M \max\{\|\varphi_0\|, \|\varphi_T\|, \|\psi\|\}. \tag{2.21}$$

3. Error analysis

Subtracting (2.3) from (2.2) we obtain the weak representation of the error $e := u_k(t, x_i) - u(t, x)$. Find $e \in C^1([0,T])^3$ with $e(0,x) = x_i - x$ such that

$$\int_0^T \partial_t e \cdot v \, dt = \int_0^T (f(u) - \partial_t u_h) \cdot v \, dt \quad \forall v \in C^1([0,T])^3. \tag{3.1}$$

With $A_h(t)$ as in (2.14) we linearize around u_k and let

$$f(u) - \partial_t u_k = e_f(u) + A_h(t)e + \eta(u_k, u) + R(u_k),$$

where we define the error in the computed velocity field,

$$e_f(u) := f(u) - f_h(u), \tag{3.2}$$

the non-linear part,

$$\eta(u_k, u) := f_h(u) - f_h(u_k) + A_h(t)e \tag{3.3}$$

and the residual to (2.3),

$$R(u_k) := f_h(u_k) - \partial_t u_k. \tag{3.4}$$

We note that the residual is orthogonal to functions in the finite element space $W_{q-1}([0,T])^3$ in the following sense,

$$\int_0^T R(u_k) \cdot v \, dt = 0 \quad \forall v \in W_{q-1}([0,T])^3. \tag{3.5}$$

We rewrite (3.1) according to the linearization above. Find $e \in C^1([0,T])^3$ with $e(0,x) = x_i - x$ such that

$$\int_0^T (\partial_t e + A_h(t)e) \cdot v \, dt = \int_0^T (e_f(u) + \eta(u_k, u) + R(u_k)) \cdot v \, dt, \tag{3.6}$$

for all $v \in C^1([0,T])^3$.

The following lemma will be useful characterizing the function $\eta(\cdot,\cdot)$. Note that ∇f_h is discontinuous across $\partial T \setminus \partial\Omega$ for $T \in \mathcal{T}$.

Lemma 3.1. *Let $u, v, w \in \Omega$ and suppose the convex hull K of $\{u, v, w\}$ is contained in Ω. Then a finite element function $f_h : \Omega \ni x \mapsto f_h(x) \in \mathbf{R}^3$*

satisfies

$$\begin{aligned}\|f_h(u) - f_h(v) + \nabla f_h(w)(u-v)\| \\ \leq C\|u-v\|\Big(h_{\min}^{-1-n/p}\big(\|u-w\| + \|v-w\|\big)\|\nabla f_h\|_{L^p(\Omega)} \\ + \max_{T\in\mathcal{T}}\|[\nabla f_h]\|_{L^\infty(\partial T\setminus\partial\Omega)}\Big),\end{aligned}$$

for some $1 \leq p \leq \infty$ *and where the constant* C *depends on* $\mathrm{card}(K \cap \mathcal{T})^2$ *and the constant in* (2.5)*, and* $[\cdot]$ *denotes the jump across* ∂T*.*

We remark that the exponent p in practice is determined by available error estimates.

Proof. Consider the line $l : [0,1] \ni s \mapsto su + (1-s)v \in \mathbf{R}^n$ and let

$$l_{\mathcal{T}} = \bigcup_{T\in\mathcal{T}} T \cap l \neq \emptyset.$$

From the identity

$$\begin{aligned}f_h(u) - f_h(v) - \nabla f_h(w)(u-v) \\ = \int_0^1 \big(\nabla f_h(su + (1-s)v) - \nabla f_h(w)\big)(u-v)\, ds,\end{aligned}$$

and by the mean value theorem there are points $\xi_T \in T$ for $T \in l_{\mathcal{T}}$ such that

$$\begin{aligned}\int_0^1 \nabla f_h(su + (1-s)v)\, ds &= \sum_{T\in l_{\mathcal{T}}} \int_{l\cap T} \nabla f_h(su + (1-s)v)\, ds \\ &= \sum_{T\in l_{\mathcal{T}}} \nabla f_h(\xi_T) \int_{l\cap T} ds.\end{aligned}$$

Hence, since $\int_{l\cap T} ds < 1$

$$\|f_h(u) - f_h(v) - \nabla f_h(w)(u-v)\| \leq \|u-v\| \sum_{T\in l_{\mathcal{T}}} \|\nabla(f_h(\xi_T) - f_h(w))\|$$

For each point ξ_T consider the line between ξ_T and w. Suppose this line crosses m_T boundaries ∂T for $T \in \mathcal{T}$ at points $\xi_{T,i}$ for $i = 1, \ldots, m_T$. Let $\xi_{T,i}^-$ and $\xi_{T,i}^+$ be the limits at $\xi_{T,i}$ going from ξ_T and w respectively, and set

$\xi_{T,0}^+ = \xi_T$ and $\xi_{T,m_T+1}^- = w$. Estimate the terms in the sum above

$$\|\nabla(f_h(\xi_T) - f_h(w))\| \le \sum_{i=0}^{m_T} \|\nabla(f_h(\xi_{T,i}^+) - f_h(\xi_{T,i+1}^-))\| + \sum_{i=1}^{m_T} \|[\nabla f_h(\xi_{T,i})]\|,$$

where $[\nabla f_h(\xi_{T,i})] = \nabla(f_h(\xi_{T,i}^-) - f_h(\xi_{T,i}^+))$ denotes the jump at $\xi_{T,i}$.

By the mean value theorem and an inverse estimate

$$\begin{aligned}\|\nabla(f_h(\xi_{T,i}^+) - f_h(\xi_{T,i+1}^-))\| &\le \big(\|u - v\| + \|v - w\|\big)\|\nabla^2 f_h\|_{L^\infty(T)} \\ &\le C h_{\min}^{-1-n/p}\big(\|u - w\| + \|v - w\|\big)\|\nabla f_h\|_{L^p(\Omega)},\end{aligned}$$

since $\|\xi_{T,i}^+ - \xi_{T,i+1}^-\| \le \|\xi_T - w\| \le \|u - v\| + \|v - w\|$.

For the jump terms we estimate

$$\|[\nabla f_h(\xi_{T,i})]\| \le \max_{T\in\mathcal{T}} \|[\nabla f_h]\|_{L^\infty(\partial T\setminus\partial\Omega)}.$$

Collecting the estimates above concludes the proof. □

For fixed u_k we consider $\eta = \eta(u_k, u_k - e)$ and $e_f = e_f(u_k - e)$ as a functions of e. Set

$$\begin{aligned} N_{0,T}(e, v) &:= \int_0^T \eta(u_k, u_k - e)\cdot v \, dt, \\ E_{0,T}(e, v) &:= \int_0^T e_f(u_k - e)\cdot v \, dt, \\ R_{0,T}(u_k, v) &:= \int_0^T R(u_k)\cdot v \, dt, \end{aligned} \tag{3.7}$$

and estimate $N_{0,T}$, $E_{0,T}$ and $R_{0,T}$. Let

$$\mathcal{B}_\rho := \{e \in C^1([0,T]) : \|e\|_{L^\infty([0,T])} \le \rho\}. \tag{3.8}$$

With $u = u_k - e$ and $v = w = u_k$ in Lemma 3.1 we get

$$\|N_{0,T}(e, v)\| \le C\|v\|_{L^1([0,T])} r_N(f_h, \rho)\rho \quad \text{for } e \in \mathcal{B}_\rho \tag{3.9}$$

where we defined

$$r_N(f_h, \rho) := \rho h_{\min}^{-1-n/p}\|\nabla f_h\|_{L^p(\Omega)} + \max_{T\in\mathcal{T}} \|[\nabla f_h]\|_{L^\infty(\partial T\setminus\partial\Omega)}. \tag{3.10}$$

Now $N_{0,T}$ is Lipschitz continuous, that is,

$$\|N_{0,T}(e_1, v) - N_{0,T}(e_2, v)\| \le C\|v\|_{L^1([0,T])} r_N(f_h, \rho)\|e_1 - e_2\|_{L^\infty([0,T])}, \tag{3.11}$$

for $e_1, e_2 \in \mathcal{B}_\rho$ and where $r_N(f_h, \rho)$ is as in (3.10).

To see this, suppose $e_1, e_2 \in \mathcal{B}_\rho$. By Hölder's inequality,

$$|N_{0,T}(e_1, v) - N_{0,T}(e_2, v)| \\ \leq \|\eta(u_k, u_k - e_1) - \eta(u_k, u_k - e_2)\|_{L^\infty([0,T])} \|v\|_{L^1([0,T])},$$

where

$$\eta(u_k, u_k - e_1) - \eta(u_k, u_k - e_2) \\ = f_h(u_k - e_1) - f_h(u_k - e_2) - \nabla f_h(u_k)(e_1 - e_2).$$

With $u = u_k - e_1$, $v = u_k - e_2$ and $w = u_k$ in Lemma 3.1, (3.11) follows.

As for $E_{0,T}$ we will use the uniform estimate

$$E_{0,T}(e, v) \leq C\|e_f\|_{L^\infty(\Omega)} \|v\|_{L^1([0,T])}. \tag{3.12}$$

We also note by taking $u = u_k - e_1$, $v = u_k - e_2$ and $w = 0$ in Lemma 3.1 that $E_{0,T}$ is Lipschitz continuous, that is,

$$\|E_{0,T}(e_1, v) - E_{0,T}(e_2, v)\| \leq C\|v\|_{L^1([0,T])} r_E(e_f) \|e_1 - e_2\|_{L^\infty([0,T])}, \tag{3.13}$$

for $e_1, e_2 \in \mathcal{B}_\rho$ and where $r_E(e_f)$ is defined by

$$r_E(e_f) := h_{\min}^{-n/p} \|\nabla e_f\|_{L^p(\Omega)}. \tag{3.14}$$

Finally, due to the Galerkin orthogonality (3.5) we may add $\mathcal{I}_{q-1} v$

$$\int_0^T R(u_k) \cdot v \, dt = \int_0^T R(u_k) \cdot (v - \mathcal{I}_{q-1} v) \, dt,$$

and hence by (2.4)

$$R_{0,T}(u_k, v) \leq C\|k^q R(u_k)\|_{L^\infty([0,T])} \|D^q v\|_{L^1([0,T])}. \tag{3.15}$$

3.1. **A posteriori error analysis.** Consider the dual problem to (3.6). Find $\varphi \in C^1([0,T])^3$ with $\varphi(T, x) = \varphi_T$ such that

$$\int_0^T \phi \cdot (-\partial_t \varphi + A_h^*(t)\varphi) \, dt = 0 \quad \forall \phi \in C^1([0,T])^3. \tag{3.16}$$

With $v = \varphi$ in (3.6) and $\phi = e$ in (3.16) subtracting the equations we get

$$\int_0^T \partial_t (e \cdot \varphi) \, dt = \int_0^T (e_f(u) + \eta(u_k, u) + R(u_k)) \cdot \varphi \, dt,$$

or with the notation in (3.7) we get

$$e(T) \cdot \varphi_T = e(0) \cdot \varphi(0) + R_{0,T}(u_k, \varphi) + E_{0,T}(e, \varphi) + N_{0,T}(e, \varphi), \tag{3.17}$$

which is a fixed point problem in e that admits a unique solution provided $N_{0,T}$ and $E_{0,T}$ has sufficiently small Lipschitz constants.

Estimating the right hand side in (3.17) we use Cauchy's inequality for the fist term and for the remaining terms we use the estimates (3.15), (3.12) and (3.9). As is usual we define the stability factors

$$\begin{aligned} S_0(T) &:= \|\varphi(0)\|, \\ S_1(T) &:= \|D^q\varphi\|_{L^1([0,T])}, \\ S_2(T) &:= \|\varphi\|_{L^1([0,T])}. \end{aligned} \tag{3.18}$$

We remark that the stability factor mentioned in (1.4) now is $S(T) = \max\{S_0, S_1, S_2\}$.

Theorem 3.2 (A priori error estimate)**.** *Let ρ, f_h and u_k be such that*

$$\begin{aligned} &S_0(T) \le 1/6, \\ &CS_2(T)r_E(e_f) \le 1/6, \\ &CS_2(T)r_N(f_h,\rho) \le 1/6, \end{aligned} \tag{3.19}$$

where C is as in Lemma 3.1, $r_N(f_h,\rho)$ and $r_E(e_f)$ as in (3.10) *and* (3.14)*, and suppose*

$$\begin{aligned} &e(0)\cdot\varphi(0) \le S_0(T)\|e(0)\| \le \frac{1}{6}\rho, \\ &R_{0,T}(u_k,\varphi) \le CS_1(T)\|k^qR(u_k)\|_{L^\infty([0,T])} \le \frac{1}{6}\rho, \\ &E_{0,T}(u,\varphi) \le S_2(T)\|e_f\|_{L^\infty(\Omega)} \le \frac{1}{6}\rho. \end{aligned} \tag{3.20}$$

Then the error $e(T) = u_k(T) - u(T)$ is bounded from above by

$$\begin{aligned} e(T)\cdot\varphi_T \le& S_0(T)\|e(0)\| + S_1(T)\|k^qR(u_k)\|_{L^\infty([0,T])} \\ &+ S_2(T)\|e_f\|_{L^\infty(\Omega)} \le \rho. \end{aligned} \tag{3.21}$$

Proof. From (3.11), (3.13) and (3.19) it follows that (3.17) is a contraction mapping on $\mathcal{B}_\rho$. From (3.19) and (3.20) we also see that the mapping is into $\mathcal{B}_\rho$. Therefore there is a unique solution $e \in \mathcal{B}_\rho$ to (3.17) that satisfies (3.21). □

We note that

$$\|\varphi(T)\| \le \|\varphi(0)\| + \|A_h^*\|_{L^\infty([0,T])}\int_0^T \|\varphi(s)\|\,ds,$$

and by Gronwall's lemma, see for example [8, p. 625] we estimate

$$\|\varphi(T)\| \leq \|\varphi(0)\|\big(1 + T\|A_h^*\|_{L^\infty([0,T])} e^{T\|A_h^*\|_{L^\infty([0,T])}}\big).$$

For flows that are dynamically unstable we do not expect any better estimates than this. Thus (3.19) and (3.20) will be very difficult or impossible to achieve in these situations.

3.2. **Shadowing.** In this section we assume that $L(t,s)$ has an exponential dichotomy on the interval $[0,T]$. We note the connection between $L(t,s)$ and $L_h(t,s)$ provided in the following roughness result. From [16, Lemma 7.4, p.133] we know that if $L(t,s)$ has an exponential dichotomy on $[0,T]$ and if

$$\|A_h(t) - A(t)\| \leq \delta \leq \delta_0(M,\beta).$$

Then $L_h(t,s)$ also has an exponential dichotomy on $[0,T]$ with constants M_h, β_h and projection $P_h(t)$ satisfying

$$0 < \beta_h < \beta \qquad \text{and} \qquad \|P_h(t) - P(t)\| \leq C\delta,$$

where M_h, β_h and C are constants only depending on M and β.

We now assume that $L_h(t,s)$ has an exponential dichotomy on the interval $[0,T]$ in the sense given in the paragraph above. It then follows that $L_h^*(s,t)$ also has an exponential dichotomy on $[0,T]$ with projection $I - P_h^*(t)$ and constants M_h and β_h. By taking the adjoint in Property 1 of Definition 2.1 and subtracting the identity we get

$$(I - P_h^*(s))L_h^*(t,s) = L^*(t,s)(I - P_h^*(t)),$$

and multiplying from left and right with $L_h^*(s,t)$ and $L_h^*(s,t)$ we obtain Property 1 for $L_h^*(s,t)$

$$L_h^*(s,t)(I - P_h^*(s)) = (I - P_h^*(t))L_h^*(s,t).$$

The other properties now follow using the identity above.

Consider the following boundary value problem related to (2.17)

$$\begin{aligned} &-\partial_s\varphi + A_h^*(s)\varphi = -\psi\delta(s-t), \quad s \in ([0,T]); \\ &(I - P_h^*(0))\varphi(0) = 0, \quad P_h^*(T)\varphi(T) = 0, \end{aligned} \tag{3.22}$$

where $\psi \in \mathbf{R}^3$ and δ is the Dirac delta distribution and thus the solution $\varphi(s)$ will have a jump $-\psi = \varphi(t)^+ - \varphi(t)^-$ at time $s = t$.

This problem is also well posed by the same arguments as for (2.17) and the solution is

$$\varphi(s,t) = -G_h^*(s,t)\psi,$$

where we explicitly added t as an argument in the solution and where $G_h^*(s,t)$ now is the Green operator

$$G_h^*(s,t) = \begin{cases} (I - P_h^*(t))L_h^*(s,t), & 0 \le t \le s, \\ -P_h^*(t)L_h^*(s,t), & s < t \le T. \end{cases} \tag{3.23}$$

In weak form (3.22) reads. Find $\varphi \in C^1([0,t))^3 \cup C^1((t,T])^3$:

$$\int_0^T \phi \cdot (-\partial_s \varphi + A_h^*(s)\varphi)\, ds = \phi(t) \cdot \psi \quad \forall \phi \in C^1([0,T])^3, \tag{3.24}$$

and by integration by parts

$$\phi(t) \cdot \psi = \int_0^T (\partial_s \phi + A_h(s)\phi) \cdot \varphi \, ds + \phi(T) \cdot \varphi(T) - \phi(0) \cdot \varphi(0), \tag{3.25}$$

where we stress that $\varphi(0)$ and $\varphi(T)$ are not equal to zero, in fact only $(I - P_h^*(0))\varphi(0) = 0$ and $P_h^*(T)\varphi(T) = 0$ ($P_h^*(0)\varphi(0)$ and $(I - P_h^*(T))\varphi(T)$ are determined by the differential equation).

Suppose $e(t) = u_k(t, x_i) - u(t, y) \in \mathcal{B}_\rho$, where $\mathcal{B}_\rho$ is the ball (3.8), and such that $P_h(0)e(0) = 0$ and $(I - P_h(T)e(T) = 0$ which implies that

$$e(T) \cdot \varphi(T) = e(T) \cdot (I - P_h^*(T))\varphi(T) = (I - P_h(T))e(T) \cdot \varphi(T) = 0,$$

and likewise $e(0) \cdot \varphi(0) = 0$.

Taking $\phi = e$ in (3.25) and with (3.6) and (3.7) we get

$$e(t) \cdot \psi = R_{0,T}(u_k, \varphi) + E_{0,T}(u, \varphi) + N_{0,T}(u, \varphi), \tag{3.26}$$

which is a fixed point problem with a similar right hand side as in (3.17) although the problem defining φ is not the same in this case. Note that the right hand side does not have any derivative in φ and hence is well defined even when φ is discontinuous as in the present case.

Estimating the right hand side in (3.26) we use Cauchy's inequality for the fist two terms and for the remaining terms we use the estimates (3.15) (with care), (3.12) and (3.9), now tanking into account that φ is discontinuous at $s = t$. As is usual we define the stability factors

$$\begin{aligned} \widetilde{S}_1(T) &:= \sup_{t\in[0,T]} \max\big\{ \|D^q\varphi(\cdot,t)\|_{L^1([0,t))}, \|D^q\varphi(\cdot,t)\|_{L^1((t,T])} \big\} \\ \widetilde{S}_2(T) &:= \sup_{t\in[0,T]} \|\varphi(\cdot,t)\|_{L^1([0,T])}, \end{aligned} \tag{3.27}$$

where now φ is the solution to the boundary value problem (3.22). We remark that the stability factor in (1.5) now is $\widetilde{S}(T) = \max\{\widetilde{S}_1, \widetilde{S}_2\}$.

Theorem 3.3 (Shadowing)**.** *Let ρ, f_h and u_k be such that*

$$\begin{aligned} C\widetilde{S}_2(T)r_E(e_f) &\le 1/4, \\ C\widetilde{S}_2(T)r_N(f_h,\rho) &\le 1/4, \end{aligned} \tag{3.28}$$

where C is as in Lemma 3.1, $r_N(f_h,\rho)$ and $r_E(e_f)$ as in (3.10) *and* (3.14)*, and suppose*

$$\begin{aligned} R_{0,T}(u_k,\varphi) &\le C\widetilde{S}_1(T)\|k^q R(u_k)\|_{L^\infty([0,T])} \le \frac{1}{4}\rho, \\ E_{0,T}(u,\varphi) &\le \widetilde{S}_2(T)\|e_f\|_{L^\infty(\Omega)} \le \frac{1}{4}\rho. \end{aligned} \tag{3.29}$$

Then the numerical solution $u_k(t,x_i)$ is shadowed by an exact solution $u(t,y_i)$ and the error $e(t) = u_k(t,x_i) - u(t,y_i)$ is bounded from above for all $t \in [0,T]$

$$|e(t)| \le \widetilde{S}_1(T)\|k^q R(u_k)\|_{L^\infty([0,T])} + \widetilde{S}_2(T)\|e_f\|_{L^\infty(\Omega)} \le \rho. \tag{3.30}$$

Proof. Set $\psi = 1$. From (3.11), (3.13) and (3.28) it follows that (3.26) is a contraction mapping on $\mathcal{B}_\rho$. From (3.28) and (3.29) we also see that the mapping is into $\mathcal{B}_\rho$. Therefore there is a unique solution $e \in \mathcal{B}_\rho$ to (3.26) that satisfies (3.30) and we get $u(t,y_i) = u_k(t,x_i) - e(t)$. □

We note that provided $L_h^*(s,t)$ has an exponential dichotomy φ will stay bounded by (2.21) and in contrast to the error estimate (3.21) the estimate in this case (3.30) will remain valid for large T. However we must show that $L_h^*(s,t)$ has an exponential dichotomy or by some means estimate $\varphi(t,\cdot)$. We discuss this matter in the next section.

3.3. **Finite time shadowing.** In this section we discuss the finite time shadowing results from [5]. We first assume that $L(t,s)$ has an exponential dichotomy as described in Sections 2.4 and 3.2.

We consider the boundary value problem (3.22) and the solution operator (3.23). From now on set $\psi = 1$.

Partition $[0,T]$ into M sub intervals $[T_m, T_{m+1}]$ for $m = 0,1,\dots,M-1$ and where $T_0 = 0$ and $T_M = T$. Let $L_m = L(T_{m+1},T_m)$ be a sequence of operators and set

$$L_{mn} = L_{m-1}\cdots L_n, \quad m > n, \qquad \text{and} \qquad L_{mm} = I.$$

If we choose $s = T_m$ and $t = T_n$ in (3.23) we get

$$\varphi(T_m,T_n) = -\mathcal{G}_{mn}^*, \tag{3.31}$$

where

$$\mathcal{G}^*_{mn} = \begin{cases} (I - P^*(T_n))L^*_{mn}, & 0 \le n \le m, \\ -P^*(T_n)L^*_{mn}, & m < n \le M. \end{cases}$$

This is the solution to the recurrence problem cf. [13, Section 3.2]

$$\begin{aligned} -\delta_{m+1,n} &= \varphi_{m+1} - L^*_m \varphi_m, \quad m = 0, \dots, M-1; \\ (I - P^*(0))\varphi_0 &= 0, \quad P^*(T_M)\varphi_M = 0, \end{aligned} \tag{3.32}$$

for $n \in [0, M-1]$ and where $\delta_{m,n} = 1$ if $m = n$ and $\delta_{m,n} = 0$ if $m \neq n$.

Let $\widehat{f} = f/\|f\|$ denote the normalization of f. Choose one (3×2) matrix Z_0 such that the (3×3) matrix

$$\begin{pmatrix} \widehat{f}_h(u(0,x)) & Z_0 \end{pmatrix}$$

is orthonormal and by QR-factorization define recursively for $m = 0, 1, \dots, M-1$

$$\begin{pmatrix} \widehat{f}_h(u(T_{m+1},x)) & L^*_m Z_m \end{pmatrix} = \begin{pmatrix} \widehat{f}_h(u(T_{m+1},x)) & Z_{m+1} \end{pmatrix} \begin{pmatrix} \cdots & \cdots \\ 0 & A_m \end{pmatrix},$$

where

$$A_m := \begin{pmatrix} a_m & b_m \\ 0 & c_m \end{pmatrix} = Z^*_{m+1} L^*_m Z_m \tag{3.33}$$

is upper triangular and with positive diagonal entries as long as matrix on the left hand side has full rank [10, Theorem 5.2.2, p. 217]. Note that $Z^*_m Z_m = I$.

Set $\varphi_m = Z_m \phi_m$ and transform (3.32)

$$\begin{aligned} -\delta_{m+1,n} Z^*_{m+1} &= \phi_{m+1} - A_m \phi_m, \quad m = 0, \dots, M-1; \\ Z^*_0 (I - P^*(0)) Z_0 \phi_0 &= 0, \quad Z^*_M P^*(T) Z_M \phi_M = 0. \end{aligned} \tag{3.34}$$

In most situations we do not know the projections $P(0)$ and $P(T)$. Nevertheless we may solve (3.34) by taking a good guess. With $\phi_m = (\phi_{m,1}, \phi_{m,2})$ and (3.33) we rewrite (3.34)

$$\begin{aligned} \phi_{m+1,1} &= a_m \phi_{m,1} + b_m \phi_{m,2} + \delta_{m+1,n} z_{m+1,1}, \\ \phi_{m+1,2} &= c_m \phi_{m,2} + \delta_{m+1,n} z_{m+1,2}, \end{aligned} \tag{3.35}$$

where $z_{m+1,i}$ is the sum of the i:th row in Z^*_{m+1}.

Considering the sequences $\{a_m\}_{m=0}^M$ and $\{c_m\}_{m=0}^M$ we distinguish six different cases and solve (3.35) accordingly. Set $a = \prod_{m=0}^M a_m$ and $c = \prod_{m=0}^M c_m$.

(1) If $a > 1.0$ and $c < 1.0$. Set $\phi_{0,2} = 0$ and solve the second equation forwards obtaining $\phi_{m,2}$, and set $\phi_{m,1} = 0$, substitute $\phi_{m,2}$ into the first equation and solve backwards obtaining $\phi_{m,1}$.
(2) If $a < 1.0$ and $c > 1.0$. Set $\phi_{0,2} = 0$ and solve the second equation backwards obtaining $\phi_{m,2}$, and set $\phi_{m,1} = 0$, substitute $\phi_{m,2}$ into the first equation and solve forwards obtaining $\phi_{m,1}$.
(3) If $a < 1.0$ and $c < 1.0$ and $a > c$. Do as in the first case.
(4) If $a < 1.0$ and $c < 1.0$ and $a < c$. Do as in the second case.
(5) If $a > 1.0$ and $c > 1.0$ and $a > c$. Do as in the first case.
(6) If $a > 1.0$ and $c > 1.0$ and $a < c$. Do as in the second case.

Cases (1) and (2) are considered as ideal and imply that $\|\phi\|$ is small. The remanding cases are not ideal and the solution may blow up and $\|\phi\|$ may be large.

Since we only guess the projections we may expect to mix the stable and unstable subspaces when solving according to the steps above. The computed solution will serve as an estimate for the true solution and hopefully this solution will be small or have a mild growth over time.

3.3.1. *Computing $\widetilde{S}_i(T)$, $i = 2, 3$, in practice.* We now substitute $L^*(s,t)$ by $L^*_h(s,t)$ in the analysis above and compute the norm to $\{\phi_m\}_{m=0}^M$ in (3.35) in two different ways.
Case I. In the first case we solve $M-2$ problems (3.35) for $n = 1, 2, \ldots, M-2$ and compute the norms from this set of solutions. The amount of work for this procedure will scale like $O(M^2)$.
Case II. In the second case we proceed as proposed in [5]. Instead of (3.35) we consider

$$\begin{aligned} \eta_{m+1,1} &= a_m \eta_{m,1} \mp |b_m| \eta_{m,2} \mp |z_{n,1}|, \\ \eta_{m+1,2} &= c_m \eta_{m,2} \pm |z_{m,2}|, \end{aligned} \tag{3.36}$$

where the $\mp$ and $\pm$ depend on whether we solve according to case (1) or (2) as described above. This procedure will imply that $|\phi_{m,1}| \leq \eta_{m,1}$ and $|\phi_{m,2}| \leq \eta_{m,2}$. The amount of work for this procedure will scale like $O(M)$.

4. Finite time shadowing in Stokes flow

Inspired by [18] where laminar fluid mixing was experimentally studied in small channels we set up the following model. Let $\Omega \subset \mathbf{R}^3$, be a polyhedral domain with periodic boundaries Γ_A and Γ_B, see Figures 4.1

and 4.2, and consider the Dirichlet Stokes problem with periodic boundary conditions in dimensionless form

$$
\begin{aligned}
-\Delta U + \nabla P &= 0 \quad \text{in } \Omega, \\
\nabla \cdot U &= 0 \quad \text{in } \Omega, \\
U &= 0 \quad \text{on } \partial\Omega \setminus (\Gamma_A \cup \Gamma_B), \\
U|_{\Gamma_A} &= U|_{\Gamma_B}, \\
P|_{\Gamma_A} &= P|_{\Gamma_B} + R,
\end{aligned}
\tag{4.1}
$$

where $U = (U_1, U_2, U_3)$ is the unknown velocity field, P the unknown pressure and R is a constant modelling the pressure drop.

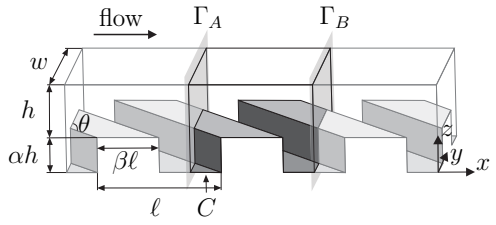


Figure 4.1: Three juxtaposed Ridge Domains. The shaded planes Γ_A and Γ_B are periodic boundaries. We choose the following values for the parameters: $\ell = w = 1$, $h = 0.3$, $\theta = 45°$, $\alpha = 2/3$, $\beta = 0.5$, and the length of the unit cell is $= 1$.

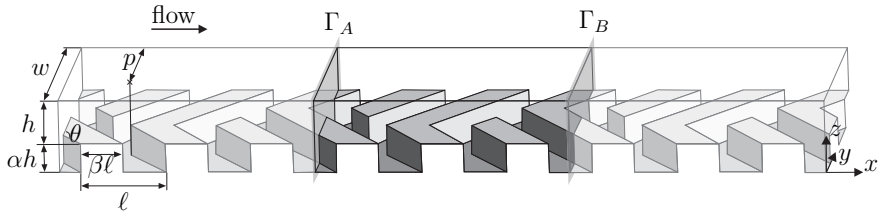


Figure 4.2: Three juxtaposed Herringbone Domains. The shaded planes Γ_A and Γ_B are periodic boundaries. We choose the following values for the parameters: $\ell = 2/3$, $w = 1$, $h = 1/5$, $\theta = 45°$, $\alpha = 2/3$, $\beta = 9/16$, $p = 2/3$, and the length of the unit cell is $= 14/9$.

From [3] and [15] we know that $U \in W^{2,4/3}(\Omega)^3 \cap W_0^{1,3}$ and thus U is continuous although not Lipschitz continuous. There will be singularities in ∇U and P along the edges and vertices of Ω. However, if we let $\Omega' \subset \Omega$ such that $\text{dist}(\Omega', \partial\Omega)$ is not too small, then we may argue that U is Lipschitz continuous in Ω' by an interior estimate as in for example [9, Theorem 4.2, p. 209]. Thus when we compute orbits using $f = U$ (or in practice $f = U_h$) in (1.1) we only consider orbits that are not too close to $\partial\Omega$.

We refer to the domains in Figures 4.1 and 4.2 as Ridge and Herringbone respectively, the names are from [18]. Accurate solutions to (4.1) in the two domains are computed by a finite element method, Hood-Taylor P_2P_1 on fine triangulations. We illustrate the solutions in Figures 4.3 and 4.4.

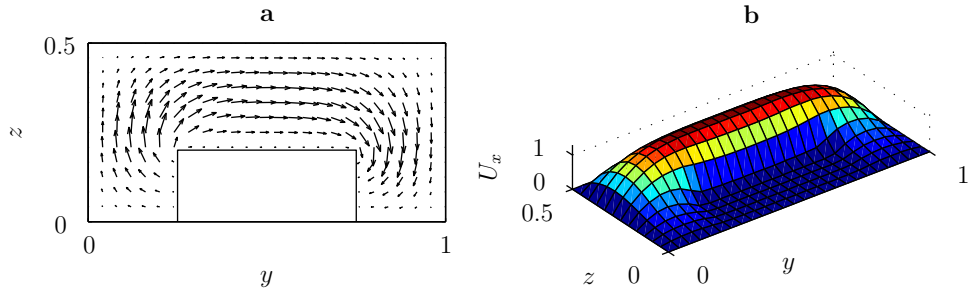


Figure 4.3: Velocity field for (4.1) solved in the Ridge Domain, Figure 4.1, at $x = 0.0$. **(a)** The y and z components of the velocity field. **(b)** The x component of the velocity field.

We compute orbits to (1.1) using the simple cG(1) method described in Section 2.2, with $f = U_h$ where U_h now is the computed solution to (4.1) . The time steps k_i for $i = 1, 2, \ldots, N$ is chosen adaptively so that the local residual is less than a small tolerance, for more details see [7]. We plot two typical orbits in Figure 4.5 for the Ridge Domain and in Figure 4.6 for the Herringbone Domain.

The dual problem (3.16) is solved by the same means but with time steps k_i for $i = 1, 2, \ldots, 2N - 1$ obtained by refining the partition of $[0, T]$ used for computing the orbits to (1.1). As φ_T we choose either of the canonical unit vectors, e.g., $(1, 0, 0)$. The stability factors $S_i(T)$ for $i = 1, 2, 3$ are then readily computed, see Figure 4.7.

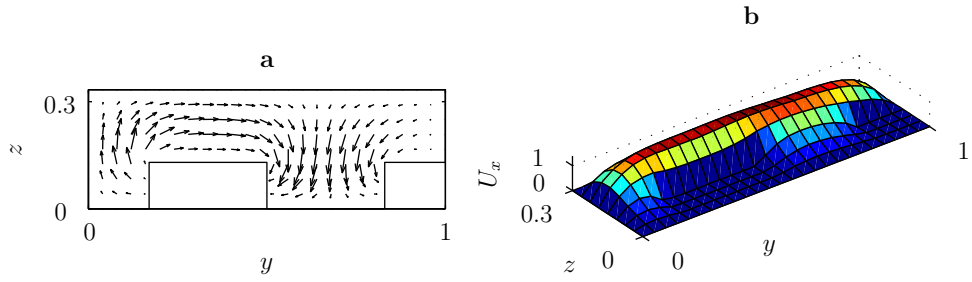


Figure 4.4: Velocity field for (4.1) solved in the Herringbone domain, Figure 4.2, at $x = 0.0$. **(a)** The y and z components of the velocity field. **(b)** The x component of the velocity field.

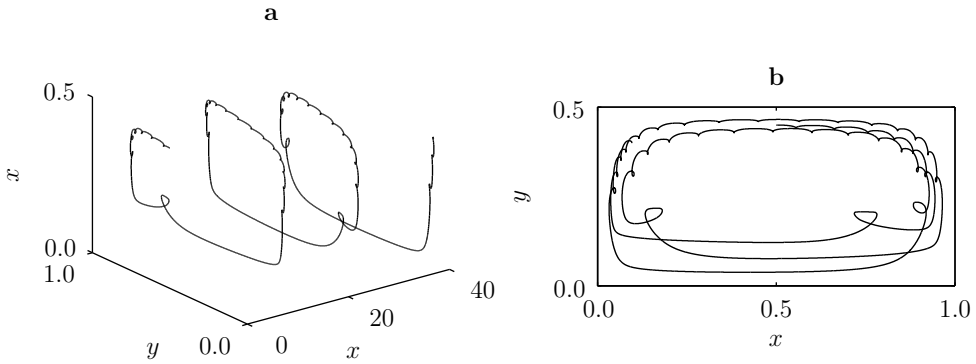


Figure 4.5: Computed orbit for $x = (0, 1/2, 9/20)$ in the velocity field U_h computed on the Ridge Domain. **(a)** Three dimensional plot. **(b)** Projection on the xy-plane.

We compute the projection matrices S_m as explained in Section 3.3 by approximating the action of L_m^* using the same method and the same time steps as for the dual problem (3.16). The recurrence problem is solved in the two different ways as described in Section 3.3, and depicted in Figures 4.8 and 4.9.

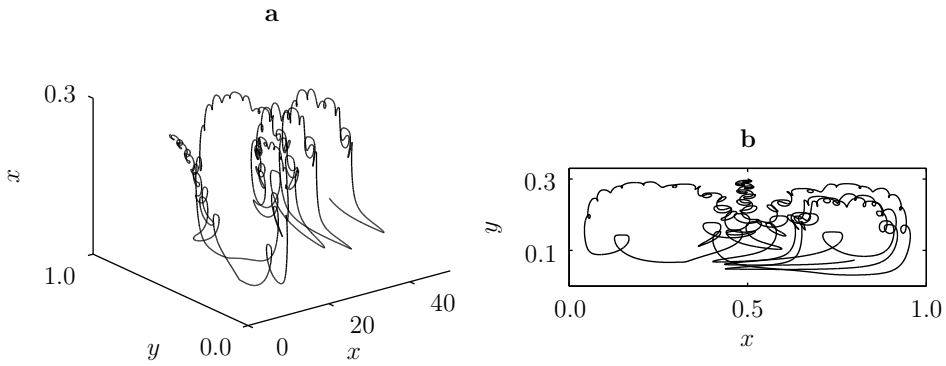


Figure 4.6: Computed orbit for $x = (0,\ 1/2,\ 1/3)$ in the velocity field U_h computed on the Herringbone Domain. **(a)** Three dimensional plot. **(b)** Projection on the xy-plane.

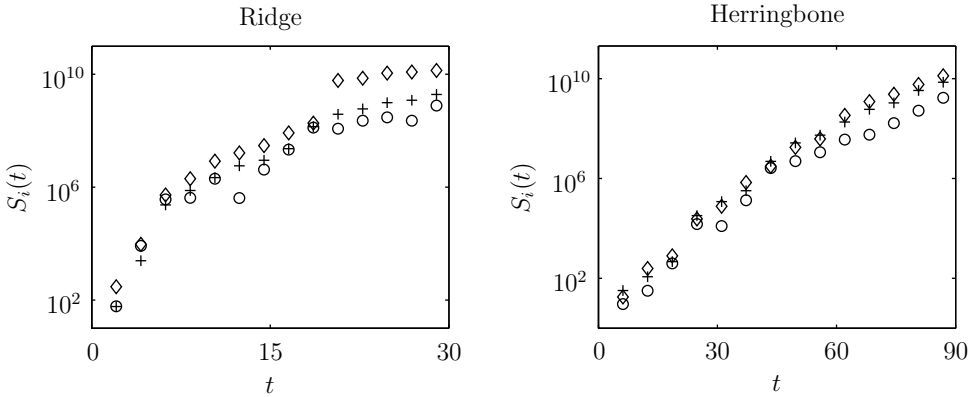


Figure 4.7: $(\circ, +, \diamond) = (S_0, S_1, S_2)$ Stability factors (3.18) for orbits in Figures 4.5 and 4.6.

5. Discussion

We have derived a shadowing error estimate (1.5) for computed orbits $u_k(t, x_i)$ to (1.1) with f replaced by a finite elements approximation f_h. Principal to the error estimate is the stability factors $\widetilde{S}_1(t)$ and $\widetilde{S}_2(t)$ which for sufficiently hyperbolic problems do not grow at any considerable rate as a function of the time t, in contrast to the stability factors S_i for the a posteriori error estimate where the stability factor grow at an exponential

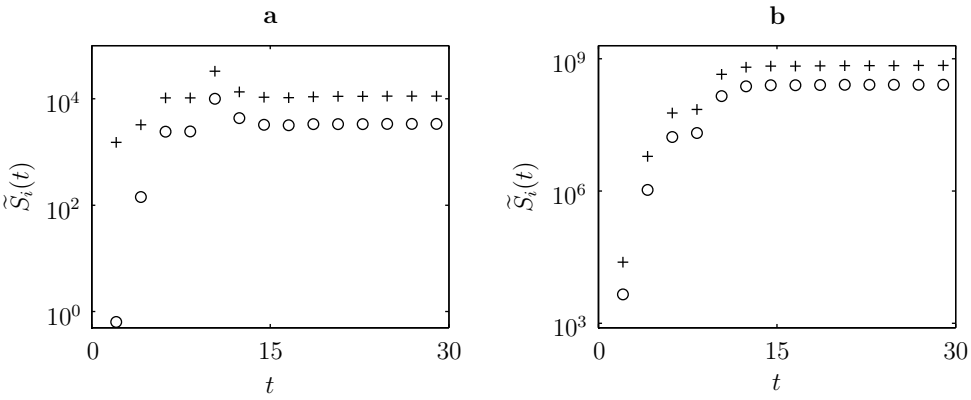


Figure 4.8: $(\circ, +) = (\widetilde{S}_1, \widetilde{S}_2)$ Stability factors (3.27) for orbit in Figure 4.5 computed as suggested in Section 3.3.1 **(a)** Case I **(b)** Case II.

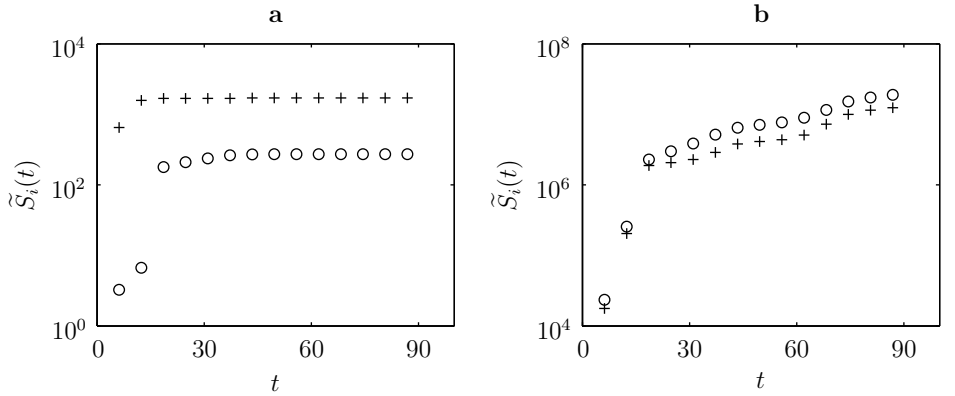


Figure 4.9: $(\circ, +) = (\widetilde{S}_1, \widetilde{S}_2)$ Stability factors (3.27) for the orbit in Figure 4.6 computed as suggested in Section 3.3.1 **(a)** Case I **(b)** Case II.

rate. We demonstrate this for orbits generated from the finite element velocity field modelled by the Stokes equations on two different domains, the Ridge Domain and the Herringbone Domain.

We note that there is an quite large difference in the way we choose to estimate the stability factors $\widetilde{S}_1(t)$ and $\widetilde{S}_2(t)$, either as in Case I or as in Case II as explained in Section 3.3.1, see Figures 4.8 and 4.9.

It is fare to say that the shadowing error estimate (1.5) is not rigorous as long as we do not control all constants in the estimate. At this stage we are not able to completely control the error in the finite element approximation

f_h. We only can provide asymptotic error estimates of e_f, that is, there is an unknown but bounded constant in the right hand side of the estimate and we can only deduce that the error goes to zero as $h \to 0$.

References

[1] L. Barreira and Y. Pesin, *Smooth Ergodic Theory and Nonuniformly Hyperbolic Dynamics*, Handbook of dynamical systems. Vol. 1B, Elsevier B. V., Amsterdam, 2006.

[2] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, second ed., Springer-Verlag, 2002.

[3] R. M. Brown and Z. Shen, *Estimates for the Stokes operator in Lipschitz domains*, Indiana Univ. Math. J. **44** (1995), 1183–1206.

[4] A. J. Chorin and J. E. Marsden, *A Mathematical Introduction to Fluid Mechanics*, second ed., Springer-Verlag, 1990.

[5] B. A. Coomes, Hüseyin Koçak, and K. J. Palmer, *Rigorous computational shadowing of orbits of ordinary differential equations*, Numer. Math. **69** (1995), 401–421.

[6] K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Introduction to adaptive methods for differential equations*, Acta Numer., 1995, pp. 105–158.

[7] ———, *Computational Differential Equations*, Cambridge University Press, 1996.

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

[9] G. P. Galdi, *An Introduction to the Mathematical Theory of the Navier-Stokes Equations. Vol. I*, Springer-Verlag, 1994.

[10] G. H. Golub and C. F. Van Loan, *Matrix Computations*, second ed., Johns Hopkins University Press, 1989.

[11] C. Grebogi, S. M. Hammel, J. A. Yorke, and T. Sauer, *Shadowing of physical trajectories in chaotic dynamics: containment and refinement*, Phys. Rev. Lett. **65** (1990), 1527–1530.

[12] W. Hayes and K. R. Jackson, *A survey of shadowing methods for numerical solutions of ordinary differential equations*, Appl. Numer. Math. **53** (2005), 299–321.

[13] S. Larsson, *Numerical analysis of semilinear parabolic problems*, The Graduate Student's Guide to Numerical Analysis '98 (Leicester), Springer, 1999, pp. 83–117.

[14] Anders Logg, *Multi-adaptive Galerkin methods for ODEs. II. Implementation and applications*, SIAM J. Sci. Comput. **25** (2004), 1119–1141.

[15] V. G. Maz'ya and J. Rossmann, *Lp estimates of solutions to mixed boundary value problems for the Stokes system in polyhedral domains*, ArXiv Mathematical Physics e-prints (2004).

[16] K. Palmer, *Shadowing in Dynamical Systems*, Kluwer Academic Publishers, 2000.

[17] Yakov B. Pesin, *Lectures on Partial Hyperbolicity and Stable Ergodicity*, European Mathematical Society (EMS), Zürich, 2004.

[18] A.D. Stroock, S.K.W. Dertinger, A. Ajdari, I. Mezic, H.A. Stone, and G.M. Whitesides, *Chaotic mixer for microchannels*, Science **295** (2002), 647 – 51.

[19] E. S. Van Vleck, *Numerical shadowing using componentwise bounds and a sharper fixed point result*, SIAM J. Sci. Comput. **22** (2000), 787–801.

[20] L-S. Young, *Developments in chaotic dynamics*, Notices Amer. Math. Soc. **45** (1998), 1318–1328.

Department of Mathematical Sciences, Chalmers University of Technology, SE-412 96 Göteborg, Sweden

E-mail address: erik.svensson@math.chalmers.se

Paper IV

OPTIMAL SEARCH IN FINITE ELEMENT TRIANGULATIONS USING BINARY TREES

ERIK D. SVENSSON

ABSTRACT. We propose a simple algorithm that, given the set S of all n-simplices, $n = 2, 3$, in a finite element triangulation and a query point $p \in \mathbf{R}^n$ will find one n-simplex or $\mathbf{R}^n \setminus S$ containing p in $O(\log N)$ search time, where N is the number of n-simplices in the triangulation. The algorithm requires $O(N \log N)$ preprocessing time and $O(N)$ storage. We apply the algorithm on two finite element triangulations and demonstrate that the search time is of the same order as the time to evaluate the barycentric coordinates of one n-simplex, which we regard a relevant time scale in many finite element applications.

1. INTRODUCTION

Given the set S of all n-simplices, $n = 2, 3$, in a finite element triangulation and a query point $p \in \mathbf{R}^n$ we pose the following search problem: *Does any n-simplex in S contain p?* This problem relates to two fundamental problems in computational geometry: the *planar subdivision search* problem, that is, given a planar subdivision in $\mathbf{R}^2$ with a number of line segments, determine which region in subdivision contains p; or the *post-office* problem, that is, given a set of points, find the point that is closest to p, see [6] and references there in. There are many solutions and suggestions how to solve the subdivision search problem, for example [5, 4, 2, 7, 6]. With N denoting the number of n-simplices in S we characterize, cf. [4], a solution or algorithm to the posed search problem by: (1) *preprocessing time* -the time to construct search structures, (2) *space* -the storage used by the method, (3) *search time* -the time required to locate the region or point in S. It is possible to solve the search problem in an optimal way,

Date: April 18, 2006.

2000 *Mathematics Subject Classification.* 68U05.

Key words and phrases. computational geometry, point location, finite element, post-processing.

that is, with $O(N)$ preprocessing time, $O(N)$ space and $O(\log N)$ search time [4, 6]. However, these methods are often considered too complicated [2] and although the search time scales linearly the constant in the linear dependence, the *query constant*, may be large [6].

In this work we propose a simple algorithm to solve the posed search problem that is characterized as optimal in search time and space and requires $O(N \log N)$ preprocessing time.

2. Preliminaries

We first introduce a few concepts used in finite element practice and theory, see for example [1].

Let A_i for $i = 1, \ldots, n+1$ be scalars not all equal to zero. A *hyperplane* π is subspace of $\mathbf{R}^n$ such that

$$\pi = \left\{ x \in \mathbf{R}^n : \sum_{i=1}^{n} A_i x_i + A_{n+1} = 0 \right\} \tag{2.1}$$

An *n-simplex* in $\mathbf{R}^n$ is the convex hull T of $n+1$ points $a_1, \ldots, a_{n+1}$, called vertices, not all contained in a hyperplane, that is, for $n = 0, \ldots, 3$: a point, a line segment, a triangle, or a tetrahedron. For $0 \leq m \leq n$, an *m-face* of the n-simplex T is an m-simplex whose vertices are also vertices of T.

Let $\Omega \subset \mathbf{R}^n$ be a polyhedral domain. A *triangulation* $\mathcal{T}$ is a partition of Ω into n-simplices T such that no vertex of any simplex lies in the interior of any m-face, for $1 < m < n$. A family of triangulations $\{\mathcal{T}_h\}_{h>0}$ is said to be shape-regular if there is a $\gamma > 0$ such that $h_T/\rho_T \leq \gamma$ for all $T \in \bigcup_h \mathcal{T}_h$, where $h_T = \max_{T \in \mathcal{T}_h} \operatorname{diam}(T)$ and $\rho_T = \sup\{\operatorname{diam}(S) : S \text{ is a ball contained in } T\}$.

An *n-rectangle* is a set of the form

$$R = \prod_{i=1}^{n} [a_i, b_i] = \{x = (x_1, \ldots, x_n) : a_i \leq x_i \leq b_i,\ 1 \leq i \leq n\}. \tag{2.2}$$

Again, let S be the set of all n-simplices, $n = 2, 3$, in a finite element triangulation $\mathcal{T}$ and set $\operatorname{card}(S) := N$, that is, we use the cardinal number to count the number of n-simplices in S.

In the complexity analysis we use a parameter N to measure the size of the search problem. We may interchangeably take N as the number of m-simplices, $0 \leq m \leq n$, in the triangulation. In $\mathbf{R}^2$ this is solely due

to the Euler relations whereas in $\mathbf{R}^3$ we will have to impose additional constraints on the triangulation.

Consider a triangulation $\mathcal{T}$, where we now assume that Ω and $\partial\Omega$ are simply connected which will only influence the Euler relations that we will use. Let N_i for $m = 0, \ldots, n$ be the number of m-simplices in the the triangulation and let N_m^∂ for $m = 0, \ldots, n-1$ be the number of m-simplices on the boundary of the triangulation.

We first consider $n = 2$. By counting the edges and triangles in the triangulation we get the identity $2N_1 - N_1^\partial = 3N_2$ and since $0 \leq N_1^\partial \leq N_1$ we may estimate

$$\frac{3}{2}N_2 \leq N_1 \leq 3N_2,$$

which shows that the number of edges and triangles are interchangeable. Inserting this into the Euler relation for triangulations in $\mathbf{R}^2$, see for example [3],

$$N_0 - N_1 + N_2 = 1,$$

we get

$$1 + \frac{1}{2}N_2 \leq N_0 \leq 1 + 2N_2,$$

which shows that the number of vertices and triangles are interchangeable.

Consider next $n = 3$. By counting the faces and tetrahedra in the triangulation we get the identity

$$2N_2 - N_2^\partial = 4N_3 \tag{2.3}$$

and since $0 \leq N_2^\partial \leq N_2$ we may estimate

$$2N_3 \leq N_2 \leq 4N_3,$$

which shows that the number of faces and tetrahedra are interchangeable.

By counting the edges and tetrahedra in the triangulation we obtain $\sum_{i=1}^{N_1} a_i = 6N_3$, where $a_i = \text{card}(\{T \in \mathcal{T} : E_i \cap T = E_i\})$ is the number of tetrahedra neighboring the edge E_i. Hence

$$\bar{a}N_1 = 6N_3, \tag{2.4}$$

where $\bar{a} = N_1^{-1} \sum_{i=1}^{N_1} a_i$ is the average of $\{a_i\}_{i=1}^{N_1}$, which shows that the number of edges and tetrahedra are interchangeable.

Also by counting the edges and faces on the boundary we get the identity $2N_1^\partial = 3N_2^\partial$ which together with the Euler relation on the boundary $N_0^\partial -$

$N_1^\partial + N_2^\partial = 2$ implies that $N_2^\partial = 2(N_0^\partial - 2)$ and with (2.3) we get the identity

$$2N_2 = 4N_3 + 2(N_0^\partial - 2). \tag{2.5}$$

Inserting (2.4) and (2.5) into the Euler relation for triangulations in $\mathbf{R}^3$

$$N_0 - N_1 + N_2 - N_3 = 1,$$

we get

$$N_0 + N_0^\partial = \Big(\frac{6}{\bar{a}} - 1\Big) N_3,$$

and since $0 \leq N_0^\partial \leq N_0$ we may estimate

$$\frac{1}{2}\Big(\frac{6}{\bar{a}} - 1\Big) N_3 + \frac{3}{2} \leq N_0 \leq \Big(\frac{6}{\bar{a}} - 1\Big) N_3 + 3,$$

which shows that the number of vertices and tetrahedra are interchangeable.

Notice that we will have to impose $\bar{a} < 6$ in order to have a use full estimate. This is often true in practice since the triangulation is generated with a shape-regularity constraint. We also remark that we may try to use the uniform estimate

$$N_1 \min_{i=1,\ldots,N_1} a_i \leq \sum_{i=1}^{N_1} a_i \leq N_1 \max_{i=1,\ldots,N_1} a_i,$$

in the analysis above but in practice this will often be useless since instead of imposing $\bar{a} < 6$ we will have to impose $\max_{i=1,\ldots,N_1} a_i < 6$ which is not likely to be true in practice –there are always a few edges where the condition fails. The mean value $\bar{a}$ is a milder condition better suited in this situation.

In conclusion, we just showed that provided $\bar{a} < 6$, we could use either $N = N_i$, for $i = 0, \ldots, n$ in the complexity analysis. We will use this fact without further notice throughout this work.

Finally, we will use the notion *binary tree* denoting a data structure devised for fast data searching [8]. The binary tree contains a number of items called *nodes* of the tree. Each node contains data and zero or two links connecting to other nodes in the tree. The first node in the tree is called the *root*. A connected set of nodes of the tree is called a *subtree* and a node that has no connections to other nodes is called a *leaf*. See the illustrations in Figure 2.1. The *height* of the tree is equal to the maximal number of nodes connecting the root and any leaf.

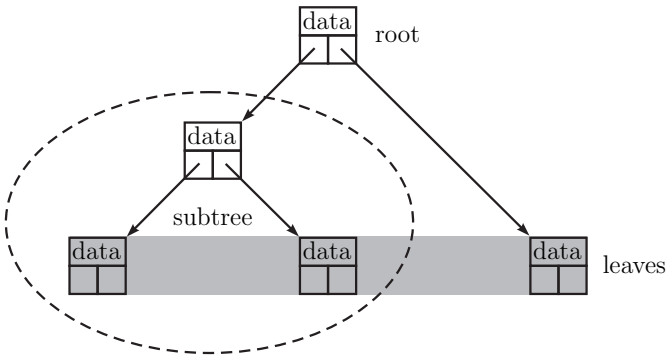


Figure 2.1: Binary tree.

3. Binary search in triangulations

We devise a binary tree that will be used to find the n-simplex containing the query point p, or no n-simplex if p is not in Ω. Each node in the tree will contain numbers A_i for $i = 1, \ldots, n+1$ representing a hyperplane π, and the subtrees *negSubtree* and *posSubtree*, also binary trees that are parts of the entire tree. The hyperplanes will partition $\mathbf{R}^n$ into negative and positive sides that will be used to sort the n-simplices in the triangulation at preprocessing. As a result of this sorting, every leaf will contain a set of n-simplices S_l where ideally S_l is such that $\text{card}\,(S_l) = 1$ or at least close to 1.

Given a query point p we use the search Algorithm 1 to find 0 or 1 n-simplex in the triangulation containing the point, 0 meaning that p is outside the Ω. In the algorithm we use a generic algorithm `inSimplex`(T, p) to test whether T contains p and we refer to Algorithms 4 or 5 in the Appendix for details.

In the sections below we describe two algorithms for constructing the binary tree. Both algorithms however suffer from different deficiencies and it is only after combining them in a new one we obtain an algorithm that will be useful in practice.

3.1. Partitioning along x_i-hyperplanes.

Let R be the smallest n-rectangle containing Ω. For a_i and b_i as in (2.2), defining the n-rectangle, set $a = (a_1, \ldots, a_n)$ and

$$dx = (dx_1, \ldots, dx_n) = (b_1 - a_1, \ldots, b_n - a_n).$$

Algorithm 1: `findSimplex`(point p, binary tree *binaryTree*)

Input: point p, binary tree *binaryTree*
Output: n-simplex T or 0 (no n-simplex)

if *no subtrees* **then** `/* at a leaf */`
  forall $T \in S_l$ **do** `/* linear search */`
    if `inSimplex`(T, p) **then** `/* see Algorithm 4 or 5 */`
      return T
  return 0 `/* no` n`-simplex was found */`
else `/* choose a subtree */`
  if $\sum_i^n A_i p_i + A_{n+1} < 0$ **then**
    return `findSimplex`$(p, negSubtree)$
  else
    return `findSimplex`$(p, posSubtree)$

Find the largest side of R, and set $i = \text{argmax}_{i=1,\dots,n}(dx_i)$ and let π denote the hyperplane with $A_i = 1$ and $A_{n+1} = -a_i - dx_i/2$ ($A_j = 0$ for $j \neq i$ and $j < n+1$). Partition R along the hyper plane π into to n-rectangles R_- and R_+. Sort the n-simplices $T \in S$, where we recall that S is the set of all n-simplices in the triangulation, now also contained in R. Add T to S_- if $T \cap R_- \neq \emptyset$ and add T to S_+ if $T \cap R_+ \neq \emptyset$. Repeat this procedure recursively for the pairs (R_-, S_-) and (R_+, S_+) until $\text{card}\,(S) < 2$ or $\text{card}\,(S) = \text{card}\,(S_-)$ or $\text{card}\,(S) = \text{card}\,(S_+)$. We summarize this procedure in Algorithm 2.

The height of the tree is $\sim \log N$ and each recursive step in the preprocessing requires sorting $\sim N$ n-simplices. Hence, the preprocessing time for the binary tree is $O(N \log N)$.

The search time will require $O(\log N)$ operations, but the query constant will be rather large since at the leafs a linear search is preformed. The number of simplices in S_l will be roughly bounded by the number of n-simplices neighboring a node in the triangulation, in practice this is ~ 10 for $n = 2$ and ~ 40 for $n = 3$. This will slow down the search and due to this Algorithm 2 is not a good choice in practice.

3.2. **Partitioning along $(n-1)$-faces.** Recall that S is the set of all n-simplices in $\mathcal{T}$ and that N_{n-1} is the number of $(n-1)$-faces in the triangulation. Let π_i for $i = 1, \dots, N_{n-1}$ be the hyperplanes defined by the $(n-1)$-faces in the triangulation. Denote the halfspaces on opposite sides

Algorithm 2: `binaryTreeRectangular`(S, a, dx)

Input: a set S of n-simplices, a and dx defining an n-rectangle R
Output: binary tree data structure
Data: the `binaryTreeRectangular` contain numbers A_i for $i = 1, \ldots, n+1$ representing the hyperplane π, subtrees *negSubtree* and *posSubtree*, and a set S_l of n-simplices.

$A_i = 0$ for $i = 1, \ldots, n+1$ `/* initialization */`
if $card\,(S) < 2$ **then** `/* if leaf */`
  $S_l = S$
  return *this* ***binaryTreeRectangular***
else
  $i = \text{argmax}_{i=1,\ldots,n}(dx_i)$
  $dx_i = dx_i/2$
  $A_i = 1$
  $A_{n+1} = -a_i - dx_i$
  forall $T \in S$ **do** `/* sort simplices */`
    if $\sum_{i=1}^{n} A_i a_j + A_{n+1} < 0$ *for one vertex* $a_j \in T$ **then** add T to S_-
    if $\sum_{i=1}^{n} A_i a_j + A_{n+1} > 0$ *for one vertex* $a_j \in T$ **then** add T to S_+
  if $card\,(S) > card\,(S_-)$ *and* $card\,(S) > card\,(S_+)$ **then** `/* new subtrees */`
    negSubtree = `binaryTreeRectangular`(S_-, a, dx)
    $a_i = a_i + dx_i$
    posSubtree = `binaryTreeRectangular`(S_+, a, dx)
  else `/* leaf */`
    $S_l = S$
    return *this* ***binaryTreeRectangular***

of π_i by $\mathbf{R}^n_{i,-}$ and $\mathbf{R}^n_{i,+}$. Now sort the simplices $T \in S$ and add T to $S_{i,-}$ if $T \cap \mathbf{R}^n_{i,-} \neq \emptyset$ and add T to $S_{i,+}$ if $T \cap \mathbf{R}^n_{i,+} \neq \emptyset$. Choose one of these hyperplanes $\pi = \pi_i$ such that

$$i = \text{argmax}_{i=1,\ldots,N_{n-1}} \begin{cases} \text{card}\,(S_{i,-})/\text{card}\,(S_{i,+}) & \text{if card}\,(S_{i,-}) < \text{card}\,(S_{i,+}), \\ \text{card}\,(S_{i,+})/\text{card}\,(S_{i,-}) & \text{otherwise,} \end{cases}$$

and set $S_- = S_{i,-}$ and $S_+ = S_{i,+}$. Repeat the procedure recursively for S_- and S_+ until $\text{card}\,(S) < 2$ or $\text{card}\,(S) = \text{card}\,(S_-)$ or $\text{card}\,(S) = \text{card}\,(S_+)$.

This procedure creates a binary tree and we summarize it in Algorithm 3.

The height of the tree is $\sim \log N$ and each recursive step in the preprocessing requires sorting $\sim N^2$ n-simplices. Hence, the preprocessing time for the binary tree is $O(N^2 \log N)$, which is far from optimal.

The search time will require $O(\log N)$ operations and the query constant will be rather good. Also, in this situation, a linear search is performed at the leafs. However, in this case the number of simplices in S_l will be small, mostly 1 and with small and rare variations. We have not made any attempts to give a rigorous upper bound for the number of simplices in S_l.

Due to the scaling of the preprocessing time this algorithm is not a good choice in practice, at least not for large triangulations.

3.3. `binaryTreeRectangular` **and** `binaryTreeFace` **combined.** We notice that the deficiencies in Algorithms 2 and 3 are complementary, small preprocessing time and large search time for Algorithm 2 but large preprocessing time and small search time for Algorithm 3. In other words it seems desirable to combine the algorithms in such way that only the favorable characteristics of the algorithms remain and cancel the deficiencies. The idea is to let Algorithm 3 continue where Algorithm 2 is terminated. We input $S = S_l$ from Algorithm 2 into Algorithm 3 and let it refine the tree further. In this way we will gain a binary tree with good query constant since the $\text{card}(S_l)$ after Algorithm 3 has terminated will be small, and since Algorithm 3 is only applied on small sets S from Algorithm 2 it will not have major impact on the total preprocessing time.

If we assume that there are $\sim N$ different leaves in the tree after Algorithm 2 has terminated and that each such leaf holds M n-simplices then the total preprocessing time will be the preprocessing time for Algorithm 2 plus the preprocessing time for Algorithm 3 applied on N sets each holding M n-simplices, that is, the total preprocessing time will scale like $O(N \log N + NM^2 \log M)$ which is close to $O(N \log N)$ for small M and large N.

Note that we may also try to apply Algorithm 3 on a smaller set $S_s \subset S$ (S outputted from Algorithm 2), chosen by some means, which will improve the preprocessing time at the expense of the search time. For example we may take S_s to be the n-simplex whose barycenter is closest to the center of mass of all barycenters of all n-simplices in S. Then $\text{card}\,(S_s) = 1$ and the total complexity will be $O(N \log N)$. This will alter $\text{card}\,(S_l)$ at the

final leafs, when Algorithm 3 has terminated, and card (S_l) will be larger but still relatively small when compared to `binaryTreeRectangular`.

4. Numerical examples

We now consider two triangulations, one in $\mathbf{R}^2$, Figure 4.1, and the other in $\mathbf{R}^3$, Figure 4.2. We build the search structure proposed in Section 3.3 and measure: the preprocessing time and the average search time for 10^6 randomly chosen query points as function of number of nodes N in the triangulations as we perform 4 and 3 uniform refinements in the $\mathbf{R}^2$ and $\mathbf{R}^3$ triangulations, respectively.

The preprocessing time is normalized with the preprocessing time for the triangulations at start and the search time is normalized with the time it takes evaluating the barycentric coordinates for one n-simplex, see the Appendix where we account for the implementation used. The motivation for the normalization of the search time is to find a time scale appropriate for finite element applications. For example, it is often necessary to evaluate the barycentric coordinates when post-processing finite element data.

In Figures 4.3 and 4.4 we visualize the search process in the two dimensional triangulation. We search for a query point contained in the shaded triangle in Figure 4.3 and marked with the bullet $\bullet$ in Figure 4.4. We also plot the the hyperplanes π (lines) used to partition the triangles in the triangulation. After 12 levels in the binary tree the triangle containing the query point could be identified. There are 9 layers from Algorithm 2 and 3 layers from Algorithm 3.

Finally we plot the results from the measurements in Figures 4.5 and 4.6, where we also make a least square data fit to the appropriate scaling, $O(N \log N)$ for the preprocessing time and $O(\log N)$ for the search.

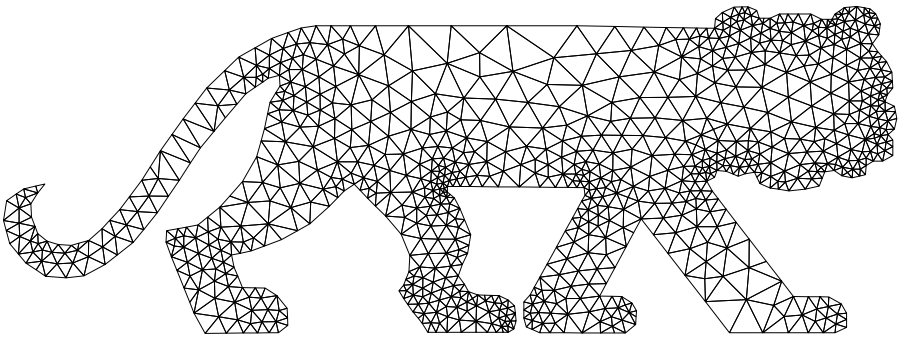


Figure 4.1: A two-dimensional triangulation with 940 nodes and 1572 triangles.

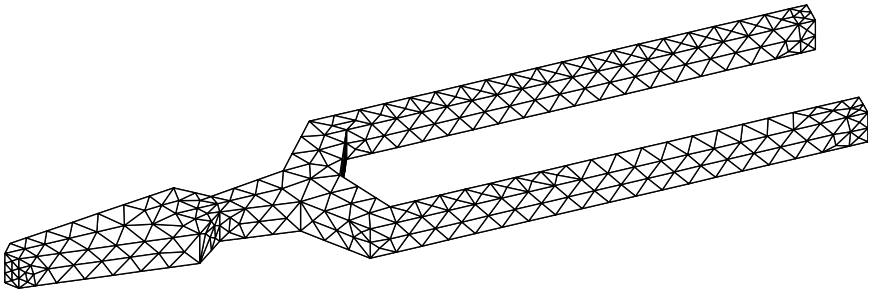


Figure 4.2: A three-dimensional triangulation with 578 nodes and 1567 tetrahedra.

Algorithm 3: `binaryTreeFace(`S`)`

Input: a set S of n-simplices
Output: binary tree data structure
Data: the `binaryTreeFace` contain numbers A_i for $i = 1, \ldots, n+1$ representing the hyperplane π, subtrees *negSubtree* and *posSubtree*, and a set S_l of n-simplices.

$A_i = 0$ for $i = 1, \ldots, n+1$ `/* initialize */`
if $card(S) < 2$ **then** `/* if leaf */`
  $S_l = S$
  return *this* ***binaryTreeFace***
else
  r = 0.0 `/* parameter do decide the best partition */`
  `/* Let` π_i `with scalars` B_i `be the hyper planes defined by the` N_{n-1} $(n-1)$`-faces in` S`. */`
  forall π_i **do**
    forall $T \in S$ **do** `/* sort simplices */`
      if $\sum_{i=1}^{n} B_i a_j + B_{n+1} < 0$ *for one vertex* $a_j \in T$ **then**
        add T to $S_{i,-}$
      if $\sum_{i=1}^{n} B_i a_j + B_{n+1} > 0$ *for one vertex* $a_j \in T$ **then**
        add T to $S_{i,+}$
    if $card(S_{i,-}) \leq card(S_{i,+})$ *and* $r < card(S_{i,-})/card(S_{i,+})$ **then**
      $r = \text{card}\,(S_{i,-})/\text{card}\,(S_{i,+})$
      $S_- = S_{i,-}$ and $S_+ = S_{i,+}$
      $A_i = B_i$ for $i = 1, \ldots, n+1$
    else if $card(S_{i,+}) < card(S_{i,-})$ *and* $r < card(S_{i,+})/card(S_{i,-})$ **then**
      $r = \text{card}\,(S_{i,+})/\text{card}\,(S_{i,-})$
      $S_- = S_{i,-}$ and $S_+ = S_{i,+}$
      $A_i = B_i$ for $i = 1, \ldots, n+1$
  if $card(S) > card(S_-)$ *and* $card(S) > card(S_+)$ **then** `/* new subtrees */`
    negSubtree = `binaryTreeFace(`S_-`)`
    posSubtree = `binaryTreeFace(`S_+`)`
  else `/* leaf */`
    $S_l = S$
    return *this* ***binaryTreeFace***

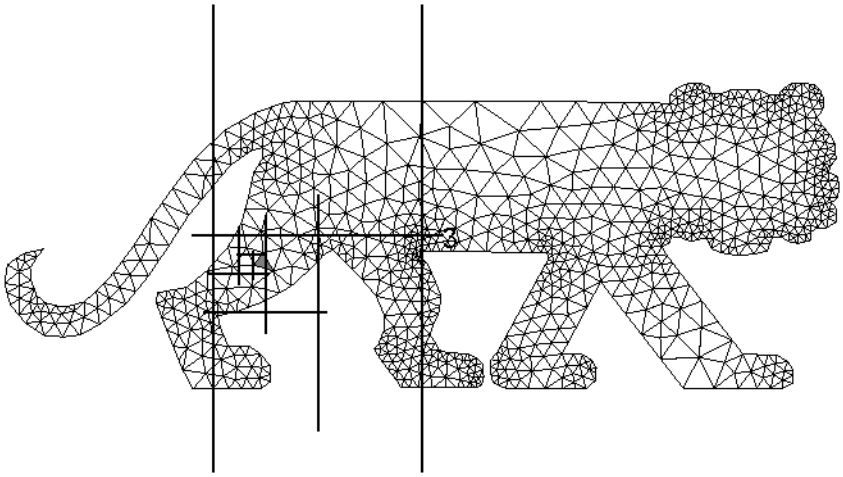


Figure 4.3: Search process using the binary tree with Algorithm 2. We are searching for a query point contained in the shaded triangle in the rear leg of the tiger. The horizontal and vertical lines are the hyperplanes π used to partition the triangles in the triangulation.

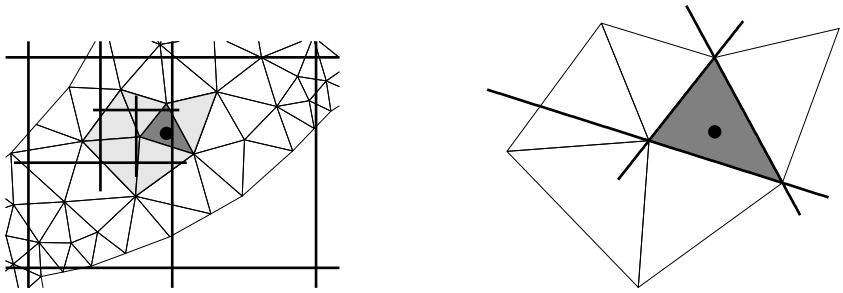


Figure 4.4: The query point is marked with the bullet $\bullet$. **(left)** Search in the tree with Algorithm 2, zoom in. The set of shaded triangles is the set S_l in the leaf from Algorithm 2. **(right)** Search in the tree with Algorithm 3. The algorithm terminates with one triangle in the final leaf.

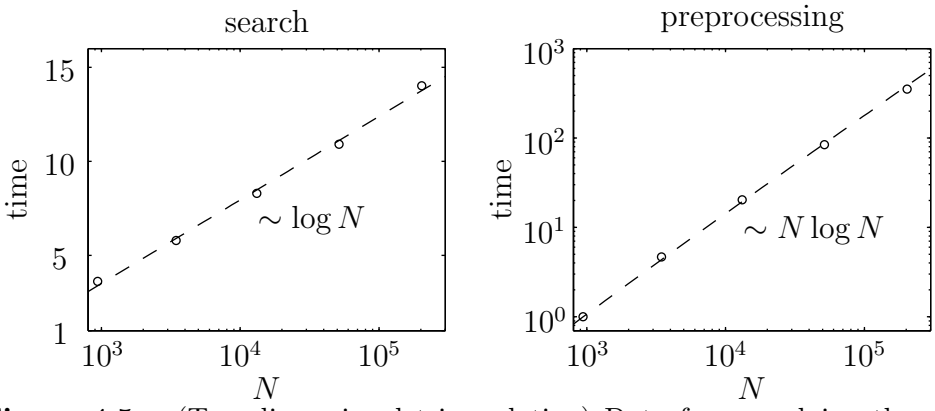


Figure 4.5: (Two-dimensional triangulation) Data from applying the algorithm in Section 3.3 to the triangulation in Figure 4.1. The dashed lines are least square fits. **(left)** Average search time for 10^6 randomly chosen query points. The time is normalized with the time to evaluate the barycentric coordinates for one triangle, which is a characteristic time scale in finite element post-processing. **(right)** Preprocessing time normalized with the preprocessing time of the triangulation at start.

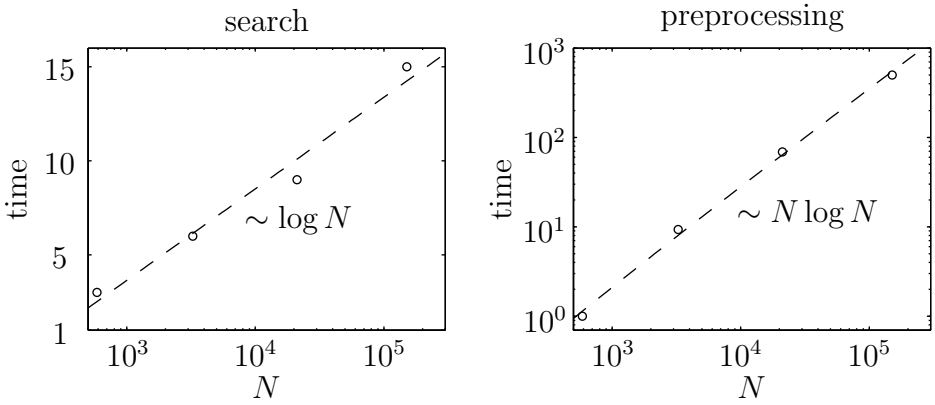


Figure 4.6: (Three-dimensional triangulation) Data from applying the algorithm in Section 3.3 to the triangulation in Figure 4.2. The dashed lines are least square fits. **(left)** Average search time for 10^6 randomly chosen query points. The time is normalized with the time to evaluate the barycentric coordinates for one tetrahedron, which is a characteristic time scale in finite element post-processing. **(right)** Preprocessing time normalized with the preprocessing time of the triangulation at start.

Appendix A. Various simple algorithms for n-simplices

In this appendix we give account for various simple algorithms or mere implementations of mathematical notions that we have used throughout this work on n-simplices with vertices in $a_i = (x_i, y_i)$ for $i = 1, \dots, 3$ (triangles) or $a_i = (x_i, y_i, z_i)$ for $i = 1, \dots, 4$ (tetrahedra) as in Figure A.1. We represents barycentric coordinates λ with the $(n+1) \times (n+1)$ matrices $\mathcal{M}$. For $x \in \mathbf{R}^n$ we then get the barycentric coordinate as $\lambda = \mathcal{M}x$.

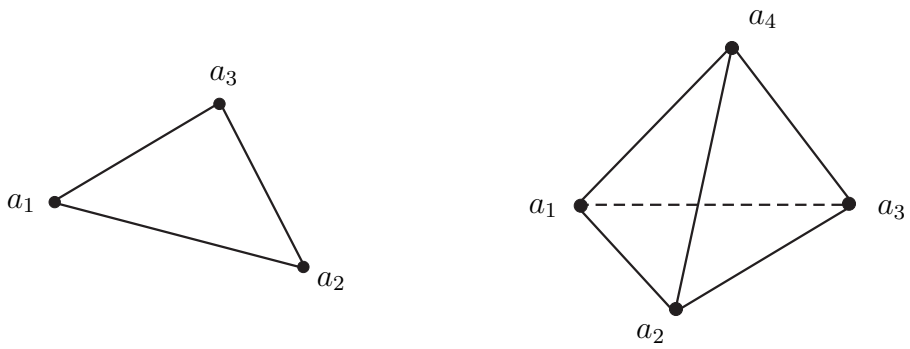


Figure A.1: n-simplices. **(left)** A triangle with vertices a_1, a_2 and a_3. **(right)** A tetrahedron with vertices a_1, a_2, a_3 and a_4.

A.1. Triangles, $n = 2$.

A.1.1. *Volume.* We compute the signed volume as the vector product $V(a_1, a_2, a_3) = ((a_2 - a_1) \times (a_3 - a_1))/2$ which in terms of the vertices is

$$V(a_1, a_2, a_3) = \frac{1}{2}\big((x_2 - x_1)(y_3 - y_1) - (x_3 - x_1)(y_2 - y_1)\big),$$

and the volume is $|V(a_1, a_2, a_3)|$.

A.1.2. *Barycentric coordinates.* The matrix $\mathcal{M}$ is the inverse to

$$\begin{pmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{pmatrix},$$

cf. [1], and we get

$$\mathcal{M} = V(T)^{-1} \begin{pmatrix} y_2 - y_3 & x_3 - x_2 & x_2y_3 - x_3y_2 \\ y_3 - y_1 & x_1 - x_3 & x_3y_1 - x_1y_3 \\ y_1 - y_2 & x_2 - x_1 & x_1y_2 - x_2y_1 \end{pmatrix},$$

A.1.3. *Point in a triangle.* In order to test whether a point p is contained in a triangle T we test if p and a_3 are on the same side of the line trough a_1 and a_2, and likewise for the other two vertices, cf. [6, Code 1.6, p. 29].

Algorithm 4: `inTriangle`(T, p)

Input: triangle T, point p
Output: true ($p \in T$) or false ($p \notin T$)

$v = V(a_1, a_2, a_3)$
if $v * V(a_1, a_2, p) < 0.0$ **then**
  return *false*
if $v * V(a_3, a_1, p) < 0.0$ **then**
  return *false*
if $v * V(a_2, a_3, p) < 0.0$ **then**
  return *false*
return *true*

A.2. **Tetrahedra,** $n = 3$**.**

A.2.1. *Volume.* We compute the signed volume as the vector triple product $V(a_1, a_2, a_3, a_4) = (a_2 - a_1) \cdot ((a_3 - a_1) \times (a_4 - a_1))/6$ which in terms of the vertices is

$$\begin{aligned} V(a_1, a_2, a_3, a_4) = \frac{1}{6}\big(&- (x_4 - x_1)(y_3 - y_1)(z_2 - z_1) \\ &+ (x_3 - x_1)(y_4 - y_1)(z_2 - z_1) \\ &+ (x_4 - x_1)(y_2 - y_1)(z_3 - z_1) \\ &- (x_2 - x_1)(y_4 - y_1)(z_3 - z_1) \\ &- (x_3 - x_1)(y_2 - y_1)(z_4 - z_1) \\ &+ (x_2 - x_1)(y_3 - y_1)(z_4 - z_1)\big), \end{aligned}$$

and the volume is $|V(a_1, a_2, a_3, a_4)|$.

A.2.2. *Barycentric coordinates.* The matrix $\mathcal{M}$ is the inverse to

$$\begin{pmatrix} x_1 & x_2 & x_3 & x_4 \\ y_1 & y_2 & y_3 & y_4 \\ z_1 & z_2 & z_3 & z_4 \\ 1 & 1 & 1 & 1 \end{pmatrix},$$

cf. [1], and we get

$$\mathcal{M} = V(T)^{-1} \begin{pmatrix} | & | & | & | \\ \mathcal{M}_1 & \mathcal{M}_2 & \mathcal{M}_3 & \mathcal{M}_4 \\ | & | & | & | \end{pmatrix},$$

where

$$\mathcal{M}_1 = \begin{pmatrix} y_4(z_3 - z_2) + y_3(z_2 - z_4) + y_2(z_4 - z_3) \\ y_4(z_1 - z_3) + y_1(z_3 - z_4) + y_3(z_4 - z_1) \\ y_4(z_2 - z_1) + y_2(z_1 - z_4) + y_1(z_4 - z_2) \\ y_3(z_1 - z_2) + y_1(z_2 - z_3) + y_2(z_3 - z_1) \end{pmatrix},$$

$$\mathcal{M}_2 = \begin{pmatrix} x_4(z_2 - z_3) + x_2(z_3 - z_4) + x_3(z_4 - z_2) \\ x_4(z_3 - z_1) + x_3(z_1 - z_4) + x_1(z_4 - z_3) \\ x_4(z_1 - z_2) + x_1(z_2 - z_4) + x_2(z_4 - z_1) \\ x_3(z_2 - z_1) + x_2(z_1 - z_3) + x_1(z_3 - z_2) \end{pmatrix},$$

$$\mathcal{M}_3 = \begin{pmatrix} x_4(y_3 - y_2) + x_3(y_2 - y_4) + x_2(y_4 - y_3) \\ x_4(y_1 - y_3) + x_1(y_3 - y_4) + x_3(y_4 - y_1) \\ x_4(y_2 - y_1) + x_2(y_1 - y_4) + x_1(y_4 - y_2) \\ x_3(y_1 - y_2) + x_1(y_2 - y_3) + x_2(y_3 - y_1) \end{pmatrix},$$

$$\mathcal{M}_4 = \begin{pmatrix} x_4(y_2z_3 - y_3z_2) + x_3(y_4z_2 - y_2z_4) + x_2(y_3z_4 - y_4z_3) \\ x_4(y_3z_1 - y_1z_3) + x_3(y_1z_4 - y_4z_1) + x_1(y_4z_3 - y_3z_4) \\ x_4(y_1z_2 - y_2z_1) + x_2(y_4z_1 - y_1z_4) + x_1(y_2z_4 - y_4z_2) \\ x_3(y_2z_1 - y_1z_2) + x_2(y_1z_3 - y_3z_1) + x_1(y_3z_2 - y_2z_3) \end{pmatrix},$$

A.2.3. *Point in a tetrahedron.* In order to test whether a point p is contained in a tetrahedron T we test if p and a_4 are on the same side of the plane trough a_1, a_2 and a_3, and likewise for the other three vertices, cf. [6, Code 1.6, p. 29].

Algorithm 5: `inTetrahedron`(T, p)

Input: tetrahedron T, point p
Output: true ($p \in T$) or false ($p \notin T$)

$v = V(a_1, a_2, a_3, a_4)$
if $v * V(a_1, a_2, a_3, p) < 0.0$ **then**
  return *false*
if $v * V(a_1, a_4, a_2, p) < 0.0$ **then**
  return *false*
if $v * V(a_1, a_3, a_4, p) < 0.0$ **then**
  return *false*
if $v * V(a_2, a_4, a_3, p) < 0.0$ **then**
  return *false*
return *true*

References

[1] P. G. Ciarlet, *Basic error estimates for elliptic problems*, Handbook of Numerical Analysis, Vol. II, North-Holland, 1991.

[2] O. Devillers, S. Pion, and M. Teillaud, *Walking in a triangulation*, Internat. J. Found. Comput. Sci. **13** (2002), 181–199.

[3] A. Ern and J-L. Guermond, *Theory and Practice of Finite Elements*, Springer-Verlag, 2004.

[4] D. Kirkpatrick, *Optimal search in planar subdivisions*, SIAM J. Comput. **12** (1983), 28–35.

[5] R. J. Lipton and R. E. Tarjan, *Applications of a planar separator theorem*, SIAM J. Comput. **9** (1980), 615–627.

[6] J. O'Rourke, *Computational Geometry in C*, second ed., Cambridge University Press, 1998.

[7] N. Sarnak and R. E. Tarjan, *Planar point location using persistent search trees*, Comm. ACM **29** (1986), 669–679.

[8] H. Schildt, *C the Complete Reference*, third ed., McGraw-Hill, 1995.

Department of Mathematical Sciences, Chalmers University of Technology, SE-412 96 Göteborg, Sweden

E-mail address: `erik.svensson@math.chalmers.se`

Paper V

MULTIGRID FOR QUADRATIC FINITE ELEMENTS

ERIK D. SVENSSON

Abstract. We investigate the convergence rate of the finite element multigrid method applied on quadratic finite element approximations for problems with full and less than full regularity.

1. Introduction

The finite element multigrid method solves linear systems of equations arising from finite element approximations to linear elliptic partial differential equations with a number of operations proportional to the number of unknowns. We say that the multigrid method has optimal complexity or scales optimally. The method is founded on solid theoretical results which are reviewed in for example [7, 13, 21]. However, it is important to note that this rather general statement is really limited to linear finite element approximations. For higher degree finite element approximations the convergence rate of the multigrid method may deteriorate see, for example, [16] and for the similar problem for the algebraic multigrid method [14].

On the other hand, for sufficiently smooth problems and for finite element approximations of degree $q > 1$ we may achieve $O(h^{q+1})$ convergence in the error $u - u_h$, measured in some suitable norm, where h is the mesh size and, u and u_h are the exact and the finite element solutions, respectively. This is appealing and motivate us to study multigrid solvers for higher degree approximations. Moreover, there are situations that for other reasons require higher degree approximations, for example, solving saddle point problems such as the Stokes equations using the Hood-Taylor finite elements.

In this work we demonstrate that the multigrid method in practice also works well for quadratic finite element approximations of problems with

Date: April 19, 2006.
2000 *Mathematics Subject Classification.* 65N55, 65N30.
Key words and phrases. multigrid, finite elements.

both full regularity and less than full regularity. We compare two different finite element approximations, the Lagrange approximation and the quadratic hierarchical approximation studied in [2], originally suggested in [3]. We use the general theory outlined in [7] to indicate how the point Gauss-Seidel smoother deteriorates as a function of the dimension n of the problem and the degree q of the approximation.

We found only a few references in the literature on multigrid methods for higher degree finite elements. For example in the monograph [7] a general theory is presented although only linear finite elements are considered explicitly.

1.1. **Preliminaries.** We assume the underlying problem is a second order linear elliptic equation on a polyhedral domain $\Omega \subset \mathbf{R}^n$ for $n = 2, 3$. Let $a(\cdot,\cdot) : V \times V \to \mathbf{R}$ be a continuous symmetric V-elliptic bilinear form, and let $f(\cdot) : V \to \mathbf{R}$ be a continuous linear form. We pose the problem in general form and consider the variational formulation

$$u \in V : \quad a(u, v) = f(v) \quad \forall v \in V, \tag{1.1}$$

where we assume that $V \subset H^1(\Omega)$ is a Hilbert space such that (1.1) is well-posed.

For any measurable set $\omega \subseteq \mathbf{R}^n$, $n = 2, 3$, let $|\omega|$ denoted its measure. We will use standard notation for the Lebesgue and Sobolev spaces with corresponding norms

$$\|\cdot\|_{L^2(\omega)} = \|\cdot\|_{0,\omega} \quad \text{and} \quad \|\cdot\|_{H^s(\omega)} = \|\cdot\|_{s,\omega},$$

and when $\omega = \Omega$, and it is clear from the context, we will simplify the notation and write

$$\|\cdot\|_{0,\Omega} = \|\cdot\|_0 \quad \text{and} \quad \|\cdot\|_{s,\Omega} = \|\cdot\|_s,$$

and likewise for the $L^2(\omega)$ scalar product

$$(u, v)_\omega = \int_\omega uv \, dx,$$

see, for example, [1] for more details.

We also use the norm defined by

$$|||v||| = a(v, v)^{1/2} \quad \forall v \in V.$$

For vectors $\tilde{v} = (\tilde{v}_1, \ldots, \tilde{v}_N) \in \mathbf{R}^N$ we will use the Euclidean norm denoted by $\|\tilde{v}\| = (\tilde{v}_1^2 + \tilde{v}_2^2 + \ldots + \tilde{v}_N^2)^{1/2}$.

Finally, throughout this work we will use C and c_i to denote various constants, not necessarily taking the same values from time to time.

1.2. **Finite elements.** We will use the notion *finite element* to denote the triplets $(T, \mathcal{P}, \mathcal{N})$ where $T \subset \Omega$ is a non-empty Lipschitz continuous set, $\mathcal{P}$ is a finite dimensional space of functions on T and $\mathcal{N} = \{N_1, N_2, \ldots, N_{m_q}\}$ is a basis for $\mathcal{P}'$, the set of nodal variables [8, 9].

As for T we only consider n-simplices with vertices $a_i \in \mathbf{R}^n$ for $i = 1, \ldots, n+1$ and $n = 2, 3$ as in Figure 1.1 and 1.2a. We set $h_T = \operatorname{diam}(T)$.

Let $\mathcal{P}_q$ to denote the space of polynomials of degree $\leq q$ and note that

$$\dim \mathcal{P}_q = \binom{n+q}{q} = \operatorname{card}(\mathcal{N}) = m_q, \tag{1.2}$$

where we use the cardinal number to count the number of elements in a set.

Let $L_q(T)$ denote the *principal lattice of order* q on T with m_q lattice points [9, Theorem 6.1, p. 70], that is,

$$L_q(T) = \left\{ x = \sum_{i=1}^{n+1} \xi_i a_i : \sum_{i=1}^{n+1} \xi_i = 1,\ \xi_i \in \left\{ 0, \frac{1}{q}, \ldots, \frac{q-1}{q}, 1 \right\} \right\}$$

For example, $L_1(T) = \{a_i\}_{i=1}^{n+1}$ is the set of vertices of the n-simplex T, and $L_2(T) = \{a_i\}_{i=1}^{n+1} \cup \{a_{ij} = (a_i + a_j)/2 : 1 \leq i < j \leq n+1\}$, see Figures 1.1 and 1.2a.

We use the common practice and refer to points in $L_q(T)$ as *local nodes*.

In order to express $\mathcal{P}_q$ we use *barycentric coordinates* on T, that is, the functions $\lambda_i \in \mathcal{P}_1$ such that $\lambda_i(a_j) = \delta_{ij}$ for $a_j \in L_1(T)$ and $i, j = 1, \ldots, n+1$, see, for example [11].

Given a basis $\{N_1, N_2, \ldots, N_{m_q}\}$ to $\mathcal{P}_q'$ we choose a basis $\{\varphi_1, \varphi_2, \ldots, \varphi_{m_q}\}$ to $\mathcal{P}_q$ so that $N_i(\varphi_j) = \delta_{ij}$ for $i, j = 1, \ldots, m_q$.

Let $(\widehat{T}, \widehat{\mathcal{P}}, \widehat{\mathcal{N}})$ denote the reference finite element where $\widehat{T}$ is either the triangle with vertices in $(0,0), (1,0), (0,1)$ or the tetrahedron with vertices in $(0,0,0), (1,0,0), (0,1,0), (0,0,1)$. We will assume that all finite elements $(T, \mathcal{P}, \mathcal{N})$ are equivalent to the reference finite element. Thus, there is an invertible affine mapping

$$F :\ \mathbf{R}^n \ni x \longmapsto F(x) = Bx + b \in \mathbf{R}^n \tag{1.3}$$

such that $F(\widehat{T}) = T$, $F^*\widehat{\mathcal{P}} = \mathcal{P}$ and $F_*\widehat{\mathcal{N}} = \mathcal{N}$ where F^* and F_* denote the pull-back and push-forward operators, see [8].

1.2.1. *Lagrange finite elements.* We recall the definition of the standard Lagrange finite element which determines a finite element space of continuous piecewise polynomials of degree $q \geq 1$. In terms of the triplet $(T, \mathcal{P}, \mathcal{N})$, $\mathcal{P} = \mathcal{P}_q$ with basis functions $\varphi_i \in \mathcal{P}_q$ for $i = 1, \ldots, m_q$ such that $\varphi_i(x_j) = \delta_{ij}$ and the nodal variables are defined by $N_j(v) = v(x_j)$ for $x_j \in L_q(T)$ and $v \in C^0$. For example: if $q = 1$, $\varphi_i = \lambda_i$, and if $q = 2$, $\varphi_i = \lambda_i(2\lambda_i - 1)$ for $i = 1, \ldots, n+1$, and $\varphi_{ij} = 4\lambda_i\lambda_j$ for $1 \leq i < j \leq n+1$ denoting the last $n+2, \ldots, m_2$ basis functions.

1.2.2. *Higher degree hierarchical finite elements.* We consider the higher degree hierarchical finite element which determines a finite element spaces of continuous piecewise polynomials of degree $q \geq 2$ as outlined in [2]. In terms of the triplet $(T, \mathcal{P}, \mathcal{N})$, $\mathcal{P} = \mathcal{P}_1 \oplus \mathcal{B}_q$ where $\mathcal{B}_q$ is the space of polynomials of degree > 1 and $\leq q$, that is, excluding the linear polynomials. For example: if $q = 2$, we choose $\varphi_i = \lambda_i$ for $i = 1, \ldots, n+1$, and $\varphi_{ij} = 4\lambda_i\lambda_j$ for $1 \leq i < j \leq n+1$ denoting the last $n+2, \ldots, m_q$ basis functions, and the nodal variables are defined by $N_i(v) = v(a_i)$ for $i = 1, \ldots, n+1$, and

$$N_{ij}(v) = v(a_{ij}) - \frac{1}{2}(v(a_i) + v(a_j)) \quad \text{for } 1 \leq i < j \leq n+1.$$

1.3. **The finite element multigrid method.** We use the notation and framework presented in [7]. Let $\mathcal{T}_1$ be a triangulation and define $\mathcal{T}_k$ for $k = 2, \ldots, K$ recursively by subdividing all n-simplices in $\mathcal{T}_{k-1}$. Triangles are subdivided into four congruent sub-triangles connecting the edge midpoints as in Figure 1.1. Tetrahedra are subdivided into eight sub-tetrahedra by the regular refinement algorithm proposed in [5] and as depicted in Figure 1.2. We remark that the all sub-tetrahedra are not congruent but on repeating the process the sub-tetrahedra will remain shape-regular [5]. Hence the family of triangulations $\{T_k\}_{k=1}^K$ will be quasi-uniform.

Set $h_1 = \max_{T \in \mathcal{T}_1} h_T$. It follows that $h_k = 2^{-k+1} h_1$ for $k = 1, \ldots, K$, and for convenience we set $h = h_K$, and recall that the family $\{\mathcal{T}_k\}$ is *quasi-uniform* [9] if there is a constant $\beta > 0$ such that

$$\frac{h}{h_T} \leq \beta \quad \forall T \in \bigcup_k \mathcal{T}_k, \tag{1.4}$$

and if there is a constant $\gamma > 0$ such that

$$\frac{h_T}{\rho_T} \leq \gamma \quad \forall T \in \bigcup_k \mathcal{T}_k, \tag{1.5}$$

where $\rho_T = \sup\{\text{diam}(S) : S \text{ is a ball contained in } T\}$. A family of triangulations satisfying (1.5) is said to be *regular*.

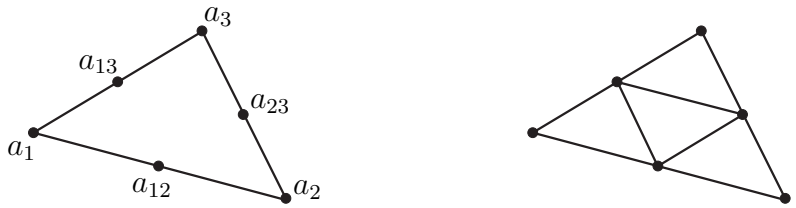


Figure 1.1: Regular triangle refinement. Original and refined triangles.

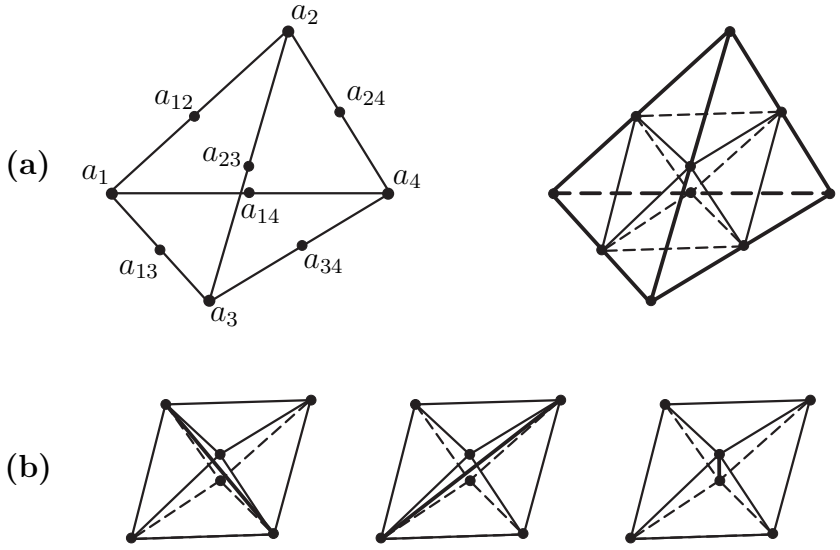


Figure 1.2: Regular tetrahedron refinement due to [5]. **(a)** Original and refined tetrahedron. **(b)** The interior octahedron is divided in one out of three ways as specified in [5].

In the usual way we define the piecewise continuous finite element spaces V_k on Ω by the finite elements $(T, \mathcal{P}_T, \mathcal{N}_T)_{T\in\mathcal{T}_k}$ with local basis functions $\{\varphi_{1,T}, \ldots, \varphi_{m_q,T}\}$ and node variables $\mathcal{N}_T = \{N_{1,T}, N_{2,T}, \ldots, N_{m_q,T}\}$.

Let $\{\phi_1, \ldots, \phi_{M_k}\}$ be a basis for V_k with

$$\dim V_k := M_k = \text{card}\,\{L_q(T) :\ T \in \mathcal{T}_k\}, \tag{1.6}$$

so that ϕ_i has support in S_i for $i = 1, \ldots, M_k$ and where

$$S_i := \bigcup\{T \in \mathcal{T}_k : x_i \in T\}, \tag{1.7}$$

for the *global nodes* $\{x_i\}_{i=1}^{M_k} = \{L_q(T) : T \in \mathcal{T}_k\}$.

For $T \in \mathcal{T}_k$ let I_T be an index set of the local nodes in the finite element $(T, \mathcal{P}_T, \mathcal{N}_T)$, for example, $I_T = \{1, 2, 3, 12, 13, 23\}$ for the quadratic Lagrange finite element in two dimensions. Let $i_j : I_T \to \{1, \ldots, M_k\}$ be the injective map that maps the local index j to the corresponding global index i_j. We express the global basis functions in terms of the local finite element base functions. For $i = 1, \ldots, M_k$ and with j so that $i_j = i$

$$\phi_i\big|_T = \begin{cases} \varphi_{j,T} & \text{if } T \in S_i, \\ 0 & \text{if } T \notin S_i. \end{cases} \tag{1.8}$$

Hence $V_k \ni v = \sum_{i=1}^{M_k} \tilde{v}_i\phi_i$, where $(\tilde{v}_1, \ldots, \tilde{v}_{M_k}) = \tilde{v} \in \mathbf{R}^{M_k}$ is the coordinate vector with respect to the basis $\{\phi_1, \ldots, \phi_{M_k}\}$.

Now $\{V_k\}_{k=1}^K$ is a nested sequence of finite element spaces, that is,

$$V_1 \subset V_2 \subset \cdots \subset V_K \subset V.$$

From equation (1.1) we obtain the finite element equations on the K:th level

$$u \in V_K : \quad a(u, v) = (f, v) \quad \forall v \in V_K, \tag{1.9}$$

where we assume that $f \in V_K$ is a finite element approximation to the linear form $f(\cdot)$ in equation (1.1).

In order to describe the multigrid method we will need the following auxiliary operators. For $k = 1, \ldots, K$ let $A_k : V_k \to V_k$ be defined by

$$(A_k v, \phi) = a(v, \phi) \quad \forall \phi \in V_k,$$

and the projectors $P_{k-1} : V_k \to V_{k-1}$ and $Q_{k-1} : V_k \to V_{k-1}$ defined by

$$a(P_{k-1}v, \phi) = a(v, \phi) \quad \forall \phi \in V_{k-1},$$

and

$$(Q_{k-1}v, \phi) = (v, \phi) \quad \forall \phi \in V_{k-1}.$$

We will also need a generic smoother $R_k : V_k \to V_k$ for $k = 1, \ldots, K$ and denote by R_k^t the adjoint of R_k with respect to $(\cdot, \cdot)$.

By the coercivity of $a(\cdot, \cdot)$ and the inverse inequality we obtain lower and upper bounds to the eigenvalues of A_k,

$$c_1 \|v\|_0^2 \leq (A_k v, v) = a(v, v) \leq c_2 h_k^{-2} \|v\|_0^2 \quad \forall v \in V_k, \tag{1.10}$$

that is, the largest eigenvalue λ_k of A_k is bounded by $c_2 h_k^{-2}$.

We consider the V-cycle multigrid algorithm. Given initial data $u^0 \in V_K$ the algorithm generates a sequence that approximates u, the solution to (1.9), by

$$u^{m+1} = \mathrm{Mg}_K(u^m, f) \quad m = 0, 1, \ldots, \tag{1.11}$$

where $\mathrm{Mg}_K(\cdot, \cdot) : V_K \times V_K \to V_K$ is defined by the following algorithm [7].

Algorithm 1: $\mathrm{Mg}_k(v, f)$

Input: multigrid level k, initial value $v = u^0$ as in (1.11) and right hand side f.

Output: u^1 in (1.11).

if $k = 1$ **then**
  return $A_0^{-1} f$ `/* exact solution */`
else
  $v' = v + R_\ell^t(f - A_k v)$ `/* presmoothing */`
  $v'' = v' + \mathrm{Mg}_{k-1}(0, Q_{k-1}(f - A_k v'))$ `/* error correction */`
  return $v'' + R_k(f - A_k v'')$ `/* postsmoothing */`

If there exists $\omega > 0$ independent of K such that

$$\omega \lambda_k^{-1} \|v\|_0^2 \leq (\overline{R}_k v, v) \quad \forall\, v \in V_k,\ k = 1, \ldots, K, \tag{1.12}$$

where $\overline{R}_k = R_k + R_k^t - R_k^t A_k R_k$ is the symmetrized smoother, and if there is a constant C_P independent of K, such that

$$\|(I - P_{k-1})v\|_0^2 \leq C_P \lambda_k^{-1} (A_k v, v) \quad \forall\, v \in V_k,\ k = 1, \ldots, K, \tag{1.13}$$

then Algorithm 1 converges [6, 7] in the following way

$$|||u - u^m||| \leq \Big(\frac{C_P}{C_P + \omega}\Big)^m |||u - u^0|||. \tag{1.14}$$

We note that the convergence deteriorates when $\omega \downarrow 0$, and in order to achieve good convergence rates it will be fundamental to understand the properties of the smoother and try to make ω as large as possible. Below we will estimate ω for $n = 2, 3$, and $q = 1, 2$, and for different finite elements. This estimate qualitatively explains the poor performance of Algorithm 1 applied to finite element equations based on higher degree basis functions.

We now consider the case when R_k is the point Gauss-Seidel smoother. Decompose the space V_k into subspaces V_k^i spanned by the basis functions ϕ_i, for $i = 1, \ldots, M_k$, that is,

$$V_k = V_k^1 \oplus \cdots \oplus V_k^{M_k}. \tag{1.15}$$

Let κ be the interaction matrix reflecting the coupling between the subspaces V_k^i and defined by

$$\kappa_{ij} = \begin{cases} 0 & \text{if } (A_k v_i, v_j) = 0, \\ 1 & \text{otherwise,} \end{cases} \quad \text{for } v_i \in V_k^i \text{ and } v_j \in V_k^j.$$

If there is a positive number C_1, independent of k, such that

$$\|\kappa\|_2 \leq \|\kappa\|_\infty \leq C_1, \tag{1.16}$$

where $\|\cdot\|$ denotes the appropriate matrix norm, and if there is a positive constant C_2, independent of k, such that

$$\sum_{i=1}^{M_k} \|v_i\|_0^2 \leq C_2 \|v\|_0^2 \quad \text{for } v \in V_k \text{ and } v_i \in V_k^i, \tag{1.17}$$

then (1.12) holds with

$$\omega = (C_2 C_1^2)^{-1}, \tag{1.18}$$

see [7, Theorem 82, p. 277] for a more general statement.

1.3.1. *Estimating C_1.* We note that C_1 is the maximal number of indices j such that $(A_k v_i, v_j) \neq 0$ for $i, j = 1, \ldots, M_q$. It is bounded by

$$C_1 \leq \max_{1 \leq i \leq M_k} \text{card } \{L_q(T) : \ T \in S_i\}, \tag{1.19}$$

where we used the notation in (1.6) and (1.7). The number of global nodes in S_i and hence C_1 will differ quit significantly as n and q varies, see Figure 1.3.

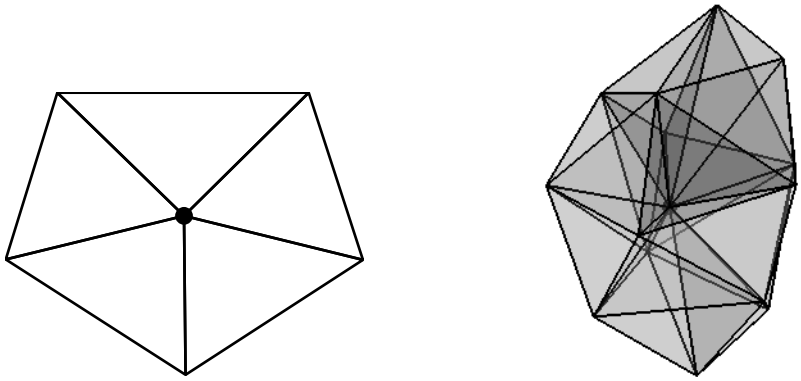


Figure 1.3: Examples of S_i from the triangulations depicted in Figure 2.1. **(left)** For $n = 2$ there are 5 triangles, 6 nodes and 10 edges. **(right)** For $n = 3$ there are 24 tetrahedra, 15 nodes and 50 edges.

1.3.2. *Estimating* C_2. Since V_k is finite dimensional the norms $\|v\|_0$ and $\|\tilde{v}\|$ are equivalent in V_k. In other words we have the following estimates

$$\alpha_1\|v\|_0^2 \le ch_k^n\|\tilde{v}\|^2 \le \alpha_2\|v\|_0^2 \quad \forall v \in V_k, \tag{1.20}$$

for some constants c, α_1 and α_2 that we will estimate below in the case the family of triangulations $\{\mathcal{T}_k\}_{k=1}^K$ is quasi-uniform. With (1.20) we readily verify (1.17) since

$$\sum_{i=1}^{M_k}\|v_i\|_0^2 \le ch_k^n\alpha_1^{-1}\sum_{i=1}^{M_k}\tilde{v}_i^2 \le \frac{\alpha_2}{\alpha_1}\|v\|_0^2,$$

and thus $C_2 = \alpha_2/\alpha_1$.

In order to derive (1.20) we first derive a similar, but local, estimate. For any $T \in \mathcal{T}_k$ we have $v|_T = \sum_{j=1}^{m_q}\tilde{v}_{i_j}\varphi_{j,T}$ and since all finite elements on $T \in \mathcal{T}_k$ are affine equivalent to the reference finite elements we get, by a change of variables,

$$\|v\|_{0,T}^2 = (v,v)_T = |\det B^{-1}|\sum_{j,\ell=1}^{m_q}(\tilde{v}_{i_j}\hat{\varphi}_{j,T}, \tilde{v}_{i_\ell}\hat{\varphi}_{\ell,T})_{\widehat{T}}, \tag{1.21}$$

where $|\det B^{-1}| = |\widehat{T}|/|T|$ and with B as in (1.3).

Let $\underline{\mu}$ and $\overline{\mu}$ denote the smallest and largest eigenvalues to the symmetric and positive definite matrix

$$[\mathcal{M}_{\widehat{T}}]_{j\ell} = |\widehat{T}|(\hat{\varphi}_{j,\widehat{T}}, \hat{\varphi}_{\ell,\widehat{T}})_{\widehat{T}} \quad \text{for } j,\ell = 1,\dots,m_q, \tag{1.22}$$

and estimate (1.21)

$$|T|\underline{\mu}\sum_{j=1}^{m_q}\tilde{v}_{i_j}^2 \le \sum_{j,\ell=1}^{m_q}(\tilde{v}_{i_j}\hat{\varphi}_{j,\widehat{T}}, \tilde{v}_{i_\ell}\hat{\varphi}_{\ell,\widehat{T}})_{\widehat{T}} \le |T|\overline{\mu}\sum_{j=1}^{m_q}\tilde{v}_{i_j}^2 \quad \forall \tilde{v}_{i_j} \in \mathbf{R}.$$

Since the triangulation is quasi-uniform

$$\frac{h_k}{\beta\gamma} \le \frac{h_T}{\gamma} \le \rho_T \le c^{-1/n}|T|^{1/n} \quad \text{and} \quad c^{-1/n}|T|^{1/n} \le h_T \le h_k,$$

where $c^n = \pi/(2n)$, we get

$$(\beta\gamma)^{-n}\underline{\mu}\sum_{j=1}^{m_q}\tilde{v}_{i_j}^2 \le \sum_{j,\ell=1}^{m_q} c h_k^{-n}(\tilde{v}_{i_j}\varphi_{j,T}, \tilde{v}_{i_\ell}\varphi_{\ell,T})_T \le \overline{\mu}\sum_{j=1}^{m_q}\tilde{v}_{i_j}^2 \quad \forall \tilde{v}_{i_j} \in \mathbf{R}$$

Let $\sigma_i = \operatorname{card}(S_i)$, that is, σ_i is the number of n-simplices the global node x_i intersect, and set

$$\underline{\sigma} = \min_{1\le i\le M_k}\sigma_i \quad \text{and} \quad \overline{\sigma} = \max_{1\le i\le M_k}\sigma_i.$$

Now (1.20) follows from the local estimate above by summing over all $T \in \mathcal{T}_k$ and taking into account that one node x_i could appear in several n-simplices which is reflected in the parameter σ_i. Thus

$$(\beta\gamma)^{-n}\underline{\sigma}\,\underline{\mu}\|\tilde{v}\|^2 \le c h_k^{-n}\|v\|_0^2 \le \overline{\sigma}\,\overline{\mu}\|\tilde{v}\|^2,$$

or

$$(\overline{\sigma}\,\overline{\mu})^{-1}\|v\|_0^2 \le c h_k^n\|\tilde{v}\|^2 \le (\beta\gamma)^n(\underline{\sigma}\,\underline{\mu})^{-1}\|v\|_0^2 \quad \forall v \in V_k,$$

where we now identify the constants in (1.20)

$$\alpha_1 = (\overline{\sigma}\,\overline{\mu})^{-1} \quad \text{and} \quad \alpha_2 = (\beta\gamma)^n(\underline{\sigma}\,\underline{\mu})^{-1},$$

and hence

$$C_2 = (\beta\gamma)^n\frac{\overline{\sigma}\,\overline{\mu}}{\underline{\sigma}\,\underline{\mu}}. \tag{1.23}$$

We note the relatively strong dependence of C_2 on β and γ. This implies that the point Gauss-Seidel smoother will deteriorate: (1) if the family of

triangulations $\{\mathcal{T}_k\}_{k=1}^K$ is not quasi-uniform, β increases with k, for example, when $\mathcal{T}_k$ is adaptively refined, or (2) if the $\{\mathcal{T}_k\}_{k=1}^K$ is not regular, γ increases with k, for example, if the refinement algorithm does not preserve the shape-regularity (1.5).

2. Numerical experiments

In matrix form (1.9) becomes

$$\mathcal{A}\tilde{u} = \mathcal{F},$$

where $\mathcal{A}$ denote the matrix $[\mathcal{A}]_{ij} = (A_K\phi_i, \phi_j)$ for $i = 1, \ldots, M_q$ and $\tilde{u} \in \mathbf{R}^{M_q}$ denote the coordinate vector with respect to the finite element basis and $\mathcal{F} = (\mathcal{F}_1, \ldots, \mathcal{F}_{M_q})$ where $\mathcal{F}_i = (f, \phi_i)$. We solve this linear system using the V-cycle Algorithm 1 with $\tilde{u}^0 = 0$ and iterate $m = 1, 2, \ldots$ until the relative residual

$$\text{Res} := \frac{\|\mathcal{F} - \mathcal{A}\tilde{u}^m\|}{\|\mathcal{F}\|}$$

is less than a specified tolerance 'Tol'. In this work we use the Tol $= 10^{-6}$. Note that the relative tolerance times a constant is always greater than $\|u - u^m\|_{1,\Omega}$ where u is the finite element solution we are approximating, cf. [11, Proposition 9.19, p. 393].

By the *work* we mean the number of arithmetic operations required for Algorithm 1 to converge or equally we measure the 'Time' for the algorithm to converge.

In order to examine the optimality of the algorithm we measure the 'Time' for different number of degrees of freedom, 'Dof', and solve the least square problem

$$\text{Time} = a(\text{Dof})^b$$

for the parameters a, b.

Below we exhibit three different numerical experiments that will elucidate the theory outline in the sections above. We use the point Gauss-Seidel smoother in all experiments and we vary $n = 2, 3$ and $q = 1, 2$ for the Lagrange finite element. For $q = 2$ also we compare with the hierarchical finite element in Section 1.2.2.

The experiments are:

- In the first experiment we consider triangulations of the n-unit cube in Figure 2.1 and estimate ω appearing in the convergence estimate (1.14).

- In the second experiment we solve the Poisson equation using the V-cycle multigrid Algorithm 1 and examine the optimality in case of problems with: (1) full regularity and (2) less than full regularly.
- In the third experiment we use the multigrid solver to precondition a Stokes solver and examine the optimality of the Stokes solver for problems with less than full regularity.

2.1. **Estimating w for the point Gauss-Seidel smoother.** In order to indicate how ω in (1.14) varies as a function of $n = 2, 3$ and $q = 1, 2$ and the type of finite element, Lagrange or hierarchical, we estimate C_1 and C_2 for triangulations of the n-unit cube in Figure 2.1 and compute ω from (1.18).

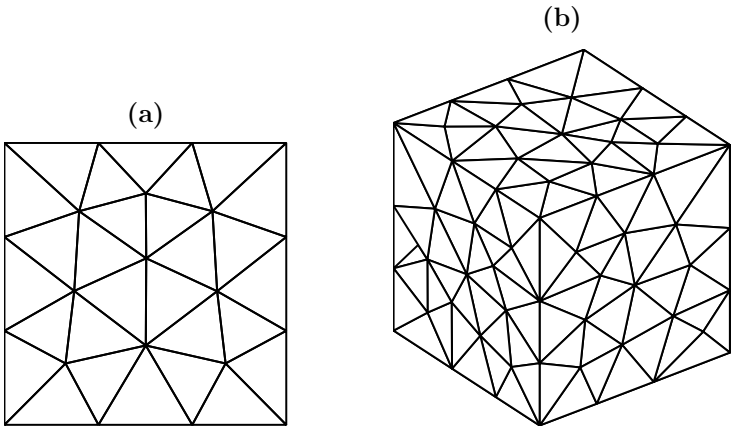


Figure 2.1: Triangulations $\mathcal{T}_0$ of the n-unit cube. **(a)** $n = 2$. **(b)** $n = 3$.

We summarize the results in Table 2.1 and give account for the estimates of C_1 and C_2 in the subsequent sections.

2.1.1. *Estimating C_1.* We estimate C_1 by (1.19) for the triangulations in Figure 2.1. When $q = 1$ we count the number of vertices and when $q = 2$ we count the number of vertices and edges for every S_i. We summarize the results in Table 2.2.

2.1.2. *Estimating C_2.* We estimate C_2 by (1.23) for the triangulations in Figure 2.1. The parameter in (1.23) are computed and the data is gathered in Tables 2.3 and 2.4 and finally we obtain C_2 in Table 2.5.

Table 2.1: ω in (1.14) computed for the triangulations in Figure 2.1. h denote hierarchical finite elements and the remaining data are for Lagrange finite elements.

(n,q)	$(2,1)$	$(2,2)$	$(2,2h)$	$(3,1)$	$(3,2)$	$(3,2h)$
ω^{-1}	$1.0 \cdot 10^4$	$3.3 \cdot 10^5$	$1.1 \cdot 10^6$	$9.0 \cdot 10^7$	$1.3 \cdot 10^{10}$	$4.6 \cdot 10^{10}$

Table 2.2: The maximum number of n-simplices, vertices and edges in S_i for $i = 1, \ldots, M_q$ with respect to i and C_1 for the triangulations in Figure 2.1.

max no. of:	n-simplices	vertices	edges	vertices+edges
$n = 2$	8	8	14	22
$n = 3$	40	23	82	105
(n,q):	(2,1)	(2,2)	(3,1)	(3,2)
C_1	8	22	23	105

Table 2.3: Parameters in (1.23) only depending on n and for the triangulations in Figure 2.1.

	β	γ	$\underline{\sigma}$	$\overline{\sigma}$
$n = 2$	2.0	1.6	2	8
$n = 3$	4.3	3.5	4	40

3. The Poisson equation

We consider the following Poisson equation with mixed Dirichlet-Neumann boundary conditions on bounded polyhedral domains $\Omega \subset \mathbf{R}^n$ for $n = 2, 3$,

$$-\Delta u = f \quad \text{in } \Omega, \qquad u = g \quad \text{on} \quad \partial\Omega_D, \quad \text{and} \quad \nu \cdot \nabla u = 0 \quad \text{on} \quad \partial\Omega_N,$$

where the boundary is partitioned so that $\partial\Omega_D \cup \partial\Omega_N = \partial\Omega$, g is a constant, ν is the outward normal to the boundary and we assume $f \in H^{-1}(\Omega)$ and

Table 2.4: Parameters in (1.23) and for the triangulations in Figure 2.1. h denote hierarchical finite elements and the remaining data are for Lagrange finite elements.

(n, q)	$(2,1)$	$(2,2)$	$(2,2h)$	$(3,1)$	$(3,2)$	$(3,2h)$
$\underline{\mu}$	0.083	0.021	0.011	0.050	0.007	0.004
$\overline{\mu}$	0.333	0.357	0.678	0.250	0.261	0.494
$\overline{\mu}/\underline{\mu}$	4	17	62	5	36	128

Table 2.5: C_2 for the triangulations in Figure 2.1. h denote hierarchical finite elements and the remaining data are for Lagrange finite elements.

(n, q)	$(2,1)$	$(2,2)$	$(2,2h)$	$(3,1)$	$(3,2)$	$(3,2h)$
C_2	164	696	$2.5 \cdot 10^3$	$1.7 \cdot 10^5$	$1.2 \cdot 10^6$	$4.2 \cdot 10^6$

thus the problem is a well posed. Let

$$V = \{u \in H^1(\Omega) : u = 0 \text{ on } \partial\Omega_D\}.$$

Now the bilinear and linear forms in Section 1.1 are

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, dx,$$

and

$$f(v) = \int_\Omega f v \, dx.$$

With $u_g \in H^1(\Omega)$ denoting the extension of g, the weak formulation to the above Poisson problem follows as usual and reads, find $u \in H^1(\Omega)$ such that

$$\begin{aligned} &u = u_g + \phi, \qquad \phi \in V, \\ &a(\phi, v) = f(v) - a(u_g, v) \qquad \forall v \in V. \end{aligned} \tag{3.1}$$

3.0.3. *Model problem I —full regularity.* In this case we let $\Omega = [0,1]^n$ be the n-unit cube depicted in Figure 2.1. Set $f = n\pi^2 \prod_{i=1}^n \sin(\pi x_i)$, $g = 0$ and $\partial\Omega_N = \emptyset$. Since Ω is convex the solution $u \in H^2(\Omega) \cap H_0^1(\Omega)$, that is

full regularity. We note that (1.13) will be satisfied also for higher degree finite elements which could be inferred from the usual duality argument.

Let $w \in H^2(\Omega) \cap H_0^1(\Omega)$ be the solution to the dual problem

$$w \in H_0^1(\Omega) \quad a(w, \phi) = (g, \phi) \quad \forall \phi \in H_0^1(\Omega),$$

where w satisfies the regularity estimate

$$\|w\|_2 \leq C\|g\|_0,$$

and where we have the error estimate

$$\|(I - P_{k-1})w\|_1 \leq Ch_{k-1}\|w\|_2$$

Thus, for $v \in H^2(\Omega) \cap H_0^1(\Omega)$ and taking $g = (I - P_{k-1})v$ and $\phi = (I - P_{k-1})v$ and due to the Galerkin orthogonality and with the above estimates

$$\begin{aligned}
\|(I - P_{k-1})v\|_0^2 &= a(w, (I - P_{k-1})v) \\
&= a((I - P_{k-1})w, (I - P_{k-1})v) \\
&\leq C\|(I - P_{k-1})w\|_1\|(I - P_{k-1})v\|_1 \\
&\leq Ch_{k-1}\|w\|_2\|(I - P_{k-1})v\|_1 \\
&\leq Ch_{k-1}\|(I - P_{k-1})v\|_0\|v\|_1
\end{aligned}$$

and (1.13) follows since $h_{k-1} \leq C\lambda_{k-1}^{-1/2}$ which follows from (1.10).

We solve the problem for different finite element approximations with the V-cycle multigrid Algorithm 1 and for $K = 6$, $n = 2$ and $K = 3$, $n = 3$. The results from these experiments are summarized in Figure 3.1 and Tables 3.1 and 3.2.

3.0.4. *Model problem II —less than full regularity.* In this case we let Ω be the L-shaped domain with one reentrant edge, $\Omega = \{(x, y) \in [0, 2]^2 \setminus [1, 2] \times [0, 1]\}$ for $n = 2$ and $\Omega = \{(x, y, z) \in [0, 2]^2 \times [0, 0.5] \setminus [1, 2] \times [0, 1] \times [0, 0.5]\}$ for $n = 3$, see Figure 3.2. Let $\partial\Omega_D = \partial\Omega_{D_0} \cup \partial\Omega_{D_1}$ where $\partial\Omega_{D_0} = \{(x, y) : x = 1, y \in [1, 2]\}$ and $\partial\Omega_{D_1} = \{(x, y) : x \in [0, 1], y = 0\}$ for $n = 2$ and $\partial\Omega_{D_0} = \{(x, y, z) : x = 2, (y, z) \in [1, 2] \times [0, 0.5]\}$ and $\partial\Omega_{D_1} = \{(x, y, z) : (x, z) \in [0, 1] \times [0, 0.5], y = 0\}$ for $n = 3$. Set $f = 0$, $g = 0$ on $\partial\Omega_{D_0}$ and $g = 1$ on $\partial\Omega_{D_1}$. Since Ω is non-convex we $u \in H^{1+\alpha}(\Omega) \cap H_0^1(\Omega)$ for $0 < \alpha \leq 1$, that is, less than full regularity. The analysis above will not immediately apply, however it is possible to generalize the analysis to include this case [7].

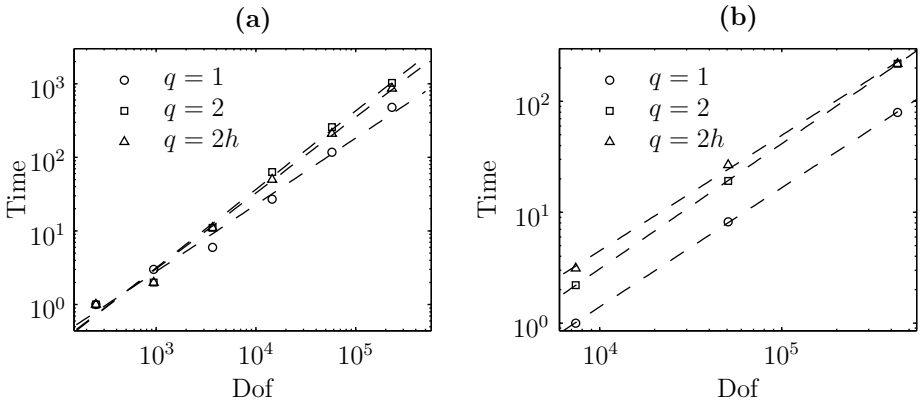


Figure 3.1: Convergence time 'Time' for the V-cycle multigrid Algorithm 1 as a function of 'Dof' and for different finite elements and the triangulations in Figure 2.1. h denote hierarchical finite elements and the remaining data are for Lagrange finite elements **(a)** $n = 2$ **(b)** $n = 3$.

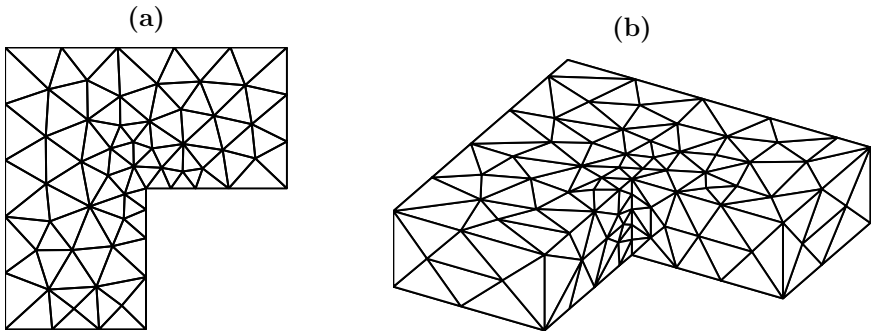


Figure 3.2: Triangulations $\mathcal{T}_0$ of the L-shaped domain. **(a)** $n = 2$. **(b)** $n = 3$

We solve the problem for different finite element approximations with the V-cycle multigrid Algorithm 1 and for $K = 6$, $n = 2$ and $K = 3$, $n = 3$. The results from these experiments are summarized in Figure 3.3 and Tables 3.1 and 3.2.

3.1. **Stokes equations with less than full regularity.** Let Ω be a the polyhedral domains illustrated in Figures 3.4 and 3.5 which we refer to as the Ridge Domain and the Herringbone Domain, respectively. Consider

Table 3.1: Convergence data for the V-cycle multigrid algorithm 1 applied to Model Problem I and II for $n = 2$ and the finite elements in Section 1.2.

Model problem I, $(n, q) = (2, 1)$
$a = 5.4 \cdot 10^{-3}$, $b = 0.91$

k	Dof	m	Res
1	249	4	$3.3 \cdot 10^{-8}$
2	945	4	$1.0 \cdot 10^{-7}$
3	3681	4	$1.6 \cdot 10^{-7}$
4	14529	4	$2.2 \cdot 10^{-7}$
5	57729	4	$2.5 \cdot 10^{-7}$
6	230145	4	$2.8 \cdot 10^{-7}$

Model problem I, $(n, q) = (2, 2)$
$a = 2.0 \cdot 10^{-3}$, $b = 1.07$

k	Dof	m	Res
1	249	4	$2.1 \cdot 10^{-7}$
2	945	4	$6.6 \cdot 10^{-7}$
3	3681	4	$9.3 \cdot 10^{-7}$
4	14529	5	$4.6 \cdot 10^{-8}$
5	57729	5	$5.0 \cdot 10^{-8}$
6	230145	5	$5.2 \cdot 10^{-8}$

Model problem I, $(n, q) = (2, 2h)$
$a = 2.5 \cdot 10^{-3}$, $b = 1.03$

k	Dof	m	Res
1	249	4	$8.0 \cdot 10^{-8}$
2	945	4	$2.3 \cdot 10^{-7}$
3	3681	4	$3.0 \cdot 10^{-7}$
4	14529	4	$3.3 \cdot 10^{-7}$
5	57729	4	$3.4 \cdot 10^{-7}$
6	230145	4	$4.4 \cdot 10^{-7}$

Model problem II, $(n, q) = (2, 1)$
$a = 2.5 \cdot 10^{-3}$, $b = 1.03$

k	Dof	m	Res
1	817	4	$1.4 \cdot 10^{-8}$
2	3169	4	$1.1 \cdot 10^{-7}$
3	12481	4	$4.0 \cdot 10^{-7}$
4	49534	4	$9.5 \cdot 10^{-7}$
5	197377	5	$4.6 \cdot 10^{-8}$
6	787969	5	$8.5 \cdot 10^{-8}$

Model problem II, $(n, q) = (2, 2)$
$a = 3.0 \cdot 10^{-3}$, $b = 0.99$

k	Dof	m	Res
1	817	4	$3.1 \cdot 10^{-7}$
2	3669	5	$4.8 \cdot 10^{-8}$
3	12481	5	$1.3 \cdot 10^{-7}$
4	49534	5	$2.5 \cdot 10^{-7}$
5	197377	5	$4.3 \cdot 10^{-7}$
6	787969	5	$6.6 \cdot 10^{-7}$

Model problem II, $(n, q) = (2, 2h)$
$a = 1.2 \cdot 10^{-3}$, $b = 1.06$

k	Dof	m	Res
1	817	4	$4.4 \cdot 10^{-7}$
2	3169	5	$2.8 \cdot 10^{-8}$
3	12481	5	$4.0 \cdot 10^{-8}$
4	49534	5	$5.7 \cdot 10^{-8}$
5	197377	5	$7.8 \cdot 10^{-8}$
6	787969	5	$1.1 \cdot 10^{-7}$

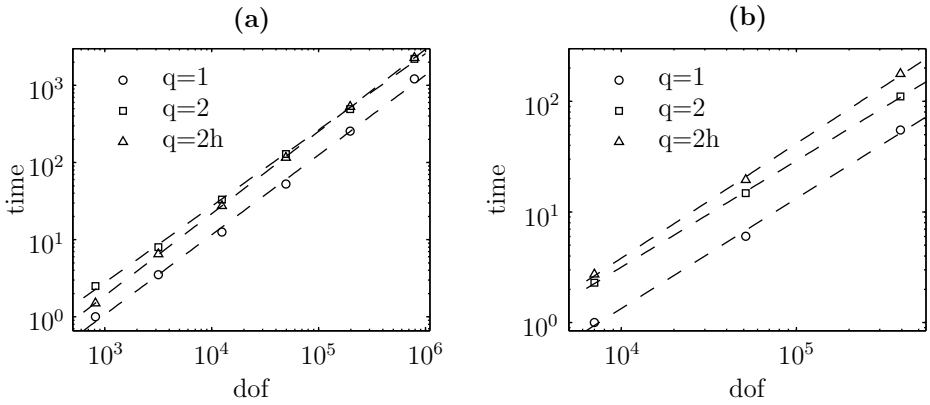


Figure 3.3: Convergence time 'Time' for the V-cycle multigrid Algorithm 1 as a function of 'Dof' and for different finite elements and the triangulations in Figure 3.2. h denote hierarchical finite elements and the remaining data are for Lagrange finite elements **(a)** $n = 2$ **(b)** $n = 3$..

the periodic Stokes problem in dimensionless form

$$
\begin{aligned}
-\Delta u + \nabla p &= 0 \quad \text{in}\,\Omega, \\
\nabla \cdot u &= 0 \quad \text{in}\,\Omega, \\
u &= 0 \quad \text{on}\,\partial\Omega \setminus (\Gamma_A \cup \Gamma_B), \\
u|_{\Gamma_A} &= u|_{\Gamma_B}, \\
p|_{\Gamma_A} &= p|_{\Gamma_B} + R,
\end{aligned}
\tag{3.2}
$$

where u is the unknown velocity field, p is the unknown pressure and R is a constant modelling the pressure drop. We note the this model is inspired by [22] where fluid mixing in micro channels was studied experimentally.

Let

$$V = \{u \in H^1(\Omega)^3 : u = 0 \text{ on } \partial\Omega \setminus (\Gamma_A \cup \Gamma_B) \text{ and } u|_{\Gamma_A} = u|_{\Gamma_B}\}.$$

and $W = L^2(\Omega)/\mathbf{R}$.

Then following the standard procedure, see for example [12, 18], we obtain the weak formulation. Find $(u, p) \in V \times W$ such that

$$a(u, \phi) + b(\phi, p) - b(u, \lambda) = Rl(v) \quad \forall\, (\phi, \lambda) \in V \times W, \tag{3.3}$$

Table 3.2: Convergence data for the V-cycle multigrid algorithm 1 applied to Model Problem I and II for $n = 3$ and the finite elements in Section 1.2.

Model problem I, $(n, q) = (3, 1)$
$a = 7.1 \cdot 10^{-5}$, $b = 1.07$

k	Dof	m	Res
1	7377	4	$2.1 \cdot 10^{-7}$
2	50713	5	$2.1 \cdot 10^{-7}$
3	432961	6	$7.8 \cdot 10^{-7}$

Model problem I, $(n, q) = (3, 2)$
$a = 9.4 \cdot 10^{-5}$, $b = 1.13$

k	Dof	m	Res
1	7377	4	$5.0 \cdot 10^{-8}$
2	50713	5	$2.1 \cdot 10^{-7}$
3	432961	7	$3.6 \cdot 10^{-7}$

Model problem I, $(n, q) = (3, 2h)$
$a = 3.0 \cdot 10^{-4}$, $b = 1.04$

k	Dof	m	Res
1	7377	6	$4.5 \cdot 10^{-7}$
2	50713	6	$9.3 \cdot 10^{-7}$
3	432961	7	$8.7 \cdot 10^{-7}$

Model problem II, $(n, q) = (3, 1)$
$a = 1.4 \cdot 10^{-4}$, $b = 1.0$

k	Dof	m	Res
1	7005	4	$1.2 \cdot 10^{-7}$
2	50713	5	$9.6 \cdot 10^{-8}$
3	393617	6	$1.3 \cdot 10^{-7}$

Model problem II, $(n, q) = (3, 2)$
$a = 4.5 \cdot 10^{-4}$, $b = 0.96$

k	Dof	m	Res
1	7005	4	$1.9 \cdot 10^{-7}$
2	51433	5	$8.0 \cdot 10^{-8}$
3	393617	5	$4.4 \cdot 10^{-7}$

Model problem II, $(n, q) = (3, 2h)$
$a = 2.8 \cdot 10^{-4}$, $b = 1.04$

k	Dof	m	Res
1	7005	6	$5.0 \cdot 10^{-7}$
2	51433	7	$2.9 \cdot 10^{-7}$
3	393617	7	$2.6 \cdot 10^{-7}$

where

$$a(u, \phi) = \int_\Omega \sum_{i,j=1}^{n} \frac{\partial u_i}{\partial x_j} \frac{\partial \phi_i}{\partial x_j} \, dx,$$

$$b(\phi, p) = -\int_\Omega (\nabla \cdot \phi) p \, dx,$$

$$l(v) = \int_{\Gamma_A} v \cdot \nu \, dS.$$

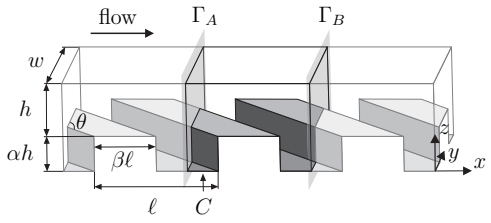


Figure 3.4: Three juxtaposed Ridge Domains. The shaded planes Γ_A and Γ_B are periodic boundaries. We choose the following values for the parameters: $\ell = w = 1$, $h = 0.3$, $\theta = 45°$, $\alpha = 2/3$, $\beta = 0.5$, and the length of the unit cell is $= 1$.

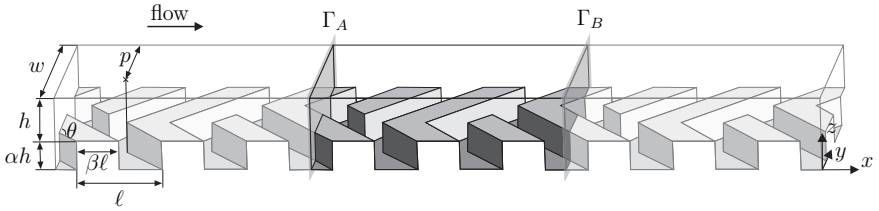


Figure 3.5: Three juxtaposed Herringbone Domains. The shaded planes Γ_A and Γ_B are periodic boundaries. We choose the following values for the parameters: $\ell = 2/3$, $w = 1$, $h = 1/5$, $\theta = 45°$, $\alpha = 2/3$, $\beta = 9/16$, $p = 2/3$, and the length of the unit cell is $= 14/9$.

We discretize (3.3) using the P_2P_1 Taylor-Hood finite elements and obtain the saddle point problem

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} u_h \\ p_h \end{pmatrix} = \begin{pmatrix} l_h \\ 0 \end{pmatrix}, \tag{3.4}$$

for matrices A, B and where ^{T} denotes the transpose. There are many plausible way to solve this problem approximately, by some iterative scheme, see the survey paper [4]. In this work we use the method proposed in [10, 17], for solving the stationary Navier-Stokes equations. The method is optimal and is based on the observation that the matrix

$$\begin{pmatrix} A & B^T \\ 0 & BA^{-1}B^T \end{pmatrix}^{-1} \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}$$

has at most three eigenvalues [19, 15]. Thus, a Krylov method applied to the preconditioned system will converge to the exact solution in less than four iterations.

In practice the matrix in the (2,2) position of the preconditioning block matrix, $BA^{-1}B^T$ (the Schur complement), is not readily inverted but since $BA^{-1}B^T$ is spectrally equivalent to the pressure mass matrix (or Gram matrix) M_p we substitute $BA^{-1}B^T$ by M_p. Hence we precondition (3.4) with

$$\begin{pmatrix} A & B^T \\ 0 & M_p \end{pmatrix}^{-1},$$

and consequently a Krylov solver will now converge in a relatively small number of iterates almost independent of the size of the problem. The method is optimal.

In this work we use a flexible GMRES algorithm [20] to solve (3.4). In the preconditioning we approximate A^{-1} by two cycles of the V-cycle Algorithm 1 with five point Gauss-Seidle smoothing iterations on each level and M_p^{-1} is approximated by a few iterations with the flexible GMRES method preconditioned by five iterations of the point Gauss-Seidle solver. We note that since the saddle point problem is symmetric we could have used a MINRES Krylov solver instead.

In Table 3.3 we summarize the data from the experiments and note that the solver is almost optimal.

Table 3.3: Convergence data for the Stokes solver with V-cycle multigrid preconditioning.

Ridge Domain

levels	dof	m	Res
0	23654	24	$6.4 \cdot 10^{-7}$
1	166599	27	$7.8 \cdot 10^{-7}$
2	1245487	27	$1.0 \cdot 10^{-6}$
3	9621069	28	$9.6 \cdot 10^{-7}$

Herringbone Domain

levels	dof	m	res
0	32999	33	$8.6 \cdot 10^{-7}$
1	232448	37	$8.6 \cdot 10^{-7}$
2	1736817	39	$8.4 \cdot 10^{-7}$

4. Discussion

We have demonstrated that the finite element multigrid method in practice works well for quadratic finite elements. Comparing the method applied to quadratic Lagrange finite element and the quadratic hierarchical finite element showed a convergence in favor of the Lagrange approximation. The estimates of ω seem to be overestimates. However, the estimates are probably qualitatively correct.

Appendix A. Eigenvalues to $\mathcal{M}_{\widehat{T}}$

We give account for the calculation of the eigenvalues to the matrix

$$\mathcal{M}_{\widehat{T}} = (\hat{\varphi}_{j,\widehat{T}}, \hat{\varphi}_{\ell,\widehat{T}})_{\widehat{T}} \quad \text{for } j, \ell = 1, \ldots, m_q,$$

for $n = 2, 3$ and Lagrange finite elements of degree $q = 1, 2$ and for $q = 2$ and the hierarchical base functions in Section 1.2.2. Note that we have omitted the factor $|\widehat{T}| = 1/(2(n-1))$ in (1.22) since in the end we are interested in the ratio $\overline{\mu}/\underline{\mu}$ of the largest to smallest eigenvalues.

We recall and use the following relation [9, eq. (25.14), p. 187]

$$\int_T \lambda_{1,T}^{m_1} \lambda_{2,T}^{m_2} \cdots \lambda_{n+1,T}^{m_{n+1}} \, dx = |T| \frac{m_1! m_2! \cdots m_{n+1}! n!}{(m_1 + m_2 + \ldots + m_{n+1} + n)!}$$

where m_j are positive integers.

q = 1 and n = 2 Lagrange.

$$\mathcal{M}_T = 1/12 \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix},$$

with $\underline{\mu} = 1/24$ and $\overline{\mu} = 1/6$, $\overline{\mu}/\underline{\mu} = 4$.

q = 1 and n = 3 Lagrange.

$$\mathcal{M}_T = 1/20 \begin{pmatrix} 2 & 1 & 1 & 1 \\ 1 & 2 & 1 & 1 \\ 1 & 1 & 2 & 1 \\ 1 & 1 & 1 & 2 \end{pmatrix},$$

with $\underline{\mu} = 1/12$ and $\overline{\mu} = 1/4$, $\overline{\mu}/\underline{\mu} = 5$.

q = 2 and n = 2 Lagrange.

$$\mathcal{M}_T = 1/180 \begin{pmatrix} 6 & -1 & -1 & 0 & -4 & 0 \\ -1 & 6 & -1 & 0 & 0 & -4 \\ -1 & -1 & 6 & -4 & 0 & 0 \\ 0 & 0 & -4 & 32 & 16 & 16 \\ -4 & 0 & 0 & 16 & 32 & 16 \\ 0 & -4 & 0 & 16 & 16 & 32 \end{pmatrix},$$

with $\underline{\mu} = (17 - \sqrt{229}/90) \approx 0.021$ and $\overline{\mu} = (17 + \sqrt{229})/90 \approx 0.357$, $\overline{\mu}/\underline{\mu} \approx 17$.

q = 2h and n = 2 hierarchical.

$$\mathcal{M}_T = 1/180 \begin{pmatrix} 0 & 30 & 30 & 48 & 24 & 48 \\ 30 & 60 & 30 & 48 & 48 & 24 \\ 30 & 30 & 60 & 24 & 48 & 48 \\ 48 & 48 & 24 & 64 & 32 & 32 \\ 24 & 48 & 48 & 32 & 64 & 32 \\ 48 & 24 & 48 & 32 & 32 & 64 \end{pmatrix},$$

with $\underline{\mu} = (31 - \sqrt{901}/90) \approx 0.011$ and $\overline{\mu} = (31 + \sqrt{901})/90 \approx 0.678$, $\overline{\mu}/\underline{\mu} \approx 62$.

q = 2 and n = 3 Lagrange.

$$\mathcal{M}_T = 1/420 \begin{pmatrix} 6 & 1 & 1 & 1 & -4 & -6 & -4 & -4 & -6 & -6 \\ 1 & 6 & 1 & 1 & -4 & -4 & -6 & -6 & -4 & -6 \\ 1 & 1 & 6 & 1 & -6 & -4 & -4 & -6 & -6 & -4 \\ 1 & 1 & 1 & 6 & -6 & -6 & -6 & -4 & -4 & -4 \\ -4 & -4 & -6 & -6 & 32 & 16 & 16 & 16 & 16 & 8 \\ -6 & -4 & -4 & -6 & 16 & 32 & 16 & 8 & 16 & 16 \\ -4 & -6 & -4 & -6 & 16 & 16 & 32 & 16 & 8 & 16 \\ -4 & -6 & -6 & -4 & 16 & 8 & 16 & 32 & 16 & 16 \\ -6 & -4 & -6 & -4 & 16 & 16 & 8 & 16 & 32 & 16 \\ -6 & -6 & -4 & -4 & 8 & 16 & 16 & 16 & 16 & 32 \end{pmatrix},$$

with $\underline{\mu} = (113 - 5\sqrt{457}/840) \approx 0.007$ and $\overline{\mu} = (113 + 5\sqrt{457}/840) \approx 0.261$, $\overline{\mu}/\underline{\mu} \approx 36$.

q = 2h and n = 3 hierarchical.

$$\mathcal{M}_T = 1/420 \begin{pmatrix} 42 & 21 & 21 & 21 & 28 & 14 & 28 & 28 & 14 & 14 \\ 21 & 42 & 21 & 21 & 28 & 28 & 14 & 14 & 28 & 14 \\ 21 & 21 & 42 & 21 & 14 & 28 & 28 & 14 & 14 & 28 \\ 21 & 21 & 21 & 42 & 14 & 14 & 14 & 28 & 28 & 28 \\ 28 & 28 & 14 & 14 & 32 & 16 & 16 & 16 & 16 & 8 \\ 14 & 28 & 28 & 14 & 16 & 32 & 16 & 8 & 16 & 16 \\ 28 & 14 & 28 & 14 & 16 & 16 & 32 & 16 & 8 & 16 \\ 28 & 14 & 14 & 28 & 16 & 8 & 16 & 32 & 16 & 16 \\ 14 & 28 & 14 & 28 & 16 & 16 & 8 & 16 & 32 & 16 \\ 14 & 14 & 28 & 28 & 8 & 16 & 16 & 16 & 16 & 32 \end{pmatrix},$$

with $\underline{\mu} = (209 - \sqrt{42337}/840) \approx 0.004$ and $\overline{\mu} = (209 + \sqrt{42337}/840) \approx 0.494$, $\overline{\mu}/\underline{\mu} \approx 128$.

References

[1] R. A. Adams, *Sobolev spaces*, Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1975.

[2] O. Axelsson and I. Gustafsson, *Preconditioning and two-level multigrid methods of arbitrary degree of approximation*, Math. Comp. **40** (1983), 219–242.

[3] R. E. Bank, T. F. Dupont, and H. Yserentant, *The hierarchical basis multigrid method*, Numer. Math. **52** (1988), 427–458.

[4] M. Benzi, G. H. Golub, and J. Liesen, *Numerical solution of saddle point problems*, Acta Numer. **14** (2005), 1–137.

[5] J. Bey, *Tetrahedral grid refinement*, Computing **55** (1995), 355–378.

[6] D. Braess and W. Hackbusch, *A new convergence proof for the multigrid method including the V-cycle*, SIAM J. Numer. Anal. **20** (1983), 967–975.

[7] J. H. Bramble and X. Zhang, *The Analysis of Multigrid Methods*, Handbook of numerical analysis, Vol. VII, North-Holland, 2000.

[8] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, second ed., Springer-Verlag, 2002.

[9] P. G. Ciarlet, *Basic error estimates for elliptic problems*, Handbook of Numerical Analysis, Vol. II, North-Holland, 1991.

[10] H. C. Elman, *Preconditioning for the steady-state Navier-Stokes equations with low viscosity*, SIAM J. Sci. Comput. **20** (1999), 1299–1316.

[11] A. Ern and J-L. Guermond, *Theory and Practice of Finite Elements*, Springer-Verlag, 2004.

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

[13] W. Hackbusch, *Multigrid Methods and Applications*, Springer-Verlag, 1985.

[14] J. J. Heys, T. A. Manteuffel, S. F. McCormick, and L. N. Olson, *Algebraic multigrid for higher-order finite elements*, J. Comput. Phys. **204** (2005), 520–532.
[15] I. C. F. Ipsen, *A note on preconditioning nonsymmetric matrices*, SIAM J. Sci. Comput. **23** (2001), 1050–1051.
[16] V. John, *Higher order finite element methods and multigrid solvers in a benchmark problem for the 3D Navier-Stokes equations*, Internat. J. Numer. Methods Fluids **40** (2002), 775–798.
[17] D. Kay, D. Loghin, and A. Wathen, *A preconditioner for the steady-state Navier-Stokes equations*, SIAM J. Sci. Comput. **24** (2002), 237–256.
[18] M. Marion and R. Temam, *Navier-Stokes Equations: Theory and Approximation*, Handbook of numerical analysis, Vol. VI, North-Holland, 1998.
[19] M. F. Murphy, G. H. Golub, and A. J. Wathen, *A note on preconditioning for indefinite linear systems*, SIAM J. Sci. Comput. **21** (2000), 1969–1972.
[20] Y. Saad, *A flexible inner-outer preconditioned GMRES algorithm*, SIAM J. Sci. Comput. **14** (1993), 461–469.
[21] V. V. Shaĭdurov, *Multigrid Methods for Finite Elements*, Kluwer Academic Publishers Group, 1995.
[22] A.D. Stroock, S.K.W. Dertinger, A. Ajdari, I. Mezic, H.A. Stone, and G.M. Whitesides, *Chaotic mixer for microchannels*, Science **295** (2002), 647–651.

Department of Mathematical Sciences, Chalmers University of Technology, SE-412 96 Göteborg, Sweden
E-mail address: erik.svensson@math.chalmers.se

Paper VI

MULTIGRID METHODS ON ADAPTIVELY REFINED TRIANGULATIONS: PRACTICAL CONSIDERATIONS

ERIK D. SVENSSON

Abstract. We outline the implementation of the finite element multigrid method on adaptively refined triangulations for Lagrange and hierarchical finite elements of degree ≤ 2 in two and three dimensions. Refining the triangulations we relax the requirement that no vertex of any n-simplex lies in the interior of an edge of another n-simplex. As a result the refinements are easy to implement and the finite element spaces can be made nested, which simplifies the multigrid implementation. The refined triangulations may however contain 'hanging' nodes which must be taken into account in order to make the finite element spaces conforming. We modify the finite elements accordingly in these situations.

1. Introduction

The finite element multigrid method is theoretically well established as outlined in for example [4, 10, 14]. In this work we consider the practical aspects implementing the method on adaptively refined triangulations for conforming linear and quadratic finite elements in two and three dimensions.

We choose to use a refinement method that produce triangulations on which we can define nested finite element spaces and thus makes the formulation of the multigrid method straight forward with well defined projection operators on the finite element spaces. This is in contrast to the situation when the finite elements spaces are non-nested [4, 13]. Moreover this choice is also motivated by the fact that the refinement algorithm becomes simple compared to the rather involved refinement algorithm proposed in [3], which also renders the finite element spaces non-nested.

Date: April 18, 2006.

2000 *Mathematics Subject Classification.* 65N55, 65N30, 65N50.

Key words and phrases. multigrid, finite elements, refinement.

The refined triangulations are irregular [8] in the sense that there will be 'hanging' nodes and the construction of conforming finite element spaces is a non-trivial task that in practice requires implementation of flexible data structures. This and the even more general aspect of *hp* refinements has already be considered in [1, 8, 12]. We partially reformulate these results using concepts from modern finite element theory.

1.1. **Preliminaries.** We assume that the underlying problem is second order linear elliptic on a polyhedral domain $\Omega \subset \mathbf{R}^n$ for $n = 2, 3$. Let $a(\cdot,\cdot) : V \times V \to \mathbf{R}$ be a continuous, symmetric and V-elliptic bilinear form, and let $f(\cdot) : V \to \mathbf{R}$ be a continuous linear form. We pose the problem as the variational formulation

$$u \in V : \quad a(u,v) = f(v) \quad \forall v \in V, \tag{1.1}$$

where we assume that $V \subset H^1(\Omega)$ is a Hilbert space such that (1.1) is well-posed.

We will use standard notation for the Lebesgue and Sobolev spaces and for any measurable set $\omega \subseteq \mathbf{R}^n$, $n = 2, 3$, we let

$$(u,v)_\omega = \int_\omega uv \, dx,$$

denote the $L^2(\omega)$ scalar product.

For vectors $\tilde{v} = (\tilde{v}_1, \ldots, \tilde{v}_N) \in \mathbf{R}^N$ we will use the Euclidean norm denoted by $\|\tilde{v}\| = (\tilde{v}_1^2 + \tilde{v}_2^2 + \ldots + \tilde{v}_N^2)^{1/2}$.

Finally, throughout this work we will use C and c_i to denote various constants, not necessarily taking the same value from time to time.

1.2. **Finite elements.** We will use the notion *finite element* to denote the triplet $(T, \mathcal{P}, \mathcal{N})$ where $T \subset \Omega$ is a non empty Lipschitz continuous set, $\mathcal{P}$ is a finite dimensional space of functions on T and $\mathcal{N} = \{N_1, N_2, \ldots, N_{m_q}\}$ is a base for $\mathcal{P}'$, the set of nodal variables [5, 6].

Remark 1.1. For a d-dimensional vector space $\mathcal{P}$ and for a subset $\{N_1, N_2, \ldots, N_d\}$ of $\mathcal{P}'$ the following two statements are equivalent [5, Lemma 3.1.4, p. 70].

(1) $\{N_1, N_2, \ldots, N_d\}$ is a basis for $\mathcal{P}'$.
(2) If $v \in \mathcal{P}$ with $N_i v = 0$ for $i = 1, \ldots, d$, then $v = 0$.

We use this to verify that a given triplet $(T, \mathcal{P}, \mathcal{N})$ is a finite element.

As for T we only consider n-simplices with vertices $a_i \in \mathbf{R}^n$ for $i = 1, \ldots, n+1$ and $n = 2, 3$ as in Figure 2.1 and 2.2a. We set $h_T = \operatorname{diam}(T)$.

Let $\mathcal{P}_q$ denote the space of polynomials of degree $\leq q$ and note that

$$\dim(\mathcal{P}_q) = \binom{n+q}{q} = \operatorname{card}(\mathcal{N}) = m_q, \tag{1.2}$$

where we use the cardinal number to count the number of elements in a set.

Let $L_q(T)$ denote the *principal lattice of order* q on T with m_q lattice points [6, Theorem 6.1, p. 70], that is,

$$L_q(T) = \Big\{ x = \sum_{i=1}^{n+1} \xi_i a_i : \sum_{i=1}^{n+1} \xi_i = 1,\ \xi_i \in \Big\{ 0, \frac{1}{q}, \ldots, \frac{q-1}{q}, 1 \Big\} \Big\}.$$

For example, $L_1(T) = \{a_i\}_{i=1}^{n+1}$ is the vertices in the n-simplex T and $L_2(T) = \{a_i\}_{i=1}^{n+1} \cup \{a_{ij} = (a_i + a_j)/2 : 1 \leq i < j \leq n+1\}$, see Figures 2.1 and 2.2a.

We use the common practice and refer to points in $L_q(T)$ as *local nodes*.

In order to express $\mathcal{P}_q$ on T we use *barycentric coordinates*, that is, $\lambda_i \in \mathcal{P}_1$ on T such that $\lambda_i(x_j) = \delta_{ij}$ for $x_j \in L_1(T)$ and $i, j = 1, \ldots, n+1$, see for example [9].

Given a basis to $\{\varphi_1, \ldots, \varphi_{m_q}\}$ to $\mathcal{P}_q$ we choose the nodal variables such that $N_i(\varphi_j) = \delta_{ij}$ for $i, j = 1, \ldots, m_q$.

1.2.1. *Lagrange finite elements.* We recall the definition of the standard Lagrange finite element which determine a finite element space of continuous piecewise polynomials of degree $q \geq 1$. In terms of the triplet $(T, \mathcal{P}, \mathcal{N})$, $\mathcal{P} = \mathcal{P}_q$ with basis functions $\varphi_i \in \mathcal{P}_q$ for $i = 1, \ldots, m_q$ such that $\varphi_i(x_j) = \delta_{ij}$ and the nodal variables are defined by $N_j(v) = v(x_j)$ for $x_j \in L_q(T)$ and $v \in C^0(T)$. For example: if $q = 1$, $\varphi_i = \lambda_i$, and if $q = 2$, $\varphi_i = \lambda_i(2\lambda_i - 1)$ for $i = 1, \ldots, n+1$ and $\varphi_{ij} = 4\lambda_i\lambda_j$ for $1 \leq i < j \leq n+1$ denoting the last $n+2, \ldots, m_2$ basis functions.

It is easily verified by Remark 1.1 that the triplet $(T, \mathcal{P}, \mathcal{N})$ is a finite element.

1.2.2. *Higher degree hierarchical finite elements.* We consider the higher degree hierarchical finite element which determine a finite element spaces of continuous piecewise polynomials of degree $q \geq 2$ as outlined in [2]. In terms of the triplet $(T, \mathcal{P}, \mathcal{N})$, $\mathcal{P} = \mathcal{P}_1 \oplus \mathcal{B}_q$, where $\mathcal{B}_q$ is the space of

polynomials of degree > 1 and $\leq q$, that is, excluding the linear functions. For example, if $q = 2$, we choose the basis functions $\varphi_i = \lambda_i$ for $i = 1, \ldots, n+1$ and $\varphi_{ij} = 4\lambda_i\lambda_j$ for $1 \leq i < j \leq n+1$ denoting the last $n+2, \ldots, m_q$ basis functions and the nodal variables are defined by $N_i(v) = v(a_i)$ for $i = 1, \ldots, n+1$ and

$$N_{ij}(v) = v(a_{ij}) - \frac{1}{2}(v(a_i) + v(a_j)) \quad \text{for } 1 \leq i < j \leq n+1.$$

In order to show that the triplet $(T, \mathcal{P}, \mathcal{N})$ is a finite element we take

$$\mathcal{P}_2 \ni v = \sum_{i=1}^{n+1} \tilde{v}_i \varphi_i + \sum_{\substack{i,j=1 \\ i<j}}^{n+1} \tilde{v}_{ij} \varphi_{ij}, \tag{1.3}$$

for constants $\tilde{v}_i, \tilde{v}_{ij} \in \mathbf{R}$. Then for $i = 1, \ldots, n+1$,

$$N_i(v) = 0 \quad \Rightarrow \quad \tilde{v}_i = 0$$

and for $1 \leq i < j \leq n+1$

$$N_{ij}(v) = 0 \quad \Rightarrow \quad \frac{1}{2}\tilde{v}_i + \frac{1}{2}\tilde{v}_j + \tilde{v}_{ij} - \frac{1}{2}(\tilde{v}_i + \tilde{v}_j) = \tilde{v}_{ij} = 0.$$

Thus $v = 0$ and from Remark 1.1 we conclude that $\{N_i\}_{i=1}^{n+1} \cup \{N_{ij} : 1 \leq i < j \leq n+1\}$ is a basis for $\mathcal{P}'$ and $(T, \mathcal{P}, \mathcal{N})$ is a finite element.

1.3. **The finite element multigrid method.** We use the notation and framework outlined in [4]. Let $\mathcal{T}_1$ be a triangulation and define $\mathcal{T}_\ell$ for $\ell = 2, \ldots, L$ recursively by subdividing all n-simplices in $\mathcal{T}_{\ell-1}$ as described in Section 2.1 below. We remark that all sub-tetrahedra are not congruent but on repeating the process the sub-tetrahedra will remain shape regular [3]. We note that since the n-simplices in $\mathcal{T}_\ell$ stay shape regular, the family of triangulations $\{T_\ell\}_{\ell=1}^L$ will be quasi-uniform.

In the usual way we define the piecewise continuous finite element spaces V_ℓ on Ω by the finite elements $(T, \mathcal{P}_T, \mathcal{N}_T)_{T \in \mathcal{T}_\ell}$ with local basis functions $\{\varphi_{1,T}, \ldots, \varphi_{m_q,T}\}$ and node variables $\mathcal{N}_T = \{N_{1,T}, N_{2,T}, \ldots, N_{m_q,T}\}$.

Let $\{\phi_1, \ldots, \phi_{M_\ell}\}$ be a basis to V_ℓ, the *global basis*, with

$$\dim(V_\ell) := M_\ell = \operatorname{card}(\{L_q(T) : \ T \in \mathcal{T}_\ell\}), \tag{1.4}$$

and such that ϕ_i has support in S_i for $i = 1, \ldots, M_\ell$ where

$$S_i := \bigcup \{T \in \mathcal{T}_\ell : x_i \in T\}, \tag{1.5}$$

for the *global nodes* $\{x_i\}_{i=1}^{M_\ell} = \{L_q(T) : T \in \mathcal{T}_\ell\}$.

For $T \in \mathcal{T}_\ell$ let I_T be an index set of the local nodes in the finite element $(T, \mathcal{P}_T, \mathcal{N}_T)$, for example, $I_T = \{1, 2, 3, 12, 13, 23\}$ for the quadratic Lagrange finite element in two dimensions. Let $i_j : I_T \to [1, M_\ell]$ be the injective map that maps the local index j to the corresponding global index i_j. We express the global basis functions in terms of the local finite element basis functions. For $i = 1, \ldots, M_\ell$ and with j so that $i_j = i$

$$\phi_i\big|_T = \begin{cases} \varphi_{j,T} & \text{if } T \in S_i, \\ 0 & \text{if } T \notin S_i. \end{cases} \tag{1.6}$$

Now $\{V_\ell\}_{\ell=1}^L$ is a nested sequence of finite element spaces, that is,

$$V_1 \subset V_2 \subset \cdots \subset V_L \subset V. \tag{1.7}$$

From equation (1.1) we obtain the finite element equation on the L:th level

$$u \in V_L : \quad a(u, v) = (f, v) \quad \forall v \in V_L, \tag{1.8}$$

where we assume that $f \in V_L$ is a finite element approximation to the linear form $f(\cdot)$ in equation (1.1).

In order to describe the multigrid method we will need the following auxiliary operators. For $\ell = 1, \ldots, L$ let $A_\ell : V_\ell \to V_\ell$ be defined by

$$(A_\ell v, \phi) = a(v, \phi) \quad \forall \phi \in V_\ell,$$

and let the projectors $P_{\ell-1} : V_\ell \to V_{\ell-1}$ and $Q_{\ell-1} : V_\ell \to V_{\ell-1}$ be defined by

$$a(P_{\ell-1} v, \phi) = a(v, \phi) \quad \forall \phi \in V_{\ell-1},$$

and

$$(Q_{\ell-1} v, \phi) = (v, \phi) \quad \forall \phi \in V_{\ell-1}.$$

We will also need a generic smoother $R_\ell : V_\ell \to V_\ell$ for $\ell = 1, \ldots, L$ and denote by R_ℓ^t the adjoint of R_ℓ with respect to $(\cdot, \cdot)$.

We consider the V-cycle multigrid algorithm. Given initial data $u^0 \in V_L$ the algorithm generates a sequence approximating u, the solution to equation (1.8), by

$$u^{m+1} = \mathrm{VMG}_L(u^m, f) \quad m = 0, 1, \ldots, \tag{1.9}$$

where $\mathrm{VMG}_L(\cdot, \cdot) : V_L \times V_L \to V_L$ is defined by the following Algorithm 1 [4].

Algorithm 1: $\mathrm{VMG}_\ell(v, f)$

Input: multigrid level ℓ, initial value $v = u^0$ as in (1.9) and right hand side f.
Output: u^1 in (1.9).

if $\ell = 0$ **then**
  return $A_0^{-1} f$ `/* exact solution */`
else
  $v' = v + R_\ell^t(f - A_\ell v)$ `/* presmoothing */`
  $v'' = v' + \mathrm{VMG}_{\ell-1}(0, Q_{\ell-1}(f - A_\ell v'))$ `/* error correction */`
  return $v'' + R_\ell(f - A_\ell v'')$ `/* postsmoothing */`

2. Irregular notions

We now relax one of the requirements in the definition of the triangulation, namely, the property that *no vertex of any n-simplex lies in the interior of an edge or face of another n-simplex* [5]. As for conforming finite element spaces we then will have to modify the finite elements accordingly.

2.1. Irregular triangulations.

Inspired by [8] we say that a *1-irregular triangulation* is a partition of Ω into n-simplices such that at the most one vertex lies in the interior of any edge of all the n-simplices in the partition. Moreover, we say that a vertex is an *irregular vertex* if it lies in the interior of another edge and that an n-simplex is an *irregular n-simplex* if it contains irregular vertices.

We note that the definition is readily generalized to *m-irregular triangulation*, for $m \geq 1$, but we will not consider this type of triangulation in this work. When we in the sequel sometimes write irregular we thus means 1-irregular.

For an irregular n-simplex there may be one or several local nodes in $L_q(T)$ that do not represent true degrees of freedom, since the nodal variables in these points will be evaluated in other, possibly global, nodes in such way that the finite element space is made conforming. In the literature this kind of nodes are called hanging, constrained or slaved, in this work we call them *hanging nodes*, see Figures 3.1 and 3.2.

Now let $\mathcal{T}_\ell$ be an 1-irregular triangulation and suppose S is a subset of n-simplices $T \in \mathcal{T}_\ell$ that we want to refine, for example, S could be

the set of n-simplices where an error estimator is larger than a certain threshold. In order to refine $\mathcal{T}_\ell$ we need to check the consistency of S, that is, the refined triangulation $\mathcal{T}_{\ell+1}$ must also be a 1-irregular triangulation and for this reason we cannot refine irregular n-simplices. We must check and modify S by adding n-simplices intersecting irregular vertices. Since some of the added n-simplices may also be irregular we must repeat the checking on the added n-simplices recursively. We describe this procedure in Algorithm 2.

Algorithm 2: CheckConsistency($\mathcal{T}$,S)

Input: a 1-irregular triangulation $\mathcal{T}$ and a set S of n-simplices $T \in \mathcal{T}$.

Output: S, possibly modified.

```
S_new = ∅
forall irregular T ∈ S do
    forall irregular vertices a_i ∈ T do
        forall T' ∈ T such that a_i ∈ T' do
            S_new = S_new ∪ {T'}
if S_new ≠ ∅ then
    CheckConsistency(T,S_new)
S = S ∪ S_new
```

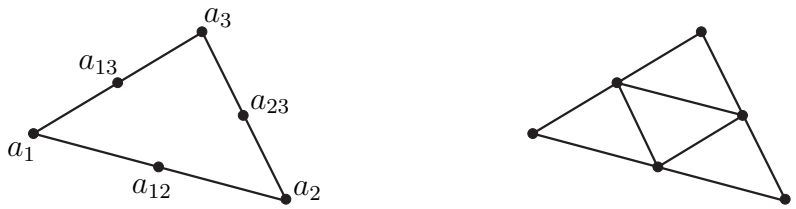


Figure 2.1: Regular triangle refinement. Original and refined triangles.

When we have checked the consistency of S we proceed with the refinement of all $T \in S$ and hence create the refined triangulation $\mathcal{T}_{\ell+1}$. We use the regular refinement algorithm from [3], where the two-dimensional case is trivial but included here for completeness. For any n-simplex let $a_{ij} = (a_i + a_j)/2$ for $1 \leq i < j \leq n+1$ denote the midpoint of the

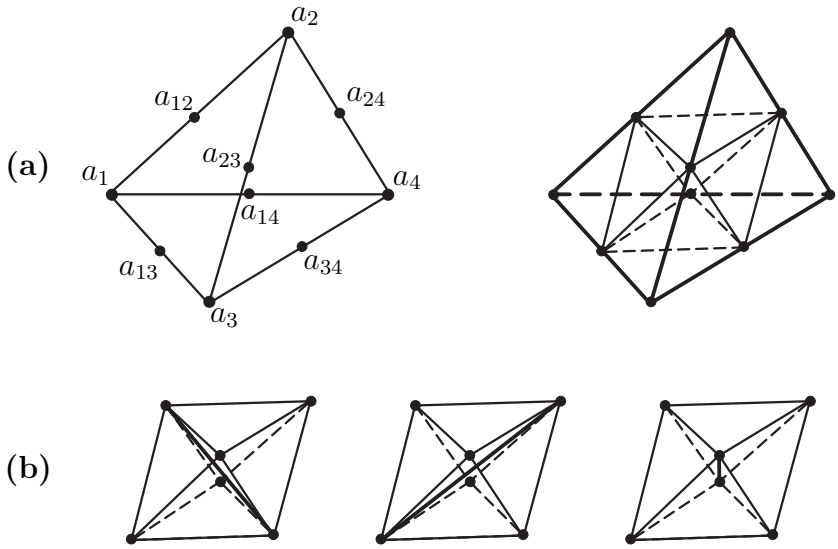


Figure 2.2: Regular tetrahedron refinement due to [3]. **(a)** Original and refined tetrahedron. **(b)** The interior octahedron is divided in one out of three ways as specified in [3].

edge connecting the vertices a_i and a_j. Now triangles are subdivided into four congruent subtriangles connecting the edge midpoints as in Figure 2.1 and as described in Algorithm 3. Tetrahedra are subdivided into eight subtetrahedra as depicted in Figure 2.2 and as described in Algorithm 4. We remark that all subtetrahedra are not congruent but on repeating the procedure the subtetrahedra will stay shape-regular [3].

Algorithm 3: RegularRefinement2D(T)

Input: a triangle T.
Output: 4 subtriangles $T_i \subset T$ for $i = 1, \ldots, 4$ such that $\bigcup_i T_i = T$.

divide $T = \{a_1, a_2, a_3\}$ into 4 subtriangles

$$T_1 = \{a_1, a_{12}, a_{13}\}, \quad T_2 = \{a_2, a_{23}, a_{12}\},$$
$$T_3 = \{a_3, a_{13}, a_{23}\}, \quad T_4 = \{a_{12}, a_{23}, a_{13}\}.$$

Algorithm 4: RegularRefinement3D(T)

Input: a tetrahedron T.
Output: 8 tetrahedra $T_i \subset T$ for $i = 1, \ldots, 8$ such that $\bigcup_i T_i = T$.

divide $T = \{a_1, a_2, a_3, a_4\}$ into 8 subtetrahedra

$$\begin{aligned} T_1 &= \{a_1, a_{12}, a_{13}, a_{14}\}, \quad T_2 = \{a_{12}, a_2, a_{23}, a_{24}\}, \\ T_3 &= \{a_{13}, a_{23}, a_3, a_{34}\}, \quad T_4 = \{a_{14}, a_{24}, a_{34}, a_4\}, \\ T_5 &= \{a_{12}, a_{13}, a_{14}, a_{24}\}, \quad T_6 = \{a_{12}, a_{23}, a_{23}, a_{24}\}, \\ T_7 &= \{a_{13}, a_{14}, a_{24}, a_{34}\}, \quad T_2 = \{a_{13}, a_{23}, a_{34}, a_{34}\}. \end{aligned}$$

2.2. **Irregular finite elements.** In order to construct a conforming finite element space from the finite elements $(T, \mathcal{P}_T, \mathcal{N}_T)_{T \in \mathcal{T}_\ell}$ where $\mathcal{T}_\ell$ is a 1-irregular triangulation we need to define a new type of finite elements on irregular n-simplices.

We say that a finite element is a *q-irregular finite element* if we evaluate one or more of the nodal variables N_i at points $x_j \in L_q(T) \pm p$ where $p = a_i - a_j$ such that the line between a_i and a_j is an edge in T. For irregular n-simplices we define q-irregular finite elements so that the generated finite element space becomes conforming. We describe this in a few examples below.

2.2.1. *1-irregular Lagrange finite elements.* In $\mathbf{R}^2$ there are finite elements with 1–3 hanging nodes as in Figure 2.3. The basis functions are as defined in Sections 1.2.1 but the nodal variables are slightly different. We consider the finite element in the case of one hanging node, the other cases are defined in the same way. The nodal variables are defined by

$$N_i(v) = v(a_i), \quad i = 1, 2, \tag{2.1}$$

and

$$N_3(v) = \frac{1}{2}(v(a_2) + v(a_4)), \tag{2.2}$$

where $a_4 = 2a_3 - a_2$ and N_3 is eliminated as a global degree of freedom.

In order to show that the triplet $(T, \mathcal{P}_1, \mathcal{N})$ is a finite element we take $\mathcal{P}_1 \ni v = \sum_{i=1}^{3} \tilde{v}_i \varphi_i$ for constants $\tilde{v}_i \in \mathbf{R}$. Then for $i = 1, 2$,

$$N_i(v) = 0 \quad \Rightarrow \quad \tilde{v}_i = 0$$

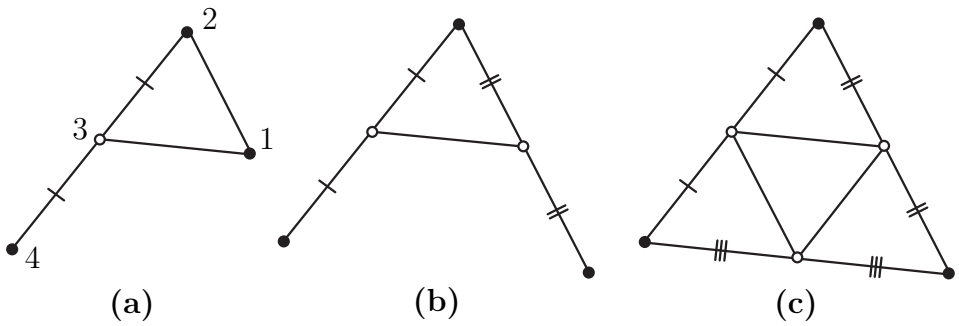


Figure 2.3: Three types of finite elements for the 1-irregular Lagrange finite elements in $\mathbf{R}^2$. • denotes a regular node and ∘ denotes a hanging node. **(a)** One hanging node. **(b)** Two hanging nodes. **(c)** Three hanging nodes.

and

$$N_3(v) = 0 \quad \Rightarrow \quad 1/2(\tilde{v}_2 - \tilde{v}_2 + 2\tilde{v}_3) = \tilde{v}_3 = 0.$$

Thus $v = 0$ and from Remark 1.1 we conclude that $\{N_i\}_{i=1}^3$ is a basis for $\mathcal{P}_1'$ and $(T, \mathcal{P}_1, \mathcal{N})$ is a finite element. Note that this construction of $\mathcal{N}$ guarantees that the global finite element functions are conforming.

In $\mathbf{R}^3$ there are finite elements with 1–4 hanging nodes and the treatment is analogous to the $\mathbf{R}^2$ case. In the case of one hanging node as in Figure 2.4 the nodal variables are defined by

$$N_i(v) = v(a_i), \quad i = 1, 2, 3, \tag{2.3}$$

and

$$N_4(v) = \frac{1}{2}(v(a_2) + v(a_5)), \tag{2.4}$$

where $a_5 = 2a_4 - a_2$, and it follows that $(T, \mathcal{P}_1, \mathcal{N})$ is a finite element.

2.2.2. *2-irregular hierarchical finite elements.* In $\mathbf{R}^2$ there are finite elements with 1–2 hanging nodes as in Figure 2.5, note that the hanging nodes now are on the the edges instead of in the vertices as for the 1-irregular Lagrange finite elements. The basis functions are as defined in Subsection 1.2.2 but the nodal variables are slightly different. We consider the finite element in the case of one hanging node, the other case is defined in the same way. The nodal variables are defined by

$$N_i(v) = v(a_i), \quad i = 1, 2, 3, \tag{2.5}$$

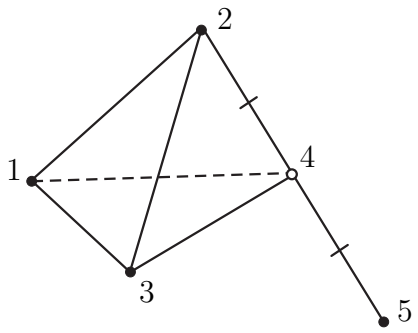


Figure 2.4: One hanging node for the 1-irregular Lagrange finite elements in $\mathbf{R}^3$. • denotes a regular node and ∘ denotes a hanging node.

and

$$
\begin{aligned}
N_{12}(v) &= v(a_{12}) - \frac{1}{2}(v(a_1) + v(a_2)), \\
N_{13}(v) &= v(a_{13}) - \frac{1}{2}(v(a_1) + v(a_3)), \\
N_{23}(v) &= \frac{1}{4}v(a_3) - \frac{1}{8}(v(a_2) + v(a_4)),
\end{aligned}
\tag{2.6}
$$

where $a_4 = 2a_3 - a_2$.

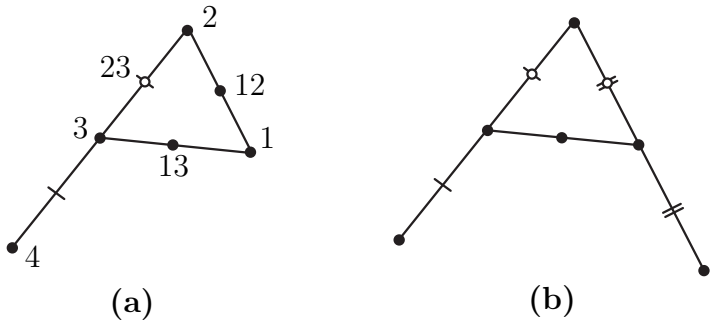


Figure 2.5: Two types of finite elements for the 2-irregular hierarchical finite element in $\mathbf{R}^2$. • denotes a regular node and ∘ denotes a hanging node. **(a)** One hanging node. **(b)** Two hanging nodes.

In order to show that the triplet $(T, \mathcal{P}_2, \mathcal{N})$ is a finite element we take v as in (1.3). Then for $i = 1, 2, 3$

$$N_i(v) = 0 \quad \Rightarrow \quad \tilde{v}_i = 0$$

and

$$\begin{aligned}
N_{12}(v) = 0 \quad &\Rightarrow \quad \frac{1}{2}\tilde{v}_1 + \frac{1}{2}\tilde{v}_2 + \tilde{v}_{12} - \frac{1}{2}(\tilde{v}_1 + \tilde{v}_2) = \tilde{v}_{12} = 0, \\
N_{13}(v) = 0 \quad &\Rightarrow \quad \frac{1}{2}\tilde{v}_1 + \frac{1}{2}\tilde{v}_3 + \tilde{v}_{13} - \frac{1}{2}(\tilde{v}_1 + \tilde{v}_3) = \tilde{v}_{13} = 0, \\
N_{23}(v) = 0 \quad &\Rightarrow \quad \frac{1}{4}\tilde{v}_3 - \frac{1}{8}(\tilde{v}_2 - \tilde{v}_2 + 2\tilde{v}_3 - 8\tilde{v}_{23}) = \tilde{v}_{23} = 0.
\end{aligned}$$

Thus $v = 0$ and from Remark 1.1 we conclude that $\{N_i\}_{i=1}^3 \cup \{N_{ij} : 1 \leq i < j \leq 3\}$ is a basis for $\mathcal{P}_2'$ and $(T, \mathcal{P}_2, \mathcal{N})$ is a finite element.

In $\mathbf{R}^3$ it is a bit more involved to maintain the continuity. There are tetrahedra with 1–5 hanging nodes. The basis functions are as defined in Sections 1.2.2 but the nodal variables are slightly different. We consider the finite element in the case of three hanging node as in Figure 2.6. The first nodal variables are defined by

$$N_i(v) = v(a_i), \quad i = 1, 2, 3, 4,$$

and

$$\begin{aligned}
N_{12}(v) &= v(a_{12}) - \frac{1}{2}(v(a_1) + v(a_2)), \\
N_{13}(v) &= v(a_{13}) - \frac{1}{2}(v(a_1) + v(a_3)), \\
N_{14}(v) &= v(a_{14}) - \frac{1}{2}(v(a_1) + v(a_4)), \\
N_{23}(v) &= \frac{1}{4}v(a_3) - \frac{1}{8}(v(a_2) + v(a_5)), \\
N_{24}(v) &= \frac{1}{4}v(a_4) - \frac{1}{8}(v(a_2) + v(a_6)), \\
N_{34}(v) &= \frac{1}{4}v(a_7) - \frac{1}{8}(v(a_5) + v(a_6)),
\end{aligned}$$

where

$$a_5 = 2a_3 - a_2, \qquad a_6 = 2a_4 - a_2, \qquad a_7 = a_3 + a_4 - a_2.$$

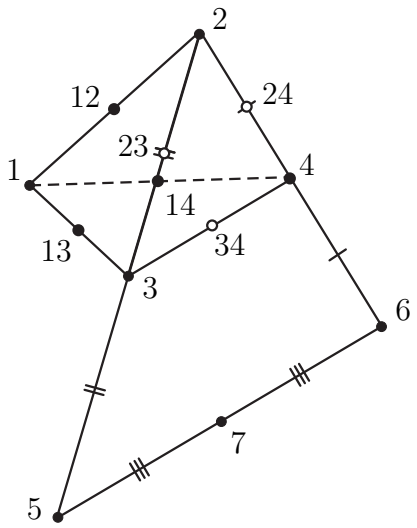


Figure 2.6: Three hanging nodes for the 2-irregular hierarchical finite element in $\mathbf{R}^3$. • denotes a regular node and ∘ denotes a hanging node.

In order to show that the triplet $(T, \mathcal{P}_2, \mathcal{N})$ is a finite element we take v as in (1.3). Then for $i = 1, 2, 3, 4$

$$N_i(v) = 0 \quad \Rightarrow \quad \tilde{v}_i = 0$$

and

$$\begin{aligned}
N_{12}(v) = 0 \quad &\Rightarrow \quad \frac{1}{2}\tilde{v}_1 + \frac{1}{2}\tilde{v}_2 + \tilde{v}_{12} - \frac{1}{2}(\tilde{v}_1 + \tilde{v}_2) = \tilde{v}_{12} = 0, \\
N_{13}(v) = 0 \quad &\Rightarrow \quad \frac{1}{2}\tilde{v}_1 + \frac{1}{2}\tilde{v}_3 + \tilde{v}_{13} - \frac{1}{2}(\tilde{v}_1 + \tilde{v}_3) = \tilde{v}_{13} = 0, \\
N_{14}(v) = 0 \quad &\Rightarrow \quad \frac{1}{2}\tilde{v}_1 + \frac{1}{2}\tilde{v}_4 + \tilde{v}_{14} - \frac{1}{2}(\tilde{v}_1 + \tilde{v}_4) = \tilde{v}_{14} = 0, \\
N_{23}(v) = 0 \quad &\Rightarrow \quad \frac{1}{4}\tilde{v}_3 - \frac{1}{8}(\tilde{v}_2 - \tilde{v}_2 - 2\tilde{v}_3 - 8\tilde{v}_{23}) = \tilde{v}_{23} = 0, \\
N_{24}(v) = 0 \quad &\Rightarrow \quad \frac{1}{4}\tilde{v}_4 - \frac{1}{8}(\tilde{v}_2 - \tilde{v}_2 - 2\tilde{v}_4 - 8\tilde{v}_{24}) = \tilde{v}_{24} = 0, \\
N_{34}(v) = 0 \quad &\Rightarrow \quad \frac{1}{4}(-\tilde{v}_2 + \tilde{v}_3 + \tilde{v}_4 - 4\tilde{v}_{23} - 4\tilde{v}_{24} + 4\tilde{v}_{34}), \\
&\qquad - \frac{1}{8}(-2\tilde{v}_2 + 2\tilde{v}_3 + 2\tilde{v}_4 - 8\tilde{v}_{23} - 8\tilde{v}_{24}) = \tilde{v}_{34} = 0.
\end{aligned}$$

Thus $v = 0$ and from Remark 1.1 we conclude that $\{N_i\}_{i=1}^4 \cup \{N_{ij} : 1 \leq i < j \leq 4\}$ is a basis for $\mathcal{P}_2'$ and $(T, \mathcal{P}_2, \mathcal{N})$ is a finite element.

2.3. **Finite element approximations on 1-irregular triangulations.** Let $\mathcal{T}_\ell$ be a 1-irregular triangulation. We define continuous finite element spaces V_ℓ on Ω by the finite elements $(T, \mathcal{P}_T, \mathcal{N}_T)_{T\in\mathcal{T}_\ell}$ as in Section 1.3, although we now use the 1 or 2-irregular finite elements defined in Section 2.2. Note that the index set I_T also changes, for example, $I_T = \{1, 2, 3, 12, 13, 4\}$ for the second order hierarchical finite element in two dimensions, Figure 2.5a.

With

$$V_\ell \ni u = \sum_{i=1}^{M_\ell} \tilde{u}_i \phi_i,$$

where $(\tilde{u}_1, \ldots, \tilde{u}_{M_\ell}) \in \mathbf{R}^{M_\ell}$ is the coordinate vector with respect to the basis $\{\phi_1, \ldots, \phi_{M_\ell}\}$ and taking $v = \phi_j$ we express (1.8), now with $u, v \in V_\ell$, as

$$\sum_{i=1}^{M_\ell} \tilde{u}_i a(\phi_i, \phi_j) = f(\phi_j) \quad \text{for } j = 1, \ldots, M_\ell. \tag{2.7}$$

Locally on each $T \in \mathcal{T}_\ell$ we have

$$\phi_i|_T = \sum_{k\in I_T} N_{k,T}(\phi_i)\varphi_{k,T},$$

and hence (2.7) is equivalent to

$$\begin{aligned} &\sum_{i=1}^{M_\ell} \sum_{T\in S_i} \sum_{k,l\in I_T} \tilde{u}_i N_{k,T}(\phi_i) a(\varphi_{k,T}, \varphi_{l,T}) N_{l,T}(\phi_j) \\ &\qquad = \sum_{T\in S_i} \sum_{l\in I_T} N_{l,T}(\phi_j) f(\varphi_{l,T}) \quad \text{for } j = 1, \ldots, M_\ell, \end{aligned} \tag{2.8}$$

where we identify $a(\varphi_{k,T}, \varphi_{l,T})$ as a local stiffness matrix and $f(\varphi_{l,T})$ as the local load vector.

The rather involved formula (2.8) in fact expresses the distribution mapping defined in [1, 12], which is useful in practice implementing finite element problems. We note that assembling the stiffness matrix and load vector we only need to know a few things for each $T \in \mathcal{T}_\ell$: (1) where to put the elements from the local stiffness matrix and load vector into the global stiffness matrix and load vector, and (2) the weights $N_{k,T}(\phi_i)$ on the local

elements. We represent this information in a set of arrays holding three numbers, $(i, k, N_{k,T}(\phi_i))_{T\in\mathcal{T}_\ell}$, where i is a global index, k is a local index on T and $N_{k,T}(\phi_i)$ is a weight, and likewise for $(j, l, N_{l,T}(\phi_j))_{T\in\mathcal{T}_\ell}$. More precisely we define the representation

$$\text{Rep}(T) = \{(i, k, N_{k,T}(\phi_i)) : i \in i_{k'},\ k, k' \in I_T,\ N_{k,T}(\phi_i) \neq 0\}. \tag{2.9}$$

Thus, provided the finite elements $(T, \mathcal{P}_T, \mathcal{M}_{\ell T})_{T\in\mathcal{T}_\ell}$ are well defined we express the finite element problem as in (2.8) and use the representation (2.9) for assembling the problem in practice.

We remark that when $V_\ell \ni f = \sum_{i=1}^{M_\ell} \tilde{f}_i\phi_i$ and (2.8) becomes

$$\begin{aligned}\sum_{i=1}^{M_\ell}\sum_{T\in S_i}\sum_{k,l\in I_T} \tilde{u}_i N_{k,T}(\phi_i)a(\varphi_{k,T},\varphi_{l,T})N_{l,T}(\phi_j)\\ = \sum_{T\in S_i}\sum_{k,l\in I_T} \tilde{f}_i N_{k,T}(\phi_i)(\varphi_{k,T},\varphi_{l,T})N_{l,T}(\phi_j) \quad \text{for } j = 1,\dots,M_\ell.\end{aligned} \tag{2.10}$$

where we identify $(\varphi_{k,T}, \varphi_{l,T})$ as the local mass matrix.

In the next four sections we explicitly compute $\text{Rep}(T)$ for the finite elements in Sections 1.2 and 2.2.

2.3.1. *Lagrange finite elements.* In this case since $N_k(v) = v(x_k)$ for $x_k \in L_q(T)$ and $k \in I_T$ as defined in Section 1.2.1, the representation (2.9) is particularly simple:

$$\text{Rep}(T) = \begin{pmatrix} i_k \\ k \\ 1 \end{pmatrix} \quad \forall\, k \in I_T,$$

which probably anyone that have implemented the Lagrange finite elements recognizes.

2.3.2. *Higher degree hierarchical finite elements.* We evaluate (2.9) for the quadratic hierarchical finite element in two dimensions. With $N_k(\cdot)$ and I_T as defined in Section 1.2.2 we get

$$\text{Rep}(T) = \begin{pmatrix} i_1 & i_2 & i_3 & i_{12} & i_{13} & i_{23} & i_1 & i_1 & i_2 & i_2 & i_3 & i_3 \\ 1 & 2 & 3 & 12 & 13 & 23 & 12 & 13 & 12 & 23 & 13 & 23 \\ 1 & 1 & 1 & 1 & 1 & 1 & -1/2 & -1/2 & -1/2 & -1/2 & -1/2 & -1/2 \end{pmatrix}.$$

The three-dimensional case is analogous.

2.3.3. *1-irregular Lagrange finite elements.* We evaluate (2.9) for the 1-irregular finite element in two dimensions. With $N_k(\cdot)$ for $x_k \in L_1(T)$ and I_T as defined in Section 2.2.1 we get

$$\operatorname{Rep}(T) = \begin{pmatrix} i_1 & i_2 & i_2 & i_4 \\ 1 & 2 & 3 & 3 \\ 1 & 1 & 1/2 & 1/2 \end{pmatrix},$$

The three-dimensional case is analogous.

2.3.4. *2-irregular hierarchical finite elements.* We evaluate (2.9) for the 2-irregular finite element in two dimensions. With $N_k(\cdot)$ and I_T as defined in Section 2.2.2 we get

$$\operatorname{Rep}(T) = \begin{pmatrix} i_1 & i_2 & i_3 & i_{12} & i_{13} & i_1 & i_1 & i_2 & i_3 & i_3 & i_2 & i_4 \\ 1 & 2 & 3 & 12 & 13 & 12 & 13 & 12 & 13 & 23 & 23 & 23 \\ 1 & 1 & 1 & 1 & 1 & -1/2 & -1/2 & -1/2 & -1/2 & 1/4 & -1/8 & -1/8 \end{pmatrix}.$$

The three-dimensional case is analogous.

2.4. **Multigrid on 1-irregular triangulations.** We need to find the projection $Q_{\ell-1} : V_\ell \to V_{\ell-1}$ as defined in Section 1.3. Since $V_{\ell-1} \subset V_\ell$ we can express the basis functions in $V_{\ell-1}$, $\{\phi_i^{\ell-1}\}_{i=1}^{M_{\ell-1}}$, in terms of the base functions in V_ℓ, $\{\phi_i^\ell\}_{i=1}^{M_\ell}$. Hence, with the definition of $Q_{\ell-1}$,

$$(Q_{\ell-1} v_\ell, \phi_i^{\ell-1}) = (v_\ell, \phi_i^{\ell-1}) = \sum_{j=J_i^\ell} \alpha_{ij}^\ell (v_\ell, \phi_j^\ell),$$

for $v_\ell \in V_\ell$ and where $J_i^\ell := \{j : \operatorname{supp}(\phi_j^\ell) \cap S_i^{\ell-1} \neq \emptyset\}$.

We use the nodal variables to express α_{ij}^ℓ

$$\alpha_{ij}^\ell = N_{k,T}^\ell(\phi_i^{\ell-1}),$$

for all $T \in \mathcal{T}_\ell$ such that $T \cap S_i^{\ell-1} \neq \emptyset$ and for all $k \in I_T$ where $j_k = j$ is the local to global mapping defined in Section 1.3.

3. Numerical experiments

In matrix form (1.8) becomes

$$\mathcal{A}\tilde{u} = \mathcal{F},$$

where $\mathcal{A}$ denotes the matrix $[\mathcal{A}]_{ij} = (A_L \phi_i, \phi_j)$ for $i = 1, \ldots, M_\ell$ and $\tilde{u} \in \mathbf{R}^{M_\ell}$ denotes the coordinate vector with respect to the finite element

basis and $\mathcal{F} = (\mathcal{F}_1, \ldots, \mathcal{F}_{M_\ell})$ where $\mathcal{F}_i = (f, \phi_i)$. We solve this linear system using the V-cycle Algorithm 1 with five iterations of a point Gauss-Seidel smoother. With $\tilde{u}^0 = 0$ we iterate $m = 1, 2, \ldots$ until the relative residual

$$\text{Res} := \frac{\|\mathcal{F} - \mathcal{A}\tilde{u}^m\|}{\|\mathcal{F}\|}$$

is less than a specified tolerance 'Tol' set to 10^{-6} in this work. Note that the relative tolerance times a constant is always greater than $\|u - u^m\|_{1,\Omega}$ where $u \in V_L$ is the finite element solution we are approximating, cf. [9, Proposition 9.19, p. 393].

3.1. **The Poisson equation.** We consider the following Poisson equation with mixed Dirichlet-Neumann boundary conditions on bounded polyhedral domains $\Omega \subset \mathbf{R}^n$ for $n = 2, 3$,

$$-\Delta u = f \quad \text{in } \Omega, \qquad u = g \quad \text{on} \quad \partial\Omega_D, \quad \text{and} \quad \nu \cdot \nabla u = 0 \quad \text{on} \quad \partial\Omega_N,$$

where the boundary is partitioned such that $\partial\Omega_D \cup \partial\Omega_N = \partial\Omega$, g is a constant, ν is the outward normal to the boundary and we assume $f \in H^{-1}(\Omega)$ and thus the problem is a well posed. Let

$$V = \{u \in H^1(\Omega) : u = 0 \text{ on } \partial\Omega_D\}.$$

Now the bilinear and linear forms in Section 1.1 are

$$a(u, v) = \int_\Omega \nabla u \cdot \nabla v \, dx,$$

and

$$f(v) = \int_\Omega f v \, dx.$$

With u_g denoting the extension of g to $H^1(\Omega)$, the weak formulation to the above Poisson problem follows as usual and reads: find $u \in H^1(\Omega)$ such that

$$\begin{aligned} u &= u_g + \phi, \qquad \phi \in V, \\ a(\phi, v) &= f(v) - a(u_g, v) \qquad \forall v \in V. \end{aligned} \tag{3.1}$$

We use the maximum norm error estimator derived in [7, 11] to adaptively refine the triangulations. For the solution u to (1.8) and $T \in \mathcal{T}_\ell$ we compute

$$\eta_T = h_T \|f + \Delta u\|_{L^\infty(T)} + \frac{1}{2} \|[\partial_{\nu_T} u]\|_{L^\infty(\partial T \setminus \partial\Omega)},$$

where $[\partial_{\nu_T} u]$ denotes the jump across ∂T in the normal derivative, $\partial_{\nu_T} u = \nu_T \cdot \nabla u$ where ν_T denotes the outward normal to ∂T.

We define $\mathcal{T}_{\ell+1}$ by refining those $T \in \mathcal{T}_\ell$ where

$$\eta_T > \overline{\eta}_T + s,$$

where $\overline{\eta}_T$ is the mean and s is the standard deviation of η_T.

3.1.1. *Model problem.* In this case we let Ω be the L-shaped domain with one re-entrant edge, $\Omega = \{(x, y) \in [0, 2]^2 \setminus [1, 2] \times [0, 1]\}$ for $n = 2$ and $\Omega = \{(x, y, z) \in [0, 2]^2 \times [0, 0.5] \setminus [1, 2] \times [0, 1] \times [0, 0.5]\}$ for $n = 3$, see Figure 3.1 and 3.2.

Let $\partial\Omega_D = \partial\Omega_{D_0} \cup \partial\Omega_{D_1}$ where $\partial\Omega_{D_0} = \{(x, y) : x = 1, y \in [1, 2]\}$ and $\partial\Omega_{D_1} = \{(x, y) : x \in [0, 1], y = 0\}$ for $n = 2$ and $\partial\Omega_{D_0} = \{(x, y, z) : x = 2, (y, z) \in [1, 2]\times[0, 0.5]\}$ and $\partial\Omega_{D_1} = \{(x, y, z) : (x, z) \in [0, 1]\times[0, 0.5], y = 0\}$ for $n = 3$. Set $f = 0$, $g = 0$ on $\partial\Omega_{D_0}$ and $g = 1$ on $\partial\Omega_{D_1}$.

We solve the problem for different finite element approximations and refine the triangulations eight times. The results from these experiments are summarized in and Tables 3.1 and 3.2.

Table 3.1: Convergence data for the V-cycle multigrid Algorithm 1 applied to the Model Problem for $n = 2$ and the finite elements in Section 2.2.

Lagrange $(n, q) = (2, 1)$				Hierarchical $(n, q) = (2, 2)$			
ℓ	M_1	m	Res	ℓ	M_2	m	Res
1	272	2	$1.3 \cdot 10^{-7}$	1	299	2	$1.2 \cdot 10^{-7}$
2	368	2	$8.3 \cdot 10^{-8}$	2	459	2	$2.7 \cdot 10^{-8}$
3	524	2	$7.1 \cdot 10^{-8}$	3	644	2	$2.0 \cdot 10^{-8}$
4	767	3	$2.2 \cdot 10^{-7}$	4	890	2	$4.3 \cdot 10^{-8}$
5	1072	2	$3.9 \cdot 10^{-8}$	5	1447	3	$3.1 \cdot 10^{-8}$
6	1642	2	$2.5 \cdot 10^{-8}$	6	2066	3	$2.7 \cdot 10^{-7}$
7	2415	3	$3.0 \cdot 10^{-8}$	7	3158	3	$2.4 \cdot 10^{-7}$
8	3577	3	$5.7 \cdot 10^{-7}$	8	4546	4	$3.8 \cdot 10^{-8}$

4. Conclusions

We outlined a methodology for implementing the finite element multigrid method on adaptively refined triangulations for various finite elements,

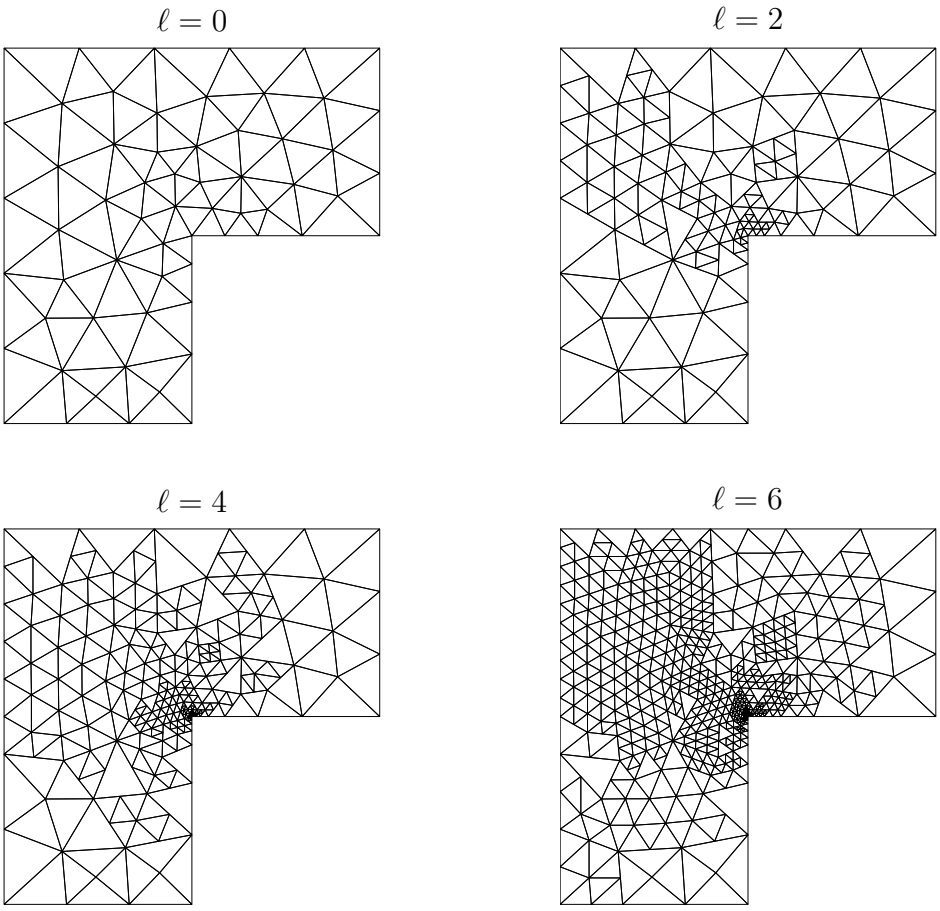


Figure 3.1: Adaptively refined triangulations $\mathcal{T}_\ell$ of the L-shaped domain in two dimensions.

Lagrange $q = 1, 2$ and hierarchical $q = 2$ for $n = 2, 3$. In a few numerical experiments we demonstrated that the methodology works in practice by solving a number of problems in two and three dimensions.

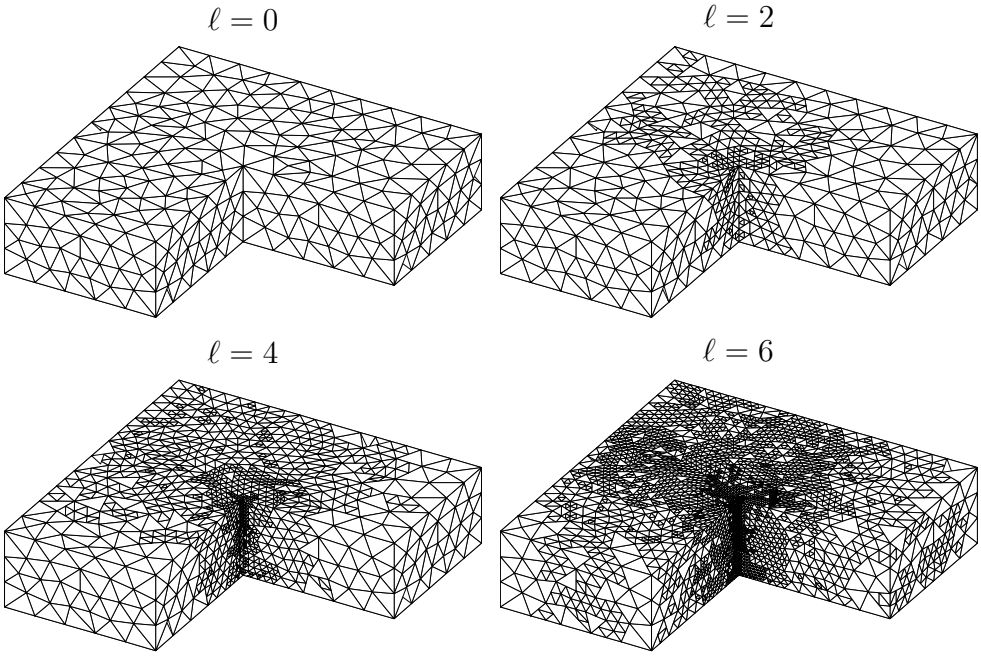


Figure 3.2: Adaptively refined triangulations $\mathcal{T}_\ell$ of the L-shaped domain in three dimensions.

Table 3.2: Convergence data for the V-cycle multigrid Algorithm 1 applied to Model Problem for $n = 3$ and the Lagrange finite element in Section 2.2.1.

Lagrange $(n, q) = (3, 1)$			
ℓ	M_1	m	Res
1	1382	2	$9.8 \cdot 10^{-9}$
2	3207	2	$1.2 \cdot 10^{-7}$
3	5378	2	$1.8 \cdot 10^{-7}$
4	11542	2	$5.2 \cdot 10^{-7}$
5	24185	2	$4.4 \cdot 10^{-7}$
6	46834	3	$2.7 \cdot 10^{-7}$
7	106711	3	$2.9 \cdot 10^{-7}$
8	225353	3	$2.3 \cdot 10^{-7}$

References

[1] M. Ainsworth and B. Senior, *Aspects of an adaptive hp-finite element method: adaptive strategy, conforming approximation and efficient solvers*, Comput. Methods Appl. Mech. Engrg. **150** (1997), 65–87.

[2] O. Axelsson and I. Gustafsson, *Preconditioning and two-level multigrid methods of arbitrary degree of approximation*, Math. Comp. **40** (1983), 219–242.

[3] J. Bey, *Tetrahedral grid refinement*, Computing **55** (1995), 355–378.

[4] J. H. Bramble and X. Zhang, *The Analysis of Multigrid Methods*, Handbook of numerical analysis, Vol. VII, North-Holland, 2000.

[5] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, second ed., Springer-Verlag, 2002.

[6] P. G. Ciarlet, *Basic error estimates for elliptic problems*, Handbook of Numerical Analysis, Vol. II, North-Holland, 1991.

[7] E. Dari, R. G. Durán, and C. Padra, *Maximum norm error estimators for three-dimensional elliptic problems*, SIAM J. Numer. Anal. **37** (2000), 683–700.

[8] L. Demkowicz, J. T. Oden, W. Rachowicz, and O. Hardy, *Toward a universal h-p adaptive finite element strategy. I. Constrained approximation and data structure*, Comput. Methods Appl. Mech. Engrg. **77** (1989), 79–112.

[9] A. Ern and J-L. Guermond, *Theory and Practice of Finite Elements*, Springer-Verlag, 2004.

[10] W. Hackbusch, *Multigrid Methods and Applications*, Springer-Verlag, 1985.

[11] Rn H. Nochetto, *Pointwise a posteriori error estimates for elliptic problems on highly graded meshes*, Math. Comp. **64** (1995), 1–22.

[12] J.T. Oden, *A general theory of finite elements. ii. applications*, Int. J. Numer. Methods Eng. (UK) **1** (1969), 247–259.

[13] L. R. Scott and S. Zhang, *Higher-dimensional nonnested multigrid methods*, Math. Comp. **58** (1992), 457–466.

[14] V. V. Shaĭdurov, *Multigrid Methods for Finite Elements*, Kluwer Academic Publishers Group, 1995.

Department of Mathematical Sciences, Chalmers University of Technology, SE-412 96 Göteborg, Sweden

E-mail address: erik.svensson@math.chalmers.se