SPECIAL TOPICS IN NUMERICS

(I. FEM for Nonlinear Problems)

Rolf Rannacher

Institute of Applied Mathematics
Heidelberg University

Lecture Notes (SS 2016)

April 20, 2017

Address of Author:

Prof. Dr. Dr.h.c. Rolf Rannacher
Institute of Applied Mathematics
Heidelberg University
Im Neuenheimer Feld 205 (MATHEMATIKON)
D-69120 Heidelberg, Germany

rannacher@iwr.uni-heidelberg.de
http://www.uni-heidelberg.de/numerik

Contents

1 Introduction

1.1 List of topics covered in this course

- Basics of FEM (Literature: Lecture Notes Rannacher [3, 47], Books of Braess [25], Ciarlet [26], Quarteroni & Valli [28], Strang & Fix [29]), Johnson [32], Brenner & Scott [34]).

- Special types of non-linear problems: variational inequalities, obstacle problem, minimal surface problem, problems of monotone type (Literature: Book of Ciarlet [26], Article of Rannacher [41]).

- General quasi-linear elliptic boundary value problems, Newton method, mesh independent convergence (Literature: Articles of Frehse & Rannacher [37, 38], Dobrowolski & Rannacher [36], Rannacher [41, 43, 44, 48], Rannacher & Scott [49]).

- The (stationary) Navier-Stokes equations, linearization procedures, stable Stokes elements, pressure and transport stabilization, solution of linear sub-systems, nested solution approaches (Literature: Lecture Notes of Rannacher [4], Books of Temam [19], Galdi [7], Girault & Raviart [27], Brenner & Scott [34], Articles of Heywood & Rannacher [39], Heywood, Rannacher & Turek [40], Rannacher & Turek [50], Rannacher [45, 46, 47]).

1.2 Basic notation

We have to deal with scalar or vector-valued functions $u = u(x)$ for arguments $x \in \mathbb{R}^n$. For derivatives of differentiable functions, we use the notation

$$\partial_x u := \frac{\partial u}{\partial x}, \quad \partial_x^2 u := \frac{\partial^2 u}{\partial^2 x}, \ \dots \ , \quad \partial_i u := \frac{\partial u}{\partial x_i}, \quad \partial_{ij}^2 u := \frac{\partial^2 u}{\partial x_i \partial x_j}, \ \dots \ ,$$

and analogously also for higher-order derivatives. With the nabla operator ∇ the "gradient" of a scalar function and the "divergence" of a vector function are written as $\operatorname{grad} u = \nabla u := (\partial_1 u, ..., \partial_d u)$ and $\operatorname{div} u = \nabla \cdot u := \partial_1 u_1 + ... + \partial_d u_d$, respectively. For a vector $\beta \in \mathbb{R}^d$ the derivative in direction β is written as $\partial_\beta u := \beta \cdot \nabla u$. Combination of gradient and divergence yields the "Laplacian operator"

$$\nabla \cdot \nabla u = \Delta u = \partial_1^2 u + ... + \partial_d^2 u.$$

The symbol $\nabla^m u$ denotes the "tensor" of all partial derivatives of order m of u, i. e., in two dimensions $u = u(x_1, x_2)$, $\nabla^2 u = (\partial_1^i \partial_2^j u)_{i+j=2}$.

1.3 Basics of finite element method for linear problems

Standard second-order model problem (Poisson diffusion equation)

$$-\Delta u = f \ \text{ in } \Omega, \quad u = 0 \ \text{ on } \partial\Omega, \tag{1.3.1}$$

in a bounded domain $\Omega \subset \mathbb{R}^d (d = 2 ord = 3)$ with sufficiently regular boundary $\partial\Omega$. Corresponding fourth-order problem (plate bending problem)

$$\Delta^2 u = f \text{ in } \Omega, \quad u = 0, \ \partial_n u = 0 \text{ on } \partial\Omega. \tag{1.3.2}$$

Variational formulation: Minimize the energy functional

$$E(u) := \tfrac{1}{2}(\nabla u, \nabla u) - (f, u)$$

on the function Hilbert space ("energy space" consisting of functions with finite energy) $V := H_0^1(\Omega)$. There exists a unique strict minimum, which is determined by the variational equation

$$(\nabla u, \nabla \varphi) = (f, \varphi) \quad \forall \varphi \in V. \tag{1.3.3}$$

We use the abstract formulation

$$a(u, \varphi) = l(\varphi) \quad \forall \varphi \in V, \quad E(u) = \tfrac{1}{2} a(u, u) - l(u), \tag{1.3.4}$$

with the "energy bilinear form" $a(u, \varphi) := (\nabla u, \nabla \varphi)$ and the "load functional" $l(\varphi) := (f, \varphi)$. For the plate bending problem there holds a similar formulation (exercise).

Finite element subspaces $V_h \subset V$ parametrized by a mesh size parameter $h > 0$. Discrete approximations $u_h \in V_h$ are determined by the variational equations

$$a(u_h, \varphi_h) = l(\varphi_h) \quad \forall \varphi_h \in V_h. \tag{1.3.5}$$

There holds "Galerkin orthogonality"

$$a(u - u_h, \varphi_h) = 0, \quad \varphi_h \in V_h, \tag{1.3.6}$$

and consequently (for symmetric energy form) the "best approximation property"

$$\|u - u_h\|_E = \min_{\varphi_h \in V_h} \|u - \varphi_h\|_E, \tag{1.3.7}$$

with the "energy norm" $\|\cdot\|_E := a(\cdot, \cdot)^{1/2}$. Notion of "Ritz projection" $R_h : V \to V_h$,

$$a(R_u, \varphi_h) = a(u, \varphi_h) \quad \forall \varphi_h \in V_h. \tag{1.3.8}$$

Improved convergence estimates in weaker norms by duality argument

$$\|u - R_h u\| \le \|u - R_h u\|_E \sup_{z \in V \cap H^2(\Omega)} \|z - R_h z\|_E. \tag{1.3.9}$$

This is general theory for Ritz-Galerkin methods.

Special case of "finite element spaces" and approximation properties:

i) types of meshes:

$h_T :=$ radius of minimal circumscribed ball ($\sim$ diam(T)),
$\rho_T :=$ radius of maximal inscribed ball.

"structural regularity" (no hoes and ', "shape regularity", "size regularity" - "quasi-uniformity". "nodal interpolation" $I_T v \in P(T)$ on cells $T \in \mathbb{T}_h$ ("unisolvance property"), "nodal interpolation" $I_h v \in V_h$ on meshes $\mathbb{T}_h = \{T\}$ of sufficiently smooth functions (e. g., $v \in C(\bar{\Omega})$. Interpolation error estimates for finite elements of order $m \geq 2$ (polynomial degree $m-1 \geq 1$, special situation for $m = 1$):

$$\|\nabla(v - I_T v)\|_T \leq c_I h_T^{m-1} \|\nabla^m v\|_T. \tag{1.3.10}$$

Consequence, $h := \max_{T \in \mathbb{T}_h} h_T$,

$$\begin{aligned} \|v - I_h v\| &= \Big(\sum_{T \in \mathbb{T}_h} \|v - I_T v\|_T^2 \Big)^{1/2} \leq c_I \Big(\sum_{T \in \mathbb{T}_h} h_T^{2(m-1)} \|\nabla^m v\|_T^2 \Big)^{1/2} \\ &\leq c_I h^{m-1} \|\nabla^m v\|. \end{aligned}$$

Other interpolation error estimates.

$$\begin{aligned} \|v - I_h v\|_\Omega &\leq c h^m \|\nabla^m v\|_\Omega, \\ \|v - I_h v\|_{\partial\Omega} &\leq c h^{m-1/2} \|\nabla^m v\|_\Omega, \\ \|v - I_h v\|_{L^\infty(\Omega)} &\leq c... \end{aligned}$$

"Inverse relation" for finite elements:

$$\|\nabla v_h\|_T \leq c \rho_T^{-1} \|v_h\|_T, \quad T \in \mathbb{T}_h. \tag{1.3.11}$$

Consequently, for quasi-uniform meshes:

$$\|\nabla v_h\|_\Omega \leq h^{-1} \|v_h\|_\Omega, \quad v_h \in V_h.$$

Questions to be considered in this context:

i) best-approximation estimates in L^p-norms $(1 \leq p \leq \infty)$?
ii) relaxation of requirements on finite element meshes?

1.4 Useful technics of mathematical analysis in the finite element method

We present some technical arguments, which are frequently used in the error analysis for finite element approximations.

1.4.1 Duality arguments ("Aubin-Nitsche trick")

From the best-approximation property and the interpolation error estimates for finite elements, we obtain the basic energy-error estimate

$$\|\nabla(u - u_h)\| \leq \min_{\varphi_h \in V_h} \|\nabla(u - \varphi_h)\|.$$

Auxiliary ("dual") problem for $e := u - u_h$,

$$a(\varphi, z) = \frac{(e, \varphi)}{\|e\|} \quad \forall \varphi \in V.$$

Since Ω is assuned to be smoothly bounded or a convex polygonal domain the "dual solution" $z \in V$ is in $H^2(\Omega)$ and satisfies the a priori estimate

$$\|\nabla^2 z\| \leq \|e\| \quad (\text{constant} c = 1).$$

Then, by Galerkin orthogonality,

$$\|e\| = a(e, z) = a(e, z - \psi_h) \leq \|\nabla e\| min_{\psi_h \in V_h} \|\nabla(z -$$

1.4.2 Inverse estimates

There holds

$$\|\nabla v_h\|_T \leq c\rho_T^{-1} \|v_h\|_T$$

Proof by transformation onto "reference unit cell" $T \to \hat{T}$ General rule:

Variant in $\mathbb{R}^2$:

$$\max_T \|v_h\| \leq ch^{-1} \|v_h\|_T$$

1.4.3 Local integral inequalities

We estimate functions in $H^1(\Omega)$ over lower dimensional faces Γ. This gives us the so-called "trace theorem"

$$\Big(\int_\Gamma |v|^2 \, do \Big)^{1/2} \leq c...$$

Estimate first proven for regular functions $v \in C^1(\overline{\Omega})$ and then extended by continuity to its completion $H^1(\Omega)$.

1.4.4 Lax-Milgram Lemma

The so-called Lax-Milgram Lemma is a generalization of the well-known Riesz representation theorem for nonsymmetric bilinear forms. Let H be a (real) Hilbert space with scalar product $(\cdot, \cdot)$ and associated norm $\|\cdot\| = (\cdot, \cdot)^{1/2}$. Then, for any linear bounded (continuous) functional $l(\cdot) : H \to \mathbb{R}$ (i. e., any element from the "dual space" H^* of H), there exists a unique element $v \in H$ such that

$$l(\varphi) = (v, \varphi), \quad \varphi \in H,$$

and

$$\|v\| = \|l\|_{V^*} := \sup_{\varphi \in H} \frac{l(\varphi)}{\|\varphi\|}.$$

The mapping $l \to v$ defines an isometric isomorphism between the Hilbert space H and its dual H^*. In this sense the two spaces may be identified. The formally more general situation

that $a(\cdot,\cdot)$ is a bounded, symmetric bilinear form on H,

$$a(u,v) = a(v,u), \quad |a(u,v)| \leq \alpha \|u\| \|v\|, \quad u,v \in H,$$

can be embedded into this situation if the bilinear form is positive definit,

$$a(v,v) \geq \gamma \|v\|^2, \quad v \in H.$$

Then, the bilinear form $a(\cdot,\cdot)$ constitutes a scalar product on H, and its associated norm is equivalent to the given norm of H. The Lax-Milgram lemma addresses this situation for the case that $a(\cdot,\cdot)$ is not symmetric, but bounded and positive definit.

1.5 Exercises (for refreshing the knowledge of some preparatory material)

Exercise 1.1: Give **short** answers to the following questions:

1. Which properties should an elliptic boundary value problem or a parabolic initial-boundary value problem possess for being called "well-posed"?

2. Let the 1st BVP of the Laplace operator on the unit square be approximated by bilinear finite elements on an equidistant cartesian quadrilateral mesh. Which finite difference scheme is obtained if the elements of the system matrix are computed using the tensor-product trapezoidal rule? What are the orders of the finite element scheme and of this related finite difference scheme?

3. What are the dimensions of the polynomial spaces $Q_1(T), P_2(T)$ and $P_5(T)$ in $\mathbb{R}^2$?

4. On a tetrahedron $T \in \mathbb{R}^3$ let a polynomial space $P(T)$ and a set of functionals $\chi_r : C^1(\overline{\Omega}) \to P(T)$ $(r = 1, \ldots, R)$ be given. What does it mean that $\{\chi_r\}_{r=1,\ldots,R}$ is "unisolvent" with respect to $P(T)$? What is the natural "nodal interpolation" $I_h v \in P(T)$ of a continuous function v?

5. What is the difference between the "Ritz projection method" and a general "Galerkin method"?

6. Explain the meaning of the terms "conformity", "Galerkin orthogonality" and "best approximation property" in the context of a finite element discretization.

7. Let the 1st BVP of the Laplace operator be discretized by a finite element method with nodal basis $\{\varphi_h^i,\, i = 1, \ldots, N_h\}$. What are the corresponding "mass matrix" and "stiffness matrix"?

8. What is the h-dependence of the condition number of the system matrices of a finite element discretization for the Laplace operator and for the biharmonic operator on a sequence of quasi-uniform meshes?

9. What is the difference between the "method of lines" and the "Rothe method" for the discretization of the parabolic heat equation? Why are the resulting linear ODE systems in the "Method of lines" generically "stiff"?

Exercise 1.2: Give the best possible powers of h in the following error estimates for the nodal interpolation $I_T v$ into the space of linear polynomials $P(T) := P_1(T)$ on a form-regular mesh $\mathbb{T}_h = \{T\}$ in 2d and $a \in T$ (no proof required):

$$\begin{aligned} (i) \qquad & \|v - I_T v\|_T \le c_i h_T^? \|\nabla^2 v\|_T, \\ (ii) \qquad & |(v - I_T v)(a)| \le c_i h_T^? \|\nabla^2 v\|_T, \\ (iii) \qquad & \|\partial_n (v - I_T v)\|_{\partial T} \le c_i h_T^? \|\nabla^2 v\|_T, \end{aligned}$$

for $v \in H^2(T)$. Why is the estimate

$$\|v - I_T v\|_T \le c_i h_T \|\nabla v\|_T$$

not possible uniformly for $v \in H^2(T)$? State the corresponding best possible estimates for the nodal interpolation into the space of quadratic polynomials.

Exercise 1.3: Consider the Dirichlet BVP

$$-\Delta u + \gamma u = f \ \text{ in } \Omega, \quad u = 0 \ \text{ on } \partial\Omega,$$

on a convex polygonal domain $\Omega \subset \mathbb{R}^2$. The data $\gamma \ge 0$ and f are supposed to be sufficiently regular. Let an approximate solution u_h be computed by using the finite element method with subspaces $V_h \subset H_0^1(\Omega)$.

a) State the corresponding discrete and continuous variational formulations.

b) For this discretization with "linear" elements on a quasi-uniform family of triangulations $\{\mathbb{T}_h\}_{h \in \mathbb{R}_+}$ state optimal-order a priori error estimates in the energy and the L^2 norm. How does the condition number of the resulting system matrices A_h depend on the mesh width h?

c) Consider part b), but now with "quadratic" elements.

Exercise 1.4: Give variational formulations in appropriate Sobolev spaces for the following boundary value problems:

$$\begin{aligned} a) \qquad & -\Delta u + u = f \ \text{ in } \Omega, \quad \partial_n u = g \ \text{ on } \partial\Omega, \\ b) \qquad & \Delta^2 u = f \ \text{ in } \Omega, \quad u = \Delta u = 0 \ \text{ on } \partial\Omega, \\ c) \qquad & -\nabla \cdot (a \nabla u) + \beta \cdot \nabla u = f \ \text{ in } \Omega, \quad u = 0 \ \text{ on } \partial\Omega, \end{aligned}$$

where the data $a, \beta, f,$ and g are assumed to be sufficiently regular.

Exercise 1.5: Let $\beta \in C^1(\bar{\Omega})^2$ be a transport vector function with the property $\nabla \cdot \beta = 0$, and $f \in L^2(\Omega)$. Prove by the Lax-Milgram lemma that the 2-d transport-diffusion problem

$$-\Delta u + \beta \cdot \nabla u = f \ \text{ in } \Omega, \quad u = 0 \ \text{ on } \partial\Omega,$$

possesses a unique weak solution in the Sobolev space $V = H_0^1(\Omega)$. Which regularity can be expected for this solution if Ω is a convex polygonal domain?

Exercise 1.6: The standard "trace inequality"

$$\|v\|_{L^2(\partial\Omega)} \le c \|v\|_{H^1(\Omega)}$$

is well known to hold for functions $v \in H^1(\Omega)$ on domains $\Omega \subset \mathbb{R}^d$ with sufficiently regular boundary $\partial\Omega$. Derive the following sharpened

version (for simplicity on the unit square $\Omega = (0,1)^2 \subset \mathbb{R}^2$:

$$\|v\|_{L^2(\partial\Omega)} \leq c\|v\|^{1/2}_{H^1(\Omega)}\|v\|^{1/2}_{L^2(\Omega)},$$

(Hint: Modify the argument for proving the first "trace inequality" given in the lecture notes of the PDE Numerics course.)

Exercise 1.7: Consider the approximation of the Neumann problem

$$-\Delta u + u = f \text{ in } \Omega, \quad \partial_n u = 0 \text{ on } \partial\Omega,$$

on a convex polygonal domain $\Omega \subset \mathbb{R}^2$ by the finite element method using piecewise linear finite elements on a quasi-uniform sequence of triangulations $\mathbb{T}_h = \{T\}$ with mesh sizes $h \to 0$. Prove the error estimates

$$\|u - u_h\|_{L^2(\Omega)} + h^{1/2}\|u - u_h\|_{L^2(\partial\Omega)} + h\|\nabla(u - u_h)\|_{L^2(\Omega)} \leq ch^2\|f\|_{L^2(\Omega)}.$$

(Hint: Use the standard arguments for proving error estimates for finite element approximations such as "best approximation property", optimal-order interpolation estimates, "duality arguments", and the trace estimate of Exercise 1.3.)

2 Some Special Types of Nonlinear Problems

This chapter deals with the finite element solution of certain classes of nonlinear problems. Typical examples are the "obstacle problem" and the "minimal surface problem". The material of this chapter and further details can largely be found in the textbook of Ciarlet [26] and the article of Rannacher [41]. To get started, we recall the basic elliptic model problem

$$-\Delta u = f \ \text{ in } \Omega, \quad u = g \ \text{ on } \partial\Omega, \tag{2.0.1}$$

which possesses a unique "weak" solution in the linear manifold $V_g := V + g$, $V := H_0^1(\Omega)$. For this, we assume that the boundary value is given as the trace of a function $g \in H^1(\Omega)$. Then, this weak solution is characterized as minimizer on V_g of the quadratic energy functional, assuming that $f \in L^2(\Omega)$,

$$E(u) := \tfrac{1}{2}\|\nabla u\|^2 - (f, u),$$

or equivalently by the variational equation

$$a(u, \varphi) = l(\varphi) \quad \forall \varphi \in V, \tag{2.0.2}$$

with the notation

$$a(u, \varphi) := (\nabla u, \nabla \varphi), \quad l(\varphi) := (f, \varphi).$$

This is the mathematical model of an elastic membrane spanned over the horizontal domain $\Omega \subset \mathbb{R}^2$ and fixed along the boundary $\partial\Omega$ at values g undergoing a vertical deflection $u(x)$ under a vertical load density $f(x)$.

2.1 Examples of nonlinear problems

2.1.1 Minimization problems

The linear boundary value problem (2.0.1) has *nonlinear* extensions of different types:

a) The energy functional $E(\cdot)$ is minimized only over a convex subset V_* of the function space V. A typical example is obtained for

$$V = H_0^1(\Omega), \quad V_* = \{v \in V \mid v \geq \psi \ \text{ a. e. in } \Omega\},$$

where ψ is a prescribed obstacle function. This so-called "obstacle problem" describes the deflection u of an elastic membrane spanned over a rigid obstacle ψ. This can also be written in a certain variational form as will be shown below.

b) The energy functional $E(\cdot)$ is not quadratic, e.g.,

i) Minimal surface problem: The membrane has the surface measure

$$F(u) = \int_\Omega \sqrt{1 + |\nabla u|^2}\, dx.$$

The problem of determining the deflection u, which results in the minimal surface is called "minimal surface problem".

$$\min \int_\Omega \sqrt{1+|\nabla u|^2}\, dx \quad \text{on } H^1(\Omega), \quad v = g \text{ on } \partial\Omega. \tag{2.1.3}$$

ii) "p-Laplace problem":

$$\min \int_\Omega \left(1+|\nabla u|^2\right)^{p/2} dx - \int_\Omega f u\, dx \quad \text{on } H_0^1(\Omega). \tag{2.1.4}$$

or, more general:

$$\min \int_\Omega F(\cdot, u.\nabla u)\, dx \quad \text{on } H_0^1(\Omega). \tag{2.1.5}$$

If $u \in H_0^1(\Omega)$ is a minimizer of the functional $E(\cdot)$, then there necessarily holds

$$\frac{d}{dt} E(\cdot, u+t\varphi, \nabla(u+t\varphi))\Big|_{t=0} = 0 \quad \forall \varphi \in H_0^1(\Omega).$$

For the minimal surface problem this reads

$$\int_\Omega \frac{\nabla u \cdot \nabla \varphi}{\sqrt{1+|\nabla u|^2}}\, dx = 0, \quad \forall \varphi \in H_0^1(\Omega),$$

from which we get as necessary optimality condition the nonlinear boundary value problem (provided that the minimizer u is sufficiently regular)

$$-\nabla \cdot \Big(\frac{\nabla u}{\sqrt{1+|\nabla u|^2}}\Big) = 0 \text{ in } \Omega, \quad u = g \text{ on } \partial\Omega.$$

This is called the "Euler equation" of the minimization problem. Analogously, we obtain for the general energy functional the variational equation

$$\int_\Omega \left(F'_{\nabla u}(x,u,\nabla u)\cdot\nabla\varphi + F'_u(x,u,\nabla u)\varphi\right) dx = 0, \quad \forall \varphi \in H_0^1(\Omega),$$

and the corresponding Euler equation

$$-\nabla \cdot F'_{\nabla u}(x,u,\nabla u) + F'_u(x,u,\nabla u) = 0 \text{ in } \Omega, \quad u = g \text{ on } \partial\Omega.$$

This is called a quasi-linear second-order differential equation in "divergence form".

2.1.2 Nonlinear diffusion-reaction-transport problems

The symmetric operator Δ is supplemented by lower order nonlinear terms.

i) Reaction-diffusion problem (mathematical model in cell biology)

$$-D\Delta u = f(\cdot, u), \quad f(\cdot,u) = \frac{u}{1+u^2}.$$

ii) Transport-diffusion problem

$$-D\Delta u + \beta(u)\cdot u = f.$$

a) Example (nonstationary) Burgers equation in $\mathbb{R}^1$,

$$\partial_t u - \nu\partial_x^2 u + u\partial_x u = 0 \quad \text{on } \mathbb{R}^1.$$

Exact solution (picture of graph of $\tanh(x)$)

$$u(x,t) = 1 - \tanh\Big(\frac{x-t}{2\nu}\Big), \qquad u(x,t)\to 2\ (x\to-\infty),\ \ u(x,t)\to 0\ (x\to\infty).$$

Derivatives

$$\begin{aligned} d_x \tanh(x) &= 1-\tanh^2(x),\\ \partial_x u(x,t) &= -\frac{1}{2\nu}\Big(1-\tanh^2\Big(\frac{x-t}{2\nu}\Big)\Big) = -\partial_t u(x,t). \end{aligned}$$

$$\begin{aligned} \nu\partial_x^2 u(x,t) &= \nu\partial_x\big(\partial_x u(x,t)\big)\\ &= -\frac{1}{2}\partial_x\Big(1-\tanh^2\Big(\frac{x-t}{2\nu}\Big)\Big) = \tanh\Big(\frac{x-t}{2\nu}\Big)\partial_x\tanh\Big(\frac{x-t}{2\nu}\Big)\\ &= -\tanh\Big(\frac{x-t}{2\nu}\Big)\partial_x\Big(1-\tanh\Big(\frac{x-t}{2\nu}\Big)\Big)\\ &= \Big(1-\tanh\Big(\frac{x-t}{2\nu}\Big)\Big)\partial_x\Big(1-\tanh\Big(\frac{x-t}{2\nu}\Big)\Big) - \partial_x\Big(1-\tanh\Big(\frac{x-t}{2\nu}\Big)\Big)\\ &= u(x,t)\partial_x u(x,t) + \partial_t u(x,t). \end{aligned}$$

The functiona $u(x) = 1-\tanh\big(\frac{x}{2\nu}\big)$ is solution of the stationary Burgers-type equation

$$-\nu\partial_x^2 u + (u-1)\partial_x u = 0.$$

b) Example (stationary) Navier-Stokes equation in $\mathbb{R}^d$ ($d=2$ or 3)

$$-\nu\Delta v + v\cdot\nabla v + \nabla p = f, \quad \nabla\cdot v = 0 \quad \text{in } \Omega, \quad v = 0 \ \text{ on } \partial\Omega.$$

v, p velocity vector and scalar pressure in an incompressible viscous Newtonian fluid. "No-slip" condition along the boundary of the flow domain. Examples "lid-driven cavity" and channel flow.

Exact solutions in special configurations:

i) Couette flow (parallel shear flow) Flow between two infinite plates (parallel to the (x_1, x_2)-plane with constant distance $L = 1$, where the bottom one is kept fixed and the upper one is moved with constant speed in x_1-direction. The velocity vector

$$v_1(x) = x_3, \quad v_2(x) = v_3(x) = 0\,,$$

is obviously divergence free and satisfies the no-slip condition at the plates. Together with the trivial pressure $p = 0$ it satifies the Navier-Stokes equation.

$$-\nu\Delta v + v\cdot\nabla v + \nabla p = 0\,.$$

ii) Poiseuille flow (parabolic channel flow) Parallel flow through an infinite pipe parallel to the x_1-axis in 3d with circular cross section with radius $R = 1$. The velocity vector

$$v_1(x) = 1 - (x_2^2 + x_3^2), \quad v_2 = v_3 = 0\,,$$

is divergence free and satisfies the np-slip condition along the wall of the pipe. We have

$$v \cdot \nabla v \equiv 0, \quad \Delta v_1 = -4\,,$$

so that this velocity together with the linear pressure $p(x) := -4\nu x_1$ satisfies the Navier-Stokes equation:

$$-\nu\Delta v + v \cdot \nabla v + \nabla p = 0\,,$$

These two examples are among the very few for which solutions of the Navier-Stokes equations can be written in closed form (without using series representations). They demonstrate that in certain circumstances there are stationary solutions for arbitrary parameter values $nu > 0$. Below, we will see that this is also the case in more general situations (e. g., domains).

2.1.3 Von Karman model in plate bending theory

In the case of "small" deflections of thin plates, $|u| \ll d \ll 1$, assuming linear material behavior (i. e., linear stress-strain relation) the governing model is the Kirchhoff plate equation , a linear fourth-order PDE. This reads with clamped boundary conditions:

$$\Delta^2 u = \frac{p}{D} \text{ in } \Omega, \quad u = \partial_n u = 0 \text{ on } \partial\Omega. \tag{2.1.6}$$

The corresponding variational formulation uses the Sobolev space $V = H_0^2(\Omega)$ and the energy form

$$a(u,\varphi) = (\Delta u, \Delta\varphi) + (1-\sigma)(2\partial_1\partial_2 u\partial_1\partial_2\varphi - \partial_1^2 u\partial_2^2\varphi - \partial_2^2 u\partial_1^2\varphi)\,dx,$$

with a materials parameter σ,. We note that that part of the nervy form which contains the parameter σ vanishes for functions in $H_0^2(\Omega)$. Therefore, the "weak" solution of the associated variational equation

$$a(u,\varphi) = (f,\varphi) \quad \forall\varphi \in V.$$

corresponds to the plate problem (2.1.6), which is formally independent of the parameter σ.

In the description of thin plates with "large" (relative to the thickness $< d$) deflection $|u| \approx d$ the deformation of the middle surface of the plate cannot be neglected anymore. The corresponding geometrically semi-linear theory goes back to von Karman (1910). It is an extension of the linear Kirchhoff theory, which results in the following system of two fourth-order equations:

$$\Delta^2 u = \frac{p}{D} + \frac{1}{D}(\partial_2^2\Psi\partial_1^2 u + \partial_1^2\Psi\partial_2^2 u - 2\partial_1\partial_2\Psi\partial_1\partial_2 u), \tag{2.1.7}$$

$$\Delta^2\Psi = Ed(\partial_1\partial_2 u^2 - \partial_1^2 u\partial_2^2 u), \tag{2.1.8}$$

for the normal deflection u of the plate's middle surface and for the so-called "stress function" Ψ, from which the horizontal distortion of the middle surface can be derived. The corresponding

boundary conditions are

$$u = \partial_n u = 0 \quad \text{on } \partial\Omega, \qquad \Psi = \partial_n \Psi = 0 \quad \text{on } \partial\Omega. \tag{2.1.9}$$

2.1.4 Eigenvalue problems

The eigenvalue problem of the Laplacian operator

$$-\Delta v = \lambda v \ \text{ in } \Omega, \quad v = 0 \ \text{ on } \partial\Omega, \tag{2.1.10}$$

has the variational formulation in the space $v = H_0^1(Omega)$: Find a pair $\{v, \lambda\} \in V \times \mathbb{R}$ such that

$$a(v, \varphi) = (\nabla v, \nabla \varphi) = \lambda (v, \varphi) \quad \forall \varphi \in V. \tag{2.1.11}$$

Because of the quadratic form λu this is a generically nonlinear problem. It corresponds to the nonlinear minimization problerm (Rayleigh quatient)

$$\min_{v \in V} R(v) = \lambda_{\min}, \qquad R(v); = \frac{a(v,v)}{\|v\|^2}.$$

Alternatively it may also be written as nonlinear system of variational equations

$$\begin{aligned} a(v, \varphi) - \lambda (v, \varphi) = 0 \quad \forall \varphi \in V, \\ (\|v\|^2 - 1)\chi = 0 \quad \forall \chi \in \mathbb{R}. \end{aligned} \tag{2.1.12}$$

This last formulation also makes sense in the nonsymmetric case, formulated over the field $\mathbb{C}$, and can be used as starting point of numerical approximations.

2.2 Convex minimization problems and variational inequalities

We consider the following abstract setting: Let be given a normed vector space V with norm $\|\cdot\|$, a continuous (bounded) bilinear form $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ and a continuous (bounded) linear form $l(\cdot) : V \to \mathbb{R}$,

$$|a(v,w)| \le \beta \|v\| \|w\|, \quad |l(v)| \le \gamma \|v\|, \quad v, w \in V,$$

and a (nonempty) subset $M \subset V$. With the abstract "energy functional"

$$J(u) := \tfrac{1}{2} a(u,u) - l(u),$$

we consider the minimization problem

$$\min J(u) \quad \text{on} \quad M. \tag{2.2.13}$$

Theorem 2.1: *Under the following conditions the abstract minimization problem (2.2.13) possesses a unique solution:*
(i) The normed vector space V is complete (i. e. a Banach space).
(ii) The subset $M \subset V$ is convex and closed.

(iii) The bilinear form $a(\cdot,\cdot)$ is symmetric and V-ellipic, i. e., there exists some constant $\alpha > 0$ such that

$$a(u,u) \geq \alpha \|u\|^2, \quad u \in V.$$

Proof 2.1: The bilinear form $a(\cdot,\cdot)$ defines a scalar product on V and the associated norm $\|\cdot\|_a := a(\cdot,\cdot)^{1/2}$ is equivalent to the norm $\|\cdot\|$ of V. Then, the space V equipped with this scalar product is a complete inner product space, i. e. a Hilbert space. By the Riesz representation theorem, there exists an element $v_l \in V$ such that

$$l(\varphi) = a(v_l, \varphi), \quad \varphi \in V.$$

Hence the functional $J(\cdot)$ can be rewritten in the form

$$J(v) = \tfrac{1}{2}a(v,v) - l(v) = \tfrac{1}{2}a(v,v) - a(v_l,v) = \tfrac{1}{2}a(v-v_l, v-v_l) - \tfrac{1}{2}a(v_l,v_l)$$

This shows that the minimization of $J(\cdot)$ on M is equivalent to minimizing the distance between the element v_f and the set M with respect to the norm $\|\cdot\|_a$, i. e., the solution is simply the projection of v_l onto the set M with respect to the scalar product $a(\cdot,\cdot)$. By the "Hilbert projection theorem" stated below such a projection exists and is uniquely determined. Q.E.D.

Lemma 2.1 (Hilbert projection theorem): *Let V be a Hilbert space and $M \subset V$ a closed convex subset. Corresponding to any vector $x \in V$, there is a unique vector $m_x \in M$ such that*

$$\|x - m_x\| \leq \min_{m \in M} \|x - m\|. \tag{2.2.14}$$

Furthermore, if $M \subset V$ is a closed subspace a necessary and sufficient condition that $m_x \in M$ is the unique minimizing vector is that $x - m_x$ is orthogonal to M.

Proof 2.2: i) Existence of m_x: Let $\delta \geq 0$ be the distance between x and M, and $(x_n)_{n\in\mathbb{N}}$ a sequence in M such that

$$\|x - x_n\|^2 \leq \delta^2 + 1/n, \quad n \in \mathbb{N}.$$

Let $n, m \in \mathbb{N}$ be arbitrary. Then,

$$\|x_n - x_m\|^2 = \|x_n - x\|^2 + \|x_m - x\|^2 - 2(x_n - x, x_m - x)$$

and

$$4\|\tfrac{1}{2}(x_n + x_m) - x\|^2 = \|x_n - x\|^2 + \|x_m - x\|^2 + 2(x_n - x, x_m - x).$$

This implies that

$$\|x_n - x_m\|^2 = 2\|x_n - x\|^2 + 2\|x_m - x\|^2 - 4\|\tfrac{1}{2}(x_n + x_m) - x\|^2.$$

Consequently, using the definition of δ and the convexity of M,

$$\|x_n - x_m\|^2 \leq 2(\delta^2 + 1/n) + 2(\delta^2 + 1/m) - 4\delta^2 = 2(1/n + 1/m).$$

This shows that the sequence $(x_n)_{n\in\mathbb{N}}$ is a Cauchy sequence in the closed subset $M \in V$ and therefore possesses a limit $m_x \in M$ with minimal distance to x,

ii) Uniqueness of m_x: Let m_x^1, m_x^2 be two minimizers. Then,

$$\|m_x^2 - m_x^1\|^2 = 2\|m_x^1 - x\|^2 + 2\|m_x^2 - x\|^2 - 4\|\tfrac{1}{2}(m_x^1 + m_x^2) - x\|^2.$$

Since $\frac{1}{2}(m_x^1 + m_x^2) \in M$, we have

$$\|\tfrac{1}{2}(m_x^1 + m_x^2) - x\|^2 \geq \delta^2,$$

and therefore,

$$\|m_x^2 - m_x^1\|^2 \leq 2\delta^2 + 2\delta^2 - 4\delta^2 = 0.$$

This means $m_x^1 = m_x^2$.
iii) Orthogonality characterization: Let $m_x \in M$ satisfy

$$(m_x - x, m) = 0, \quad m \in M.$$

Then,

$$\|x - m\|^2 = \|m_x - x\|^2 + \|m - m_x\|^2 + 2(m_x - x, m - m_x) = \|m_x - x\|^2 + \|m - m_x\|^2,$$

which shows that m_x is a minimizer. On the other hand, for a minimizer m_x, there holds with any $m \in M$ and $t \in \mathbb{R}$ that

$$\|(m_x + tm) - x\|^2 - \|m_x - x\|^2 = 2t(m_x - x, m) + t^2\|m\|^2 = 2t(m_x - x, m) + \mathcal{O}(t^2) \geq 0.$$

This implies that necessarily $(m_x - x, m) = 0$.

Theorem 2.2: *a) An element $u \in M \subset V$ is the solution of the abstract minimization problem (2.2.13) if ans only if it satisfies the variational inequality*

$$a(u, v - u) \geq l(v - u) \quad \forall v \in M. \tag{2.2.15}$$

b) In case that M is a closed subspace the necessary and sufficient relation becomes a variational equality,

$$a(u, v) = l(v) \quad \forall v \in M. \tag{2.2.16}$$

Proof 2.3: a) Let $u \in M$ be the minimizer of (2.2.13). Since M is convex, for any $v \in M$ all elements of the form $u + t(v - u) = tv + (1 - t)u$, $t \in [0, 1]$, are in M. Then, the continuously differentiable function

$$j(t) := J(u + t(v - u)), \quad t \in [0, 1],$$

has a minimum at $t = 0$, and therefore,

$$\frac{d}{dt} j(t)\Big|_{t=0} = \frac{d}{dt}\big(a(u + t(v - u), u + t(v - u)) - l(u + t(v - u))\big)\Big|_{t=0} \geq 0.$$

This implies the asserted variational inequality

$$a(u, v - u) \geq l(v - u) \quad \forall v \in M.$$

Let now this variational inequality be satisfied by some $u \in M$. Then, for any $v \in M$ there

holds

$$\begin{aligned} 0 &\le a(u, v-u) - l(v-u) \\ &= -a(u,u) + l(u) + a(u,v) - l(v) \\ &= -\tfrac{1}{2}a(u,u) + l(u) - \tfrac{1}{2}a(u,u) + a(u,v) - \tfrac{1}{2}a(v,v) - l(v) + \tfrac{1}{2}a(v,v) \\ &= -J(u) - \tfrac{1}{2}\|u-v\|_a^2 + J(v), \end{aligned}$$

and thus

$$J(u) + \tfrac{1}{2}\|u-v\|_a^2 \le J(v).$$

This shows that u is a strict minimizer of $J(\cdot)$ on M.

b) We use the inequality (2.2.15) with $u+v \in M$ for $v, -v \in M$ to obtain

$$a(u,v) \ge l(v), \quad a(u,v) \le l(v) \quad \forall v \in M.$$

This implies the asserted variational equation (2.2.16).

Remark 2.1: In case that the subset $M \subset V$ is a closed convex cone with vertex at the origin, i. e., $u, v \in M$ implies $\alpha u + \beta v \in M$ for $\alpha, \beta \in \mathbb{R}_+$, then necessary and sufficient variational inequality for the minimizer $u \in M$ takes the form

$$a(u,v) \ge l(v) \quad \forall v \in M, \qquad a(u,u) = f(u). \tag{2.2.17}$$

This is an intermediate result of the above proof of the variational equation (2.2.16).

Remark 2.2: If the bilinear form $a(\cdot,\cdot)$ is nonsymmetric, we may consider the variational inequality

$$a(u, v-u) \ge l(v-u) \quad \forall v \in M,$$

though, in this case, it is not related to a minimization problem. The result of Theorem 2.1 can be extended to this situation (if V itself is a Hilbert space) in such that the variational inequality has a unique solution in M. In the case that M is a linear subspace or even more special that $M = V$ this is just the content of the well-known "Lax-Milgram lemma".

2.2.1 Approximation of abstract variational inequalities

First, we develop an error estimate for the above abstract variational inequality, from which afterwards concrete error estimates for the approximation of obstacle problems will be derived.

We consider the following somewhat more special setting: Let V be a Hilbert space with norm $\|\cdot\|_V$, $a(\cdot,\cdot)$ a bounded, symmetric, V-elliptic bilinear form and $f(\cdot)$ a bounded linear form on V. Then, for any closed convex subset $M \subset V$ there exists unique minimizer $u \in M$ of the abstract energy functional $J(\cdot) := \frac{1}{2}a(\cdot,\cdot) - f(\cdot)$ on M, which is characterized by the variational inequality

$$a(u, v-u) \ge f(v-u) \quad \forall v \in M. \tag{2.2.18}$$

For discretizing this problem, we consider a finite dimensional subspace $V_h \subset V$ and a (non empty) closed, convex subset $M_h \subset V_h$ not necessarily contained in M. The approximation

$u_h \in M_h$ to u is then defined as the unique solution of the "discrete" variational inequality

$$a(u_h, v_h - u_h) \geq f(v_h - u_h) \quad \forall v_h \in M_h. \tag{2.2.19}$$

The bilinear form $a(\cdot,\cdot)$ defines an operator $A : V \to V^*$ through the relation

$$Au(\varphi) := a(u, \varphi), \quad \varphi \in V.$$

In this sense $Au - f$ may be considered as a functional in the dual space V^*. Notice that in the present situation, we generally do not have $Au = f$.

Theorem 2.3: *For the approximation of the abstract variational inequality, there holds*

$$\begin{aligned} \|u - u_h\|_V \leq c\Big(&\inf_{v_h \in M_h} \{\|u - v_h\|_V^2 + |(Au - f)(u - v_h)|\} \\ &+ \inf_{v \in M} |(Au - f)(u_h - v)|\Big)^{1/2}, \end{aligned} \tag{2.2.20}$$

with a constant c independent of the choice of V_h and M_h.

Proof: We have

$$\alpha\|u - u_h\|_V^2 \leq a(u - u_h, u - u_h) = a(u,u) + a(u_h,u_h) - a(u,u_h) - a(u_h,u),$$

and by the variational inequalities satisfied by u and u_h:

$$\begin{aligned} a(u,u) &\leq a(u,v) + f(u - v), \quad v \in M, \\ a(u_h,u_h) &\leq a(u_h,v_h) + f(u_h - v_h), \quad v_h \in M_h. \end{aligned}$$

Therefore, we conclude that for all $v \in M$ and $v_h \in M_h$:

$$\begin{aligned} \alpha\|u - u_h\|_V^2 &\leq a(u, v - u_h) + a(u_h, v_h - u) + f(u - v) + f(u_h - v_h) \\ &= a(u, v - u_h) - f(v - u_h) + a(u, v_h - u) - f(v_h - u) + a(u_h - u, v_h - u) \\ &= (f - Au, u_h - v) + (f - Au, u - v_h) + a(u - u_h, u - v_h), \end{aligned}$$

and, consequently,

$$\alpha\|u - u_h\|_V^2 \leq |(f - Au)(u - v_h)| + |(f - Au)(u_h - v)| + \beta\|u - u_h\|_V\|u - v_h\|_V.$$

Since

$$\|u - u_h\|_V\|u - v_h\|_V \leq \frac{1}{2}\Big\{\frac{\alpha}{\beta}\|u - u_h\|_V^2 + \frac{\beta}{\alpha}\|u - v_h\|_V^2\Big\},$$

we obtain, by combining the two previous inequalities,

$$\frac{\alpha}{2}\|u - u_h\|_V^2 \leq |(f - Au)(u - v_h)| + |(f - Au)(u_h - v)| + \frac{\beta^2}{2\alpha}\|u - v_h\|_V^2.$$

From this, we conclude the asserted estimate. Q.E.D.

Remark 2.3: (i) The above proof also covers the case of a nonsymmetric bilinear form $a(\cdot,\cdot)$.

(ii) If the inclusion $M_h \subset M$ holds, then the difficult term $\inf_{v\in M} |(Au-f)(u_h-v)|^{1/2}$ vanishes.

(iii) In the case $M = V$, then $Au = f$ and (2.2.20) reduces to the usual best-approximation extimate of the linear situation.

2.2.2 Application to obstacle and Signorini problem

Obstacle problem

We consider the obstacle problem on a convex polygonal domain $\Omega \subset \mathbb{R}^2$, with the setting

$$V = H_0^1(\Omega), \quad M = \{v \in V,\ v \geq \psi \text{ a. e. on } \Omega\}, \quad H = L^2(\Omega),$$

with norms $\|v\|_V := \|\nabla v\|_{L^2} = \|v\|_1$, $\|v\|_H := \|v\|_{L^2} = \|v\|$, and

$$a(u,\varphi) = (\nabla u, \nabla\varphi), \quad f(\varphi) = (f,\varphi),$$

with a function $f \in L^2(\Omega)$ and a smooth obstacle function $\psi \in H_0^1(\Omega) \cup H^2(\Omega)$. Further, let $V_h|$, be the usual spaces of piecewise linear finite elements on a quasi-uniform sequence of triangulations $\{\mathbb{T}_h\}_{h>0}$. We choose the discrete set

$$M_h := \{v_h \in V_h \,|\, v_h \geq \psi_h\},$$

where $I_h\psi$ is the usual piecewise linear nodal interpolation of ψ, which is well defined as $\psi \in H^2(\Omega)$. Generally, M_h is not contained in M

Corollary 2.1 (Obstacle problem): *Suppose for the minimizer of the obstacle problemthat $u \in H^2(\Omega)$. Then, for quasi-uniform sequence of triangulations there holds the error estimate*

$$\|u - u_h\|_1 \leq c(f,u,\psi)h, \tag{2.2.21}$$

with a constant $c(f,u,\psi)$ independent of h.

Proof: By assumption, we have $f \in H = L^2(\Omega)$ and $u \in H^2(\Omega)$. Therefore,

$$Au(v) := (\nabla u, \nabla v) = -(\Delta u, v), \quad v \in V = H_0^1(\Omega),$$

and thus

$$|(Au)(v)| \leq \|\Delta u\|\|v\|.$$

This implies that indeed $Au \in H$. Further, by construction $I_h u \in M_h$ and therefore

$$\begin{aligned}\inf_{v_h\in M_h} \left\{\|u - v_h\|_1^2 + \|Au - f\|\|u - v_h\|\right\} &\leq \|u - I_h u\|_1^2 + \|Au - f\|\|u - I_h u\| \\ &\leq ch^2\|u\|_2^2 + c\|Au - f\|h^2\|u\|_2.\end{aligned}$$

It remains to evaluate the term $\inf_{v\in M}\|u_h - v\|$. To this end, we introduce the function $u_h^* := \max\{u_h, \psi\}$, which satisfies $u_h^* \geq \psi$ on Ω. Since both u_h, $\psi \in H_0^1(\Omega)$ also $u_h^* \in H_0^1(\Omega)$ (nontrivial result) and then by construction also $u_h^* \in M$. With the set $\Lambda_h := x \in \Omega,\ u_h \leq \psi\}$

observing $u_h - u_h^* = 0$ on $\Omega \setminus \Gamma_h$, there holds,

$$\|u_h - u_h^*\|^2 = \int_{\Lambda_h} |u_h - \psi|^2 \, dx.$$

By construction, we have $u_h - I_h\psi \geq 0$ and consequently on Λ_h:

$$0 < |\psi - u_h| = \psi - u_h \leq \psi - I_h\psi = |\psi - I_h\psi|.$$

Thus,

$$\|u_h - u_h^*|^2 \leq \int_{\Lambda_h} |\psi - I_h\psi|^2 \, dx \leq \|\psi - I_h\psi\|^2 \leq ch^4\|\psi\|_2^2,$$

and finally,

$$\inf_{v \in M} \|u_h - v\| \leq ch^2\|\psi\|_2,$$

which concludes the proof. Q.E.D.

Signorini problem

We consider a special Signorini problem again on a convex polygonal domain $\Omega \subset \mathbb{R}^2$, with the setting

$$V = H^1(\Omega), \quad M = \{v \in V,\ v \geq 0 \text{ a. e. on } \partial\Omega\}, \quad H = L^2(\Omega),$$

with norms $\|v\|_V := \|v\|_1 = \left(\|v\|^2 + \|\nabla v\|^2\right)^{1/2}$, $\|v\|_H := \|v\|_{L^2} = \|v\|$, and

$$a(u, \varphi) = (\nabla u, \nabla\varphi) + (u, \varphi), \quad f(\varphi) = (f, \varphi),$$

with a function $f \in L^2(\Omega)$. Further, let V_h be again the usual spaces of piecewise linear finite elements on a quasi-uniform sequence of triangulations $\{\mathbb{T}_h\}_{h>0}$. We choose the discrete set

$$M_h := \{v_h \in V_h \,|\, v_h \geq 0 \text{ on } \partial\Omega\},$$

which this time is contained in M.

Corollary 2.2 (Signorini problem): *Suppose for the minimizer of the Signorini problem that $u \in H^2(\Omega)$. Then, for quasi-uniform sequence of triangulations there holds the error estimate*

$$\|u - u_h\|_1 \leq c(f, u)h^{3/4}, \tag{2.2.22}$$

with a constant $c(f, u)$ independent of h.

Proof: In this special situation the abstract error estimate (2.2.20) reduces to

$$\|u - u_h\|_1 \leq c\Big(\inf_{v_h \in M_h} \left\{\|u - v_h\|_1^2 + |(Au - f)(u - v_h)|\right\}\Big)^{1/2}.$$

Again by assumption, we have $f \in H = L^2(\Omega)$ and $u \in H^2(\Omega)$. Therefore,

$$Au(v) := (\nabla u, \nabla v) + (u, v) = -(\Delta u, v) + (\partial_n u, v)_{\partial\Omega} + (u, v), \quad v \in V = H^1(\Omega),$$

and thus, employing the trace inequality $\|\partial_n u\|_{\partial\Omega} \le c\|u\|_2$,

$$|Au(v)| \le c\|u\|_2(\|v\| + \|v\|_{\partial\Omega}).$$

Further, by construction $I_h u \in M_h$ and therefore

$$\begin{aligned}\inf_{v_h\in M_h}\left\{\|u-v_h\|_1^2 + |(Au-f)(u-v_h)|\right\} &\le \|u-I_hu\|_1^2 + c\|u\|_2\|u-I_hu\|_{\partial\Omega}\\ &\le ch^2\|u\|_2^2 + c\|u\|_2 h^{3/2}\|u\|_2,\end{aligned}$$

which completes the proof. Q.E.D.

2.3 The minimal surface problem

Let $\Omega \subset \mathbb{R}^2$ be a convex polygonal domain. We consider the minimization of the functional

$$J(u) := \int_\Omega \sqrt{1+|\nabla u|^2}\,dx$$

on the manifold $V_g := \{v \in V := H^1(\Omega)\,|\,v = g \text{ on } \partial\Omega\}$ for a prescribed function $g \in H^2(\Omega)$. This functional represents the surface content of the graph of the function $u = u(x)$, $x \in \Omega$, what suggests the name "minimal surface problem". In contrast to the obstacle problems the solvability analysis of this problem is difficult and not discussed here. For the following, we rather assume the existance of a minimizer in the manifold V_g.

Above, we have already seen that any solution $u \in V_g$ of the minimal surface problem necessarily satisfies the variational equation

$$\int_\Omega \frac{\nabla u\cdot\nabla\varphi}{\sqrt{1+|\nabla u|^2}}\,dx = 0 \quad \forall\varphi \in V_0 := H_0^1(\Omega). \tag{2.3.23}$$

This result can be expressed in the language of the Calculus of Variations in function spaces as follows: The gradient and the Hessian matrix of the function

$$f(x) := \sqrt{1+|x|^2}, \quad x \in \mathbb{R}^2,$$

act as follows:

$$\begin{aligned}(f'(x),\xi) &= \sum_{i=1}^2 \partial_i f(x) x_i\xi_i = \frac{x\cdot\xi}{(1+|x|^2)^{1/2}},\\ (f''(x)\xi,\xi) &= \sum_{i,j=1}^2 \partial_i\partial_j f(x)\xi_i\xi_j = \frac{|\xi|^2 + (x_2\xi_1 - x_1\xi_2)^2}{(1+|x|^2)^{3/2}}.\end{aligned}$$

Therefore, we see that

$$\frac{|\xi|^2}{(1+|x|^2)^{3/2}} \le (f''(x)\xi,\xi) \le |\xi|^2, \quad \xi \in \mathbb{R}^2,$$

and particularly that $f(\cdot)$ is strictly convex on bounded sets. With this notation by Taylor

expansion,

$$J(v+\varphi) - J(v) = \int_\Omega \left\{ \sqrt{1+|\nabla(v+\varphi)|} - \sqrt{1+|\nabla v|} \right\} dx$$
$$= \int_\Omega \frac{\nabla v \cdot \nabla \varphi}{\sqrt{1+|\nabla v|^2}}\, dx + \mathcal{R}(v,\varphi),$$

with a bounded remainder term

$$|\mathcal{R}(v,\varphi)| \le \frac{1}{2} \int_\Omega \|\nabla \varphi\|^2\, dx \le \tfrac{1}{2} \|\varphi\|_1^2.$$

Thus, the functional $J(\cdot)$ is differentiable on V with so-called "Fréchet derivative" $J'(\cdot)$ acting as a bounded functional on V like

$$J'(v)\varphi = \int_\Omega \frac{\nabla v \cdot \nabla \varphi}{\sqrt{1+|\nabla v|^2}}\, dx.$$

In this framework the above necessary optimality condition for the minimizer of the functional $J(\cdot)$ can be written in the form

$$J'(u)\varphi = 0, \quad \varphi \in H_0^1(\Omega). \tag{2.3.24}$$

Furthermore, there holds

$$J(u+\varphi) - J(u) \ge \int_\Omega \frac{|\nabla \varphi|^2}{(1+|\nabla u|^2)^{3/2}}\, dx,$$

so that any minimizer $u \in V + g$ (if it exists) is unique.

2.3.1 Finite element approximation

Let $\{\mathbb{T}_h\}_{h>0}$ be a quasi-uniform family of triangulations of the (for simplicity) polygonal domain Ω and $V_h \subset H^1(\Omega)$ the corresponding spaces of piecewise linear finite elements. With the natural nodal interpolation $g_h := I_h g \in V_h$ of g (well defined since $g \in H^2(\Omega)$ by assumption), we introduce the manifolds

$$V_{h,g_h} := \{ v_h \in V_h \,|\, v_h - g_h = 0 \ \text{ on } \partial\Omega \}.$$

Then, the discrete approximations u_h are defined as minimizers of the functional $J(\cdot)$ on V_{h,g_h}. For later purposes, we also introduce the discrete spaces $V_{h,0} := V_h \cap V_0$.

Theorem 2.4 (Discrete minimal surface problem): *There are uniquely determined minimizers $u_h \in V_h$ of the functional $J(\cdot)$ on the manifolds V_{h,g_h}.*

Proof: Let $v_h^* \in V_{h,g_h}$ be an arbitrary but fixed function. Then, since on V_h all norms are equivalent, there exists some $R > 0$ such that

$$J(v_h^*) \le J(v_h), \quad v_h \in V_{h,g_h},\ \|v_h\|_{H^1} \ge R.$$

Hence, minimizing $J(\cdot)$ on V_{h,g_h} is equivalent to minimizing it on the bounded subset $\{v_h \in V_{h,g_h} \,|\, \|v_h\|_{H^1} \le R\}$. Since this set is compact, the continuous functional $J(\cdot)$ has there a minimum. This minimum is also unique since the functional $J(\cdot)$ is strictly convex on bounded sets. Q.E.D.

Theorem 2.5: *Suppose that the solution $u \in V_g$ of the minimal surface problem exists and is in $H^2(\Omega) \cap W^{1,\infty}(\Omega)$. Then there holds the error estimate*

$$\|u - u_h\|_{H^1} \le c(u)h, \tag{2.3.25}$$

with a constant $c(u)$ independent of h.

Proof: (i) As consequences of the minimization property of the functions $u \in V_g$ and $u_h \in V_{h,g_h}$, we have the relations

$$\begin{aligned} J'(u)\varphi &= 0, \quad \varphi \in V_0, \\ J'(u_h)\varphi_h &= 0, \quad \varphi_h \in V_{h,0}. \end{aligned}$$

(ii) Next, we consider the quantity

$$\Delta_h := \Big(\int_\Omega \frac{|\nabla(u - u_h)|^2}{\sqrt{1 + |\nabla u_h|^2}}\, dx \Big)^{1/2},$$

and want to prove that $\Delta_h \le c(u)h$. Let $v_h \in V_{h,g_h}$ be arbitrary, so that $w_h := v_h - u_h \in V_{h,0}$. Then, again setting $f(x) := \sqrt{1 + |x|^2}$ and observing $J'(u)w_h = 0$,

$$\begin{aligned} \Delta_h^2 &= \int_\Omega \frac{|\nabla(u - u_h)|^2}{\sqrt{1 + |\nabla u_h|^2}}\, dx \\ &= \int_\Omega \frac{\nabla(u - u_h) \cdot \nabla(u - v_h)}{f(\nabla u_h)}\, dx + \int_\Omega \frac{\nabla(u - u_h) \cdot \nabla w_h}{f(\nabla u_h)}\, dx \\ &= \int_\Omega \frac{\nabla(u - u_h) \cdot \nabla(u - v_h)}{f(\nabla u_h)}\, dx + \int_\Omega \frac{\nabla u \cdot \nabla w_h}{f(\nabla u_h)}\, dx \\ &= \int_\Omega \frac{\nabla(u - u_h) \cdot \nabla(u - v_h)}{f(\nabla u_h)}\, dx + \int_\Omega \Big(\frac{1}{f(\nabla u_h)} - \frac{1}{f(\nabla u)} \Big) \nabla u \cdot \nabla w_h\, dx\,. \end{aligned}$$

For the first integral on the right, observing that $f(x) \ge 1$, we have

$$\Big| \int_\Omega \frac{\nabla(u - u_h) \cdot \nabla(u - v_h)}{f(\nabla u_h)}\, dx \Big| \le \int_\Omega \frac{|\nabla(u - u_h)|}{\sqrt{f(\nabla u_h)}}\, |\nabla(u - v_h)|\, dx \le \Delta_h \|\nabla(u - v_h)\|.$$

In order to estimate the second integral on the right, we note that

$$\begin{aligned}\frac{1}{f(\nabla u_h)} - \frac{1}{f(\nabla u)} &= \frac{f(\nabla u) - f(\nabla u_h)}{f(\nabla u_h) f(\nabla u)} \\ &= \frac{f(\nabla u)^2 - f(\nabla u_h)^2}{f(\nabla u_h) f(\nabla u)(f(\nabla u_h) + f(\nabla u))} \\ &= \frac{|\nabla u|^2 - |\nabla u_h|^2}{f(\nabla u_h) f(\nabla u)(f(\nabla u_h) + f(\nabla u))} \\ &= \frac{\nabla(u - u_h)}{f(\nabla u_h) f(\nabla u)} \cdot \frac{\nabla(u + u_h)}{f(\nabla u_h) + f(\nabla u)},\end{aligned}$$

and thus

$$\Big|\frac{1}{f(\nabla u_h)} - \frac{1}{f(\nabla u)}\Big| \le \frac{|\nabla(u - u_h)|}{f(\nabla u_h) f(\nabla u)}.$$

This is used to estimate

$$\begin{aligned}\Big|\int_\Omega \Big(\frac{1}{f(\nabla u_h)} - \frac{1}{f(\nabla u)}\Big)\nabla u \cdot \nabla w_h \, dx\Big| &\le \int_\Omega \frac{|\nabla u|}{f(\nabla u)} \frac{|\nabla(u - u_h)|}{\sqrt{f(\nabla u_h)}} \frac{|\nabla w_h|}{\sqrt{f(\nabla u_h)}} \, dx \\ &\le \gamma(u) \Delta_h \Big(\int_\Omega \frac{|\nabla w_h|^2}{f(\nabla u_h)} \, dx\Big)^{1/2} \\ &\le \gamma(u) \Delta_h \big(\Delta_h + \|\nabla(u - v_h)\|\big),\end{aligned}$$

where, since by assumption $u \in W^{1,\infty}(\Omega)$ (crucial at this point),

$$\gamma(u) := \operatorname{ess\,sup}_\Omega \frac{|\nabla u|}{f(\nabla u)} < 1.$$

Combining the foregoing estimates, we obtain

$$\Delta_h \le \gamma(u) \Delta_h + (1 + \gamma(u)) \|\nabla(u - v_h)\|).$$

Since $\gamma(u) < 1$, it follows that

$$\Delta_h \le c(u) \inf_{v_h \in V_h + I_h g} \|\nabla(u - v_h)\|,$$

where $c(u) := (1 + \gamma(u))((1 - \gamma(u))$. Since by assumption $u \in H^2(\Omega)$ the nodal interpolation $I_h u \in V_h$ is well defined and by construction satisfies $I_h u \in V_{h,g_h}$. Thus,

$$\inf_{v_h \in V_{h,g_h}} \|\nabla(u - v_h)\| \le ch \|u\|_2,$$

and therefore, $\Delta_h \le c(u) h$.

(iii) Next, we want to show that $\sup_\Omega |\nabla u_h| \le c(u)$. Let $T \in \mathbb{T}_h$ be an arbitrary cell. By the triangle inequality and the result of (ii),

$$\begin{aligned}\Big(\int_T \frac{|\nabla u_h|^2}{\sqrt{1 + |\nabla u_h|^2}} \, dx\Big)^{1/2} &\le \Delta_h + \Big(\int_T \frac{|\nabla u|^2}{\sqrt{1 + |\nabla u_h|^2}} \, dx\Big)^{1/2} \\ &\le c(u) h + \operatorname{ess\,sup}_T |\nabla u| \operatorname{meas}(T)^{1/2} \le c(u) h,\end{aligned}$$

Since $\nabla u_h|_T$ is constant, we can write

$$\int_T \frac{|\nabla u_h|^2}{\sqrt{1+|\nabla u_h|^2}}\,dx = \frac{|\nabla u_h|_T|^2}{\sqrt{1+|\nabla u_h|_T|^2}}\,\mathrm{meas}(T) \geq c\frac{|\nabla u_h|_T|^2}{\sqrt{1+|\nabla u_h|_T|^2}}\,h^2.$$

The foregoing estimates imply that necessarily

$$\max_{T\in\mathbb{T}_h} \frac{|\nabla u_h|_T|^2}{\sqrt{1+|\nabla u_h|_T|^2}} \leq c(u).$$

Observing $x^2/\sqrt{1+x^2} \to \infty$ for $x \to \infty$, we conclude the desired bound

$$\sup_\Omega |\nabla u_h| \leq c(u).$$

(iv) Finally, combining the foregoing results, we estimate as follows:

$$\begin{aligned}\|\nabla(u-u_h)\| &= \Big(\int_\Omega \frac{|\nabla(u-u_h)|^2}{\sqrt{1+|\nabla u_h|^2}}\,\sqrt{1+|\nabla u_h|^2}\Big)^{1/2}\\ &\leq \Delta_h\Big(\max_{T\in\mathbb{T}_h}\sqrt{1+|\nabla u_h|_T|^2}\Big)^{1/2} \leq c(u)h.\end{aligned}$$

(v) For estimating the full H^1-norm of the error, we use the well-known inequality

$$\|v\|_\Omega \leq c\big(\|\nabla v\|_\Omega + \|v\|_{\partial\Omega}\big), \quad v \in H^1(\Omega),$$

which, as generalization of the Poincaré inequality, can be proven by a similar argument as has been used in deriving the trace inequality. Employing this inequality for $u-u_h$ and observing that

$$\|u-u_h\|_{\partial\Omega} = \|g - I_h g\|_{\partial\Omega} \leq c(u)h^{3/2},$$

we obtain the asserted estimate $\|u-u_h\|_{H^1} \leq c(u)h$. Q.E.D.

Remark 2.4: By a much more involved argument one can prove the optimal-order L^2-error estimate

$$\|u-u_h\| \leq c(u)h^2. \tag{2.3.26}$$

2.4 Problems of monotone type

In this section, we consider convex minimization problems of the form

$$\min J(v) \quad \text{on} \quad V, \quad J(v) := \Big\{\frac{1}{p}\int_\Omega |\nabla v|^p\,dx - \int_\omega f v\,dx\Big\}, \tag{2.4.27}$$

or, as already stated above,

$$\min J(v) \quad \text{on} \quad V, \quad J(v); = \Big\{\frac{1}{p}\int_\Omega \big(1+|\nabla v|^2\big)^{p/2}\,dx - \int_\omega f v\,dx\Big\}, \tag{2.4.28}$$

for some $p \in (1,\infty)$ and an appropriate function space V. For $p = 2$ this corresponds to the usual quadatic minimization problem associated with the Laplacian operator. The limit case $p = 1$ is related to the minimal surface problem, which has turned out to be particularly difficult for theoretical analysis. Here, we will restrict ourselves to the case $p \geq 2$ and the first example corresponding to the so-called "p-Laplacian operator". The case $p < 2$ and the other example can be treated with slightly different arguments leading to similar results.

The natural solution space for this minimization problems corresponding to homogeneous boundary values is the Sobolev (Banach) space $V := H_0^{1,p}(\Omega)$, which may be defined as the closure of the space $C_0^\infty(\Omega)$ of test functions with respect to the norm

$$\|v\|_{1,p} := \big(\|v\|_{0,p}^p + \|v\|_{0,p}^p\big)^{1/p},$$

where $\|\cdot\|_p = \|\cdot\|_{0,p}$ denotes the usual L^p-norm over the domain Ω. As usual, for simplifying the analysis and the finite element approximation, the domain Ω is assumed to be two-dimensional, polygonal and convex. Further, the force term is assumed to be $f \in L^2(\Omega)$. The space V is equipped with the norm $\|v\|_V := \|\nabla v\|_p$, which inview of the L^p-version of the Poincaré inequality

$$\|v\|_p \leq c(p,\Omega)\|\nabla v\|_p, \quad v \in H^{1,p}(\Omega), \;\; 1 < p < \infty,$$

is equivalent to the norm $\|\cdot\|_{1,p}$.

Remark 2.5: The dual V^* (space of all linear continuous functionals) of the space $V = H_0^{1,p}(\Omega)$ denoted by $H^{-1,q}(\Omega)$ is likewise a Banach space. For $1 < p < \infty$ the bi-dual $(V^*)^*$ has a natural linear embedding into V, which is an isomophism, i. e., V is "reflexive". In reflexive Banach spaces any bounded sequence is "weakly compact", i. e., contains a weakly convergent subsequence, a property which will be used below in proving existence of minimizers for the functional $J(\cdot)$. The difficulty with the minimal surface problem is partially due to the fact that for $p = 1$ the corresponding Banach space $H_0^{1,1}(\Omega)$ is not reflexive.

As argued before, any minimizer of $J(\cdot)$ on V necessarily satisfies the variational equation (exercise)

$$\int_\Omega |\nabla u|^{p-2}\nabla u \cdot \nabla\varphi \, dx = \int_\Omega f\varphi \, dx \quad \forall \varphi \in V. \tag{2.4.29}$$

We want to embed this problem into a more abstract functional analytic setting. To this end, we note that the left-hand side of the equation (2.4.29) defines a (nonlinear) operator $A : V \to V^*$ by

$$Au(\varphi) := \int_\Omega |\nabla u|^{p-2}\nabla u \cdot \nabla\varphi \, dx, \quad \varphi \in V.$$

For seeing this, we use the general Hölder inequality (exercise)

$$\Big|\int_\Omega vw \, dx\Big| \leq \Big(\int_\Omega |v|^p \, dx\Big)^{1/p}\Big(\int_\Omega |w|^q \, dx\Big)^{1/q}, \quad v \in L^p(\Omega),\; w \in L^q(\Omega), \quad 1/p + 1/q = 1.$$

This is the same concept as that already used in the case of a linear bilinear form $a(\cdot,\cdot)$ leading to a likewise linear operator $A : V \to V^*$. The following lemma provides natural generalizations of the properties "V-ellipticity" and "boundedness" (resp. "Lipschitz continuity") for the nonlinear operator A.

Lemma 2.2: *For the operator* $A : V \to V^*$ *there hold the following inequalities:*

$$\alpha\|u-v\|_V^p \le (Au-Av)(u-v), \quad u.v \in V, \tag{2.4.30}$$

$$\|Au-Av\|_{V^*} \le \beta\big(\|u\|_V + \|v\|_V\big)^{p-2}\|u-v\|_V, \quad u.v \in V, \tag{2.4.31}$$

with some constants $\alpha, \beta > 0$.

Proof: (i) We introduce the auxiliary function

$$\varphi(\xi,\eta) := \frac{(|\xi|^{p-2}\xi - |\eta|^{p-2}\eta)\cdot(\xi-\eta)}{|\xi-\eta|^p}, \quad \xi,\eta \in \mathbb{R}^2,\ \xi \ne \eta.$$

We want to show that there is some constant $\alpha > 0$, such that

$$\alpha \le \varphi(\xi,\eta), \quad \xi,\eta \in \mathbb{R}^2,\ \xi \ne \eta. \tag{2.4.32}$$

From this, we obtain

$$\alpha|\xi-\eta|^p \le (|\xi|^{p-2}\xi - |\eta|^{p-2}\eta)\cdot(\xi-\eta),$$

and then by integration over Ω the estimate (2.4.30). Since

$$\varphi(0,\eta) = 1 \quad \text{for } \eta \ne 0,$$

it suffices to consider the case $\xi \ne 0$. Next, we prove that

$$\varphi(\xi,\eta) > 0 \quad \text{for } \xi \ne \eta,$$

This follows from the relations

$$\begin{aligned}(|\xi|^{p-2}\xi - |\eta|^{p-2}\eta)\cdot(\xi-\eta) &= |\xi|^p - (|\xi|^{p-2}+|\eta|^{p-2})(\xi\cdot\eta) + |\eta|^p \\ &\ge |\xi|^p - |\xi|^{p-1}|\eta| - |\eta|^{p-1}|\xi| + |\eta|^p \\ &= (|\xi|^{p-1} - |\eta|^{p-1})(|\xi| - |\eta|) \\ &> 0 \quad \text{for } |\xi| \ne |\eta|.\end{aligned}$$

Since the penultimate inequality is an equality if and only if $\eta = \mu\xi$ for some $\mu \in \mathbb{R}$, the only remaining case is that where $\eta = -\xi$, But then

$$(|\xi|^{p-2}\xi - |\eta|^{p-2}\eta)\cdot(\xi-\eta) = 4|\xi|^p > 0.$$

We may restrict ourselves without loss of generality to the case $\xi = \bar{\xi} = (1,0)$ since $\varphi(\lambda\xi,\lambda\eta) = \varphi(\xi,\eta)$ for all $\lambda > 0$ and since the Euclidean scalar product is invariant under rotations around the origin. In view of

$$\lim_{|\eta|\to\infty} \varphi(\bar{\xi},\eta) = 1,$$

it remains to to study the behavior of the function $\varphi(\bar{\xi},\eta)$ in the neighborhood of the point $\bar{\xi}$. To this end, let

$$\eta_1 = 1 + \rho\cos\theta, \quad \eta_2 = \rho\sin\theta.$$

Then, a simple computation shows that

$$\varphi(\bar{\xi},\eta) = \frac{1+(p-2)\cos^2\theta+\varepsilon(\rho,\theta)}{\rho^{p-2}},$$

with $\lim_{\rho\to 0}\varepsilon(\rho,\theta)=0$ uniformly for $\theta\in[9,2\pi)$. Therefore,

$$\lim_{\eta\to\bar{\xi}}\varphi(\bar{\xi},\eta)=1 \ \text{ for } \ p=2, \quad \lim_{\eta\to\bar{\xi}}\varphi(\bar{\xi},\eta)=\infty \ \text{ for } \ p>2,$$

and the desired inequality (2.4.32) follows from the foregoing results.

(ii) To prove the inequality (2.4.31), we introduce the auxiliary function

$$\psi(\xi,\eta) := \frac{\big||\eta|^{p-2}\eta-|\xi|^{p-2}\xi\big|}{|\eta-\xi|(|\eta|+|\xi|)^{p-2}}, \quad \xi,\eta\in\mathbb{R}^2, \ \ \xi\neq\eta.$$

We want to show that there is some constant $\beta>0$ such that

$$\psi(\xi,\eta)\le\beta, \quad \xi,\eta\in\mathbb{R}^2, \ \ \xi\neq\eta. \tag{2.4.33}$$

Since

$$\psi(0,\eta)=1 \quad \text{for } \ \eta\neq 0,$$

we can without loss of generality assume that $\xi\neq 0$. Further, it suffices to consider the case $\xi=\bar{\xi}=(1,0)$, since $\psi(\lambda\xi,\lambda\eta)=\psi(\xi,\eta)$ for $\lambda>0$ and since the Euclidian norm is invariant under rotations around the origin. There also holds

$$\lim_{|\eta|\to\infty}\psi(\bar{\xi},\eta)=1.$$

To study the behavior of the function $\psi(\bar{\xi},\eta)$ in the neighborhood of of the point $\bar{\xi}$, we let again

$$\eta_1=1+\rho\cos\theta, \quad \eta_2=\rho\sin\theta,$$

For this, we obtain

$$\psi(\bar{\xi},\eta)=2^{2-p}\big(1+p(p-2)\cos^2\theta\big)^{1/2}+\varepsilon(\rho,\theta),$$

with $\lim_{\rho\to 0}\varepsilon(\rho.\theta)=0$ uniformly for $\theta\in[0,2\pi)$ and therefore,

$$\limsup_{\eta\to\bar{\xi}}\psi(\bar{\xi}.\eta)<\infty.$$

These relations finally imply the desired estimate (2.4.33). As a consequence, we have

$$\big||\eta|^{p-2}\eta-|\xi|^{p-2}\xi\big|\le\beta|\xi-\eta|(|\xi|+|\eta|), \quad \xi,\eta\in\mathbb{R}^2. \tag{2.4.34}$$

To prove (2.4.31), we use the characterization

$$\|Au-Av\|_{V^*}=\sup_{w\in V}\frac{|(Au-Av)(w)|}{\|w\|_V}. \tag{2.4.35}$$

From the last inequality above, we conclude that

$$\begin{aligned}
|(Au - Av)(w)| &= \Big| \int_\Omega (|\nabla u|^{p-2}\nabla u - |\nabla v|^{p-2}\nabla v) \cdot \nabla w \, dx \Big| \\
&\le \int_\Omega \big| |\nabla u|^{p-2}\nabla u - |\nabla v|^{p-2}\nabla v \big| \,]\nabla w| \, dx \\
&\le \beta \int_\Omega |\nabla(u-v)|(|\nabla u| + |\nabla v|)^{p-2}|\nabla w| \, dx \\
&\le \Big(\int_\Omega |\nabla(u-v)|^p \, dx \Big)^{1/p} \Big(\int_\Omega (|\nabla u| + |\nabla v|)^p \, dx \Big)^{(p-2)/p} \Big(\int_\Omega |\nabla w|^p \, dx \Big)^{1/p} \\
&\le \beta \|u - v\|_V (\|u\|_V + \|v\|_V)^{p-2} \|w\|_V,
\end{aligned}$$

where the generalized Hölder inequality has been used with the exponents $1/p+(p-2)/p+1/p = 1$. In view of the characterization (2.4.35) this completes the proof of inequality (2.4.31). Q.E.D.

2.4.1 An abstract error analysis

Motivated by the concrete problem presented above, we now consider the following abstract setting. Let V be a reflexive Banach space with norm $\|\cdot\|_V$ and corresponding dual space V^* with norm $\|\cdot\|_{V^*}$. Further let be given a (generally nonlinear) operator $A : V \to V^*$, such that $A0 = 0$ (for simplicity), and an element $f \in V^*$ for which the equation

$$Au(\varphi) = f(\varphi) \quad \forall \varphi \in V, \tag{2.4.36}$$

is to be solved. For discretizing problem (2.4.36), we consider finite dimensional subspaces $V_h \subset V$ and the corresponding discrete problems

$$Au_h(\varphi_h) = f(\varphi_h) \quad \forall \varphi_h \in V_h. \tag{2.4.37}$$

Here, the existence of solutions $u \in V$ and $u_h \in V_h$ to these problems is assumed. Below, we will prove this to be actually true for the special case of the p-Laplace problem. We want to derive an error estimate for the present abstract setting. To this end, we pose the following conditions:

i) The mapping $A : V \to V^*$ is "strongly monotone". i. e., there exists a strictly increasing function $\chi : [0, \infty) \to \mathbb{R}$ with the properties $\chi(0) = 0$ and $\chi(t) \to \infty$ $(t \to \infty)$, such that

$$(Av - Aw)(v - w) \ge \chi(\|v - w\|_V)\|v - w\|_V, \quad v, w \in V. \tag{2.4.38}$$

In view of Lemma 2.2 the operator A corresponding to the p-Laplacian is strongly monotone with $\chi(t) = \alpha t^{p-1}$.

ii) The mapping $A : V \to V^*$ is "Lipschitz continuous (for bounded arguments)", i. e., for any ball $B_R(0) = \{v \in V \,|\, \|v\|_V \le R\}$ there exist a constant $\Gamma(R)$, such that

$$\|Av - Aw\|_{V^*} \le \Gamma(R)\|v - w\|_V, \quad v, w \in B_R. \tag{2.4.39}$$

In view of Lemma 2.2 the operator A corresponding to the p-Laplacian is Lipschitz continuous for bounded arguments with $\Gamma(R) = \beta(2R)^{p-2}$.

Theorem 2.6 (Abstract error estimate): *Under the above assumptions there holds the error estimate*

$$\chi(\|u-u_h\|_V) \le c \inf_{v_h \in V_h} \|u-v_h\|_V, \tag{2.4.40}$$

with a constant c independent of the choice of V_h.

Proof: From the monotonicity and Lipschitz-continuity of A, we conclude that

$$\begin{aligned}\chi(\|u\|_V)\|u\|_V &= \chi(\|u-0\|_V)\|u-0\|_V \le (Au-A0)(u-0)\\ &= f(u) \le \|f\|_{V^*}\|u\|_V,\end{aligned}$$

and consequently,

$$\chi(\|u\|_V) \le \|f\|_{V^*}.$$

In the same way, we obtain a corresponding estimate for u_h. The assumed properties of the function $\chi(\cdot)$ imply that it is invertible, so that

$$\|u\|_V, \|u_h\|_V \le \chi^{-1}(\|f\|_{V^*}).$$

Next, we fix an arbitrary element $v_h \in V_h$. Using the equations satisfied by u and u_h, we obtain

$$(Au-Au_h)(\varphi_h) = 0, \quad \varphi_h \in V_h,$$

and, in particular,

$$(Au-Au_h)(u_h-v_h) = 0.$$

Hence, combining the foregoing results,

$$\begin{aligned}\chi(\|u-u_h\|_V)\|u-u_h\|_V &\le (Au-Au_h)(u-u_h) = (Au-Au_h)(u-v_h)\\ &\le \|Au-Au_h\|_{V^*}\|u-v_h\|_V \le \Gamma(\chi^{-1}(\|f\|_{V^*}))\|u-u_h\|_V\|u-v_h\|_V,\end{aligned}$$

from which we conclude the asserted estimate. Q.E.D.

Remark 2.6: The estimate (2.4.40) is another generalization of the "best-approximation" result in the linear case where $\chi(t) = \alpha t$.

2.4.2 Application to the p-Laplace problem

Now, we use the abstract error estimate stated in Theorem 2.6 for the p-Laplace problem. For a quasi-uniform family of meshes $\{\mathbb{T}_h\}_{h>0}$ of $\overline{\Omega}$ let $V_h \subset V$ be the usual finite element subspaces of piecewise linear elements.

First, we prove the existence of unique solutions for problem (2.4.36) and its discrete analogue (2.4.37). Recall the definition of the norm $\|v\|_V := \|\nabla v\|_p$ of the Banach space V.

Theorem 2.7 (Existence of solutions): *In case of the p-Laplacian operator, $2 \le p < \infty$, problem (2.4.36) and its discrete analogue (2.4.37) possess unique solutions $u \in V = H_0^{1,p}(\Omega)$ and $u_h \in V_h$, which are the unique minimizers of the corresponding functional $J(\cdot)$ on V and V_h, respectively.*

Proof: (i) For the functional $J(\cdot)$ there holds

$$J(v) = \tfrac{1}{p}\|v\|_V^p - f(v) \geq \tfrac{1}{p}\|v\|_V^p - \|f\|_{V^*}\|v\|_V,$$

from which we deduce that $J(\cdot)$ is bounded from below and that

$$J(v) \to \infty \quad \text{for } \|v\| \to \infty.$$

(ii) We show the strict convexity of $J(\cdot)$. Since the functional $f(\cdot)$ is linear and therefore convex, it suffices to establish the strict convexity of the mapping

$$v \in V \to \int_\Omega F(\nabla v)\, dx, \quad F : \xi \in \mathbb{R}^d \to \tfrac{1}{p}|\xi|^p.$$

Let v and w be two different functions in V such that

$$\text{meas}\, \Omega_* > 0, \quad \Omega_* := \{x \in \Omega \,|\, \nabla v \neq \nabla w\}$$

and let $\theta \in (0,1)$ be given. Then, by the strict convexity of the mapping $t \in \mathbb{R} \to |t|^p$,

$$\begin{aligned}\int_\Omega F(\theta\nabla v + (1-\theta)\nabla w)\, dx &= \int_{\Omega_*} \ldots dx + \int_{\Omega\setminus\Omega_*} \ldots dx \\ &< \int_\Omega \left\{\theta F(\nabla v) + (1-\theta)F(\nabla w)\right\} dx.\end{aligned}$$

We note at this stage that the strict convexity of the functional $J(\cdot)$ implies the uniqueness of possible minimizers on V and V_h.

(iii) We show the (Fréchet) differentiability of $.J(\cdot)$. The mapping F defined above is twice differentiable, with

$$\begin{aligned}\partial_i F(\xi) &= |\xi|^{p-2}\xi_i, \quad \xi \in \mathbb{R}^d, \\ \partial_i\partial_j F(\xi) &= (p-2)|\xi|^{p-4}\xi_i\xi_j, \quad \xi \in \mathbb{R}^d.\end{aligned}$$

Consequently, we can write

$$F(\xi+\eta) - F(\xi) = |\xi|^{p-2}\xi\cdot\eta + R(\xi,\eta),$$

with bounded remainder term

$$|R(\xi,\eta)| \leq c(p)(|\xi] + |\eta|)^{p-2}|\eta|^2.$$

Thus,

$$\int_\Omega F(\nabla(u+v))\, dx - \int_\Omega F(\nabla u)\, dx = \int_\Omega |\nabla u|^{p-2}\nabla u\cdot\nabla v\, dx + \mathcal{R}(u,v),$$

with

$$|\mathcal{R}(u,v)| \leq c(p)\int_\Omega (|\nabla u| + |\nabla v|)^{p-2}|\nabla v|^2\, dx.$$

Since

$$\Big|\int_\Omega |\nabla u|^{p-2}\nabla u\cdot\nabla v\, dx\Big| \leq \|\nabla u\|_V^{p-1}\|v\|_V,$$

the linear mapping

$$v \in V \;\to\; \int_\Omega |\nabla u|^{p-2}\nabla u \cdot \nabla v \, dx$$

is continuous for a fixed $u \in V$. Further, since

$$\int_\Omega (|\nabla u| + |\nabla v|)^{p-2}|\nabla v|^2 \, dx \le (\|u\|_V + \|v\|_V)^{p-2}\|v\|_V^2,$$

this shows that the mapping considered in (ii) and by that also the functional $J(\cdot)$ is differentiable. Its derivative has the form

$$J'(u)(v) = \int_\Omega |\nabla u|^{p-2}\nabla u \cdot \nabla v \, dx - f(v), \quad v \in V. \tag{2.4.41}$$

Hence, the (unique) minimizers u and u_h of $J(\cdot)$ on V and V_h must necessarily satisfy the variational equations

$$J'(u)(v) = 0 \quad \forall v \in V, \tag{2.4.42}$$

$$J'(u_h)(v_h) = 0 \quad \forall v_h \in V_h. \tag{2.4.43}$$

Further, in view of the convexity of the functional $J(\cdot)$, any solutions of these equations are also minimizers.

(iv) The "discrete" minimization problems have solutions. This is a consequence of the strict convexity of the functional $J(\cdot)$ and the property $J(v_h) \to \infty$ for $\|v_h\|_V \to \infty$. The argument is analogous to that already used in the context of the minimal surface problem. Further, letting $v_h = u_h$ in the variational equation (2.4.43) gives us $\|u_h\|_V^p = f(u_h)$ and, consequently, the uniform bound

$$\|u_h\|_V \le \|f\|_{V^*}^{1/(p-1)}.$$

(v) We use the previous results for extablishing the existence of a (unique) minimizer $u \in V$. The space $V = H^{1,p}(\Omega)$, $1 < p < \infty$, is reflexive. Then, the bounded set of approximate minimizers $u_h \in V_h$ contains a sequence $(u_{h_k})_k \in \mathbb{N}$, with $h_k \to 0$ $(k \to \infty)$, which converges weakly to some element $u \in V$,

$$\varphi \in V^* : \quad \varphi(u_{h_k}) \to \varphi(u), \quad k \to \infty.$$

We want to show that this limit u is a minimizer of $J(\cdot)$. To this end let $\psi \in C_0^\infty(\Omega)$ be arbitarily chosen and $I_{h_k}\psi \in V_{h_k}$ its nodal interpolant. For that there holds

$$J(u_h) \le J(I_{h_k}\psi).$$

Since the functional $J(\cdot)$ is continuous and convex, it is weakly lower semicontinuous, and consequently,

$$J(u) \le \liminf_{k\to\infty} J(u_{h_k}) \le \liminf_{k\to\infty} J(I_{h_k}\psi).$$

Observing that $\|\psi - I_{h_k}\psi\|_V \to 0$ $(k \to \infty)$ by the continuity of $J(\cdot)$ it follows that

$$\lim_{k\to\infty} J(I_{h_k}\psi) = J(\psi),$$

and thus

$$J(u) \le J(\psi), \quad \psi \in C_0^\infty(\Omega)$$

Since $C_0^\infty(\Omega)$ is dense in $H^{1,p}(\Omega)$ it follows that u is indeed minimizer of $J(\cdot)$. Q.E.D.

Corollary 2.3 (Qualitative convergence): *The minimizers $u_h \in V_h$ converge to the minimizer $u \in V = H^{1,p}(\Omega)$:*

$$\|u_h - u\|_V \to 0 \quad (h \to 0). \tag{2.4.44}$$

Proof: In the proof of Theorem 2.7, we showed that there is a sequence $(h_{h_k})_{k\in\mathbb{N}}$ of discrete solutions, which converges weakly to the minimizer of $J(\cdot)$, Since this limit is unique the whole family $(u_h)_{h>0}$ of discrete solutions converges weakly to this solution $u \in V$ as $h \to 0$. Therefore,

$$f(u) = \lim_{h\to 0} f(u_h).$$

Furthermore, we have for arbitrary $\psi \in C_0^\infty(\Omega)$:

$$\limsup_{h\to 0} J(u_h) \le \limsup_{h\to 0} J(I_h\psi) = \lim_{h\to 0} J(I_h\psi) = J(\psi).$$

Since $C_0^\infty(\Omega)$ is dense in $V = H^{1,p}(\Omega)$ the function ψ can be chosen arbitrarily close to u in the norm of V. Hence, we deduce that

$$J(u) \le \liminf_{h\to 0} J(u_h) \le \limsup_{h\to 0} J(u_h) \le J(u),$$

and consequently,

$$J(u) = \lim_{h\to 0} J(u_h).$$

From this, we also get

$$\|u\|_V = \lim_{h\to 0} \|u_h\|_V.$$

Since the space $V = H^{1,p}(\Omega)$ is uniformly convex this implies the strong convergence

$$\|u - u_h\|_V \to 0 \quad (h \to 0).$$

Q.E.D.

Remark 2.7: In the previous proofs some deep results from abstract Functional Analysis have been used:

0) The Banach space $H^{1,p}(\Omega)$, $1 < p < \infty$, is reflexive and uniformly convex.

1) In a reflexive Banach space bounded sets are weakly compact, i. e., contain weakly convergent sequences.

2) On a reflexive Banach space continuous and convex functionals are also weakly lower semicontinuous.

3) In a reflexive, uniformly convex Banach space a weakly convergent sequence $(u_k)_{kin\mathbb{N}}$ converges also strongly if the corrsponding sequence of norms converges, $\|u_k\| \to \|u\|$ $(k \to \infty)$.

For the proofs of these results the reader may consult the standard textbooks on Functional Analysis, particularly those on Convex Analysis.

Corollary 2.4 (Error estimate): *For the finite element approximation of the p-Laplac problem there holds the error estimate*

$$\|u - u_h\|_{1,p} \leq ch^{1/(p-1)}, \tag{2.4.45}$$

with a constant $c(\|f\|_{V^*}, \|u\|_{2,p})$.

Proof: The operator $A : V \to V^*$ corresponding to the p-Laplacian operator is monotone with the function $\chi(t) = \alpha t^{p-1}$. Hence the abstract error estimate of Theorem 2.6 yields

$$\alpha\|u - u_h\|_V^{p-1} \leq c \inf_{v_h \in V_h} \|u - v_h\|_V.$$

Further,

$$\inf_{v_h \in V_h} \|u - v_h\|_V \leq \|u - I_h u\|_V \leq ch\|u\|_{2,p},$$

which concludes the proof. Q.E.D.

Remark 2.8: The assumption $u \in H^{2,p}(\Omega)$ made in Corollary 2.4 may not be realistic in general situations. Solutions of nonlinear equations do not need to be smooth for smooth data (see exercise).

2.5 Exercises

Exercise 2.1: Consider the nonhomogeneous boundary value problem

$$-\Delta u = f \ \text{ in } \Omega, \quad u = g \ \text{ on } \partial\Omega,$$

on a sufficiently regular domain $\Omega \subset \mathbb{R}^n$. The right-hand side f is in $L^2(\Omega)$ and the boundary function g is given as the trace of a function $g \in H^1(\Omega)$. The corresponding energy forms

$$a(u, v) := (\nabla u, \nabla v), \quad E(u) := \tfrac{1}{2}a(u, u) - (f, u),$$

are defined on the Sobolev space $H^1(\Omega)$. Show by the arguments introduced in class that the minimization problem

$$\min E(u) \quad \text{on} \quad V_g := \{v \in H^1(\Omega),\ v|_{\partial\Omega} = g\},$$

possesses a unique solution u , which as "weak solution" of the boundary value problem is characterized by the variational equation

$$a(u, \varphi) = (f, \varphi) \quad \forall \varphi \in H_0^1(\Omega).$$

Exercise 2.2: Consider the minimization problem

$$\min \Big(\frac{1}{p} \int_\Omega \big(1 + |\nabla u|^2\big)^{p/2}\, dx - \int_\Omega f u\, dx \Big) \quad \text{on} \quad H_0^{1,p}(\Omega),$$

for some $p \in [1, \infty)$ and a given function $f \in L^2(\Omega)$. Derive the corresponding variational formulation (first-order necessary optimality condition) and, in case of a sufficiently regular minimizer u , the resulting nonlinear boundary value problem.

Exercise 2.3: The Sobolev space $H_0^2(\Omega)$ is defined as the closure of the space $C_0^\infty(\Omega)$ of all C^∞-functions with compact support in Ω with respect to the norm

$$\|v\|_{H^2} := \Big(\|\nabla^2 v\|_\Omega^2 + \|\nabla v\|_\Omega^2 + \|v\|_\Omega^2\Big)^{1/2}.$$

a) Show that this norm is equivalent to the semi-norm $\|\nabla^2 v\|_\Omega$ and on a convex (pogonal) domain also equivalent to the semi norm $\|\Delta v\|_\Omega$.

b) Show that for functions in $H_0^2(\Omega)$ there holds

$$2(\partial_1\partial_2 u, \partial_1\partial_2 v)_\Omega - (\partial_1^2 u, \partial_2^2 v)_\Omega - (\partial_2^2 u, \partial_1^2 v)_\Omega = 0.$$

Exercise 2.4: Apply the existence theorem for convex minimization problems provided in class to show existence and uniqueness of solutions for the obstacle problem in plate bending theory,

$$\min E(u) \quad \text{on } V_\psi := \{v \in H_0^2(\Omega),\ v \geq \psi\}$$

where, with some $\sigma \in [0,1]$,

$$E(u) := \tfrac{1}{2}\|\Delta u\|_\Omega^2 + (1-\sigma)\big(\|\partial_1\partial_2 u\|_\Omega^2 - (\partial_1^2 u, \partial_2^2 u)_\Omega\big) - (f,u),$$

and ψ is an admissible (i. e., positive only on a compact subset of Ω) smooth function describing an obstacle.

Exercise 2.5: Recall that the unique minimizer $u \in M$ of the quadratic functional $J(u) = \frac{1}{2}a(u,u) - l(u)$ on a convex, closed subset M of a Hilbert space V is characterized by the associated variational inequality

$$a(u, v-u \geq l(v-u) \quad \forall v \in M.$$

Show that if the subset M is a closed convex cone with vertex at the origin, i. e., $u, v \in M$ implies $\alpha u + \beta v \in M$ for $\alpha, \beta \in \mathbb{R}_+$, then the characterizing variational inequality takes the simplified form

$$a(u,v) \geq l(v) \quad \forall v \in M, \quad a(u,u) = f(u).$$

An examples of such a cone is the set $M = \{v \in H^1(\Omega) \,|\, u \geq 0 \text{ a. e. on } \Omega\} \subset H^1(\Omega)$.

Exercise 2.6: Consider the obstacle problem

$$\min\big\{\tfrac{1}{2}\|\nabla u\|^2 - (f,u)\big\} \quad \text{on} \quad M := \{v \in H_0^1(\Omega) \,|\, v \geq \psi \text{ a. e. in } \Omega\},$$

with given $f \in L^2(\Omega)$ and obstacle $\psi \in H_0^1(\Omega) \cap H^2(\Omega)$. Show that the unique solution $u \in M$ of this problem corresponds to the formal solution of the following boundary value problem:

$$\begin{aligned} -\Delta u &\geq f \ \text{ in } \Omega, \\ u &\geq \psi \ \text{ in } \Omega, \\ (-\Delta u - f)(u-\psi) &= 0 \ \text{ in } \Omega, \\ u &= 0 \ \text{ on } \partial\Omega, \\ u &= \psi, \ \ \partial_n u = \partial_n \psi \ \text{ on } \Gamma, \end{aligned}$$

where Γ is the (unknown) "interface" (free boundary) between the subsets of Ω where the constraint $u \geq \psi$ is active or inactive, respectively.

Exercise 2.7: Consider the special Signorini problem

$$\min J(u) \quad \text{on} \quad M := \{v \in V := H^1(\Omega) \,|\, v \geq 0 \ \text{ on } \partial\Omega\},$$

where

$$J(v) := \tfrac{1}{2}a(v,v) - (f,v), \quad a(v,\varphi) = (\nabla v, \nabla\varphi) + (v,\varphi).$$

on a convex polygonal domain $\Omega \subset \mathbb{R}^2$. This problem is approximated by the finite element method with piecewise linear elements on a quasi-uniform family of meshes $\{\mathbb{T}_h\}_{h>0}$:

$$\min J(u_h) \quad \text{on} \quad M_h := \{v_h \in V_h \,|\, v_h \geq 0 \ \text{ on } \partial\Omega\}.$$

Both problems, the continuous as well as the discrete ones, possess unique solutions, which are characterized by variational inequalities.

a) State the variational inequalities corresponding to this problem.

b) Derive a variant of the abstract V-error estimate developed in class, which is suited for the present special situation.

c) On the basis of this abstract error estimate prove the estimate

$$\|u - h_h\|_{H^1} \leq c(f,u)h^{3/4}.$$

Exercise 2.8: The error estimate $\|u - u_h\|_{H^1} \leq c(u)h$ for the finite element approximation of the minimal surface problem has been proven in class for piecewise linear elements on quasi-uniform families of triangulations.

a) Convince yourself that the proof remains valid also for the case of mesh families not satisfying the uniform-size condition, which allows for local mesh refinement.

b) Discuss whether the proof carries over to the case of piecewise bilinear elements on quadrilateral meshes.

Exercise 2.9: Show that the relation

$$Au(\varphi) := \int_\Omega \|\nabla u\|^{p-2}\nabla u \cdot \nabla\varphi \, dx, \quad \varphi \in V := H^{1,p}(\Omega),$$

for some $p \in (1,\infty)$, defines an operator $A : V \to V^*$. For this use the general Hölder inequality for (scalar) functions $v \in L^p(\Omega)$, $w \in L^q(\Omega)$, $1/p + 1/q = 1$:

$$\Big|\int_\Omega vw \, dx\Big| \leq \Big(\int_\Omega |v|^p \, dx\Big)^{1/p}\Big(\int_\Omega |w|^q \, dx\Big)^{1/q}, \quad v \in L^p(\Omega),\ w \in L^q(\Omega).$$

Exercise 2.10: Consider the special one-dimensional p-Laplace problem on the domain $\Omega := (-1,1)$ where

$$J(v) := \frac{1}{p}\int_\Omega |v'|^p \, dx - \int_\Omega v \, dx.$$

Show that the unique minimizer $u \in H_0^{1,p}(\Omega)$ of the functional $J(\cdot)$ is given by

$$u(x) = (1 - 1/p)\big(1 - |x|^{p/(p-1)}\big),$$

and that

$$u \in H^{2,p}(\Omega), \quad 1 < p < \frac{3+\sqrt{5}}{2}, \qquad u \notin H^{2,p}(\Omega), \quad \frac{3+\sqrt{5}}{2} \le p < \infty.$$

This demonstrates the possible limits in the regularity of solutions to strongly nonlinear problems.

Exercise 2.11: Consider the approximation of the (linear) boundary value problem

$$-\Delta u = f \ \text{ in } \Omega, \quad u = 0 \ \text{ on } \partial\Omega,$$

by "linear" finite elements. For this approximation we have the usual H^1-norm error estimate

$$\|u - u_h\|_{H^1} \le ch\|u\|_{H^2},$$

provided that $u \in H_0^1(\Omega) \cap H^2(\Omega)$. Show in case that only the minimal regularity $u \in H_0^1(\Omega)$ is known that still qualitative convergence holds

$$\|u - u_h\|_{H^1} \to 0 \quad (h \to 0).$$

However, this convergence is not uniform with respect to $\|u\|_{H^1}$. (Hint: One may use the fact that the space $C_0^\infty(\Omega)$ is by definition dense in $H_0^1(\Omega)$.)

3 General Quasilinear Elliptic Problems

In this chapter, we discuss the finite element approximation of general so-called "quasi-linear" elliptic boundary value problems in "divergence form":

$$-\sum_{i=1}^{d} \partial_i F_i(\cdot, u, \nabla u) + F_0(\cdot, u, \nabla u) = 0 \quad \text{in } \Omega, \quad u = g \text{ on } \partial\Omega, \tag{3.0.1}$$

in $\mathbb{R}^d$, and to systems of such equations in which case u is a vector function. The material of this chapter and further details can be found in the articles Frehse & Rannacher [37, 38], Dobrowolski & Rannacher [36], Rannacher [41, 43, 44, 48], and Rannacher & Scott [49].

The above setting includes the following special cases:

1. Minimal surface problem, with

$$F_i(\cdot, u, \nabla u) := (1 + |\nabla u|^2)^{-2/2} \partial_i, \quad F_0(\cdot, u, \nabla u) = 0,$$

2. Modified p-Laplace problem for $1 < p < \infty$, with

$$F_i(\cdot, u, \nabla u) := (\gamma + |\nabla u|^2)^{(p-2)/2} \partial_i, \quad F_0(\cdot, u, \nabla u) = -f,$$

   where $\gamma > 0$, the limit case $\gamma = 0$ being excluded,

3. Nonlinear elasticity problem (geometrically and statically), with

$$F_i(\cdot, u, \nabla u) := \sum_{j,k=1}^{d} (\delta_{jk} + \partial_k u_j) s_{ik}, \quad F_0(\cdot, u, \nabla u) = -f,$$

   where the stress components s_{ik} are given in terms of the strain components $\varepsilon_{ik} = \frac{1}{2}(\partial_k u_i + \partial_i u_k) + \frac{1}{2}\partial_i u_k \partial_k u_i$ via a strain energy functional $\Phi[\varepsilon(u)]$ (so-called "hyper-elastic" material):

$$s_{ik}(u) = \frac{\partial \Phi}{\partial \varepsilon_{ik}}[\varepsilon(u)].$$

4. Nonlinear diffusion problem, with

$$F_i(\cdot, u, \nabla u) := a(u) \partial_i, \quad F_0(\cdot, u, \nabla u) = -f,$$

5. Nonlinear diffusion-transport problem ("vector Burgers equation"), with

$$F_i(\cdot, u, \nabla u) := \nu \partial_i, \quad F_0(\cdot, u, \nabla u) = u \cdot \nabla u - f,$$

6. Nonlinear Diffusion-reaction problem, with

$$F_i(\cdot, u, \nabla u) := D \partial_i, \quad F_0(\cdot, u, \nabla u) = -f(u).$$

We emphasize that the considered problem does not need to originate from a minimization problem. In contrast to the treatment of the p-Laplace problem in Section 2.4, which was for-

mulated in an appropriate Sobolev (Banach) space $H^{1,p}(\Omega)$, here we prefer to use the standard Sobolev (Hilbert) space $V := H^1_0(\Omega)$ but intersected with $W^{1,\infty}(\Omega)$. We will use again the notation $\|\cdot\|_p := \|\cdot\|_{L^p}$ and $\|\cdot\|_{m,p} := \|\cdot\|_{H^{m,p}}$ for the norms of $L^p(\Omega)$, $1 \le p \le \infty,$, $H^{m,p}(\Omega)$, $m \in \mathbb{N}$, $1 \le p < \infty,$, and $W^{m,\infty}(\Omega)$, respectively. The scalar product and norm of $L^2(\Omega)$ are written without subscript as $(\cdot,\cdot)$ and $\|\cdot\|$. This notation is also used with the obvious interpretation for vector and tensor functions, e. g., $(\nabla v, \nabla w), \|\nabla v\|_1, \|\nabla^2 v\|_\infty, etc.$.

3.1 Quasi-linear problems

For technical simplicity, we only consider the special situation of a (convex) polygonal domain $\Omega \subset \mathbb{R}^2$ and homogeneous Dirichlet data $g = 0$. The case of nonhomogeneous Dirichlet data, which would be required in treating the minimal surface problem can be covered by the standard technical modifications and all results presented below remain balid for this case. Since the analysis for problem (3.0.1) in its full generality is rather technical and lengthy, for clarity of presentation, we prefer to restrict the following analysis to the prototypical special case of scalar problems in $\mathbb{R}^2$ of the form

$$-\nabla \cdot F(\cdot, \nabla u) = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega. \tag{3.1.2}$$

For the treatment of more general situations, we refer to the literature [38], [36], and [43, 44], from which much of the contents of this chapter is taken.

The function $F(x,\eta)$ is assumed to be sufficiently regular (i. e. differentiable) with respect to all its arguments $x \in \overline{\Omega}$ and $\eta \in \mathbb{R}^2$. In particular, its Jacobian matrix $F'(\cdot,\eta) = \nabla_\eta F(\cdot,\eta)$ is positive definite and Lipschitz continuous for bounded arguments, uniformly for $x \in \Omega$,

$$(F'(\cdot,\eta)\xi,\xi) \ge \alpha(\eta)|\xi|^2, \quad \xi \in \mathbb{R}^2, \qquad |F'(\cdot,\eta) - F'(\cdot,\eta')| \le \gamma(\eta,\eta')|\eta - \eta'|, \quad \eta,\eta' \in \mathbb{R}^2,$$

with a constant $\alpha(\eta) > 0$.

The variational formulation of problem (3.1.2) reads as follows:

(P) Find a function $u \in V \cap W^{1,\infty}(\Omega)$, such that

$$a(u;\varphi) = (f,\varphi) \quad \forall \varphi \in V, \tag{3.1.3}$$

where we use the semi-linear form (nonlinear in its first and linear in its second argument)

$$a(u;\varphi) := (F(\cdot,\nabla u), \nabla\varphi),$$

which is well-defined on $(V \cap W^{1,\infty}(\Omega)) \times V$.

We are not interested here in the question of existence of solutions of problem (3.1.3) but in their numerical approximation by the finite element Galerkin method. Therefore, we simply presuppose the existance of a solution $u \in V \cap W^{1,\infty}(\Omega)$, which is unique in a certain $W^{1,\infty}$-ball

$$B_R(u) = \{v \in V \cap W^{1,\infty}(\Omega),\ |||u - v|||_\infty \le R\}, \qquad |||v|||_\infty := \operatorname{ess\,sup}_\Omega |\nabla v|.$$

The local uniqueness of the solution u is related to the assumption that the "tangent form"

$$a'(u;v,w) := (F'(\cdot,\nabla u)\nabla v, \nabla w), \quad v,w \in V,$$

corresponding to the semilinear form $a(\cdot;\cdot)$ is uniformly V-elliptic for arguments $v \in B_R(u)$:

$$a'(v;\varphi,\varphi) \geq \alpha(R)\|\nabla\varphi\|^2, \quad \varphi \in V. \tag{3.1.4}$$

Further, by our assumptions on $F(\cdot,\cdot)$ the tangent form $a'(\cdot;\cdot,\cdot)$ has the following Lipschitz continuity property, uniformly on $B_R(u)$,

$$|a'(u;\varphi,\psi) - a'(v;\varphi,\psi)| \leq c(R)|||u-v|||_\infty|||\varphi||_\infty|||\nabla\psi\|_1, \tag{3.1.5}$$

for $v \in V \cap B_R(u), \varphi \in V \cap W^{1,\infty}(\Omega), \psi \in V$. The data of the problem, particularly the domain Ω, should be sufficiently regular, such that for some $q > 2$ the linear operator $A'(u): V \cap H^{2,q}(\Omega) \to L^q(\Omega)$ generated by the bilinear form $a'(u;\cdot,\cdot)$ on V is onto and satisfies the a priori estimate

$$\|v\|_{2,q} \leq c\|A'(u)v\|_q. \tag{3.1.6}$$

The link between the bilinear form $a'(u;\cdot,\cdot)$ and the operator $A'(u)$ is given by the relation

$$(A'(u)v,\varphi) := a'(u;v,\varphi), \quad v,\varphi \in V.$$

It is known that this condition is satisfied particularly on convex polygonal domains in $\mathbb{R}^2$.

Remark 3.1: We have avoided making any assumption which would guarantee the global monotonicity of problem (3.1.3). Only local regularity and continuity properties are required in a $W^{1,\infty}$-neighborhood of a sufficiently regular solution. Clearly our assumptions exclude problems with solutions having unbounded gradients, e. g., in the neighborhood of reentrant corners or at points where the ellipticity degenerates.

3.1.1 Finite element discretization

For the discretization of problem (3.1.3), we consider only the simplest finite element Galerkin method. Let again $\{\mathbb{T}_h\}_{h>0}$ be a quasi-uniform family of triangulations covering the polygonal domain Ω and $V_h \subset V$ the usual subspaces of piecewise linear functions:

$$V_h = \{v_h \in V,\, v_h|_T \in P_1(T),\; T \in \mathbb{T}_h\} \subset W^{1,\infty}(\Omega).$$

Then, the discrete problems read as follows:

(P_h) Find $u_h \in V_h$, such that

$$a(u_h;\varphi_h) = (f,\varphi_h) \quad \forall\varphi_h \in V_h. \tag{3.1.7}$$

The existence of solutions to the discrete problems (P_h) will be obtained from the assumed existence of a solution to the continuous problem (P).

Remark 3.2: For the following, we have assumed the mesh family $\{\mathbb{T}_h\}_{h>0}$ to be "quasi-uniform". i. e., particularly to satisfy the "uniform size" and "uniform shape" condition. This assumption is mainly for technical convenience. Most of the presented results remain essentially valid if only a "weak size-uniformity" condition (for the precise definition see below) is satisfied, which allows for local mesh refinement. This is important if systematic mesh adaptivity is used

in solving such nonlinear problems. We will address the question of required mesh regularity explicitly below.

Theorem 3.1: *Let the above conditions be satisfied, particularly let problem (P) possess a solution* $u \in V \cap H^{2,p}(\Omega)$ *for some* $p > 2$, *which is unique in some* $W^{1,\infty}$*-ball* $B_R(u)$. *Then, for sufficiently small* $h, h \le h_0$, *the discrete problems* (P_h) *also possess locally unique solutions* $u_h \in V_h \cap B_R(u)$ *and there holds the basic* $W^{1,\infty}$*-error estimate*

$$|||u - u_h|||_\infty \le c(u)L(h)h^{1-2/p}, \quad 2 < p \le \infty, \tag{3.1.8}$$

where $L(h) := \max\{\log(1/h), 1\}$. *Further, under the same assumptions there hold the improved* L^∞*- and* L^2*-error estimates*

$$\|u - u_h\|_\infty \le c(u)h^{2-2/p}L(h), \quad 2 < p \le \infty, \tag{3.1.9}$$

$$\|u - u_h\|_2 \le c(u)h^2. \tag{3.1.10}$$

Proof: The proof employs a so-called "homotopy argument". In this argument the given nonlinear problems are embedded into a parameter-dependent family of problems containing, for one special parameter value, a linear problem, for which the asserted error estimate is known. By a continuation argument that estimate is then carried over to the given nonlinear problems. We shall give the details only for the basic $W^{1,\infty}$-error estimate (3.1.8) leave the modification of the argument for covering also the other estimates as an exercise. Below, the symbol c will be used as a "generic" constant, which may vary with the context, but is always independent of all critical parameters involved.

(i) We begin with some technical preliminaries. We introduce the linear "Ritz projection" $R_h : V \to V_h$ defined by

$$a'(u; R_h v, \varphi_h) = a'(u; v, \varphi_h) \quad \forall \varphi_h \in V_h,\ v \in V. \tag{3.1.11}$$

This construction is well-defined, since by assumption the bilinear form $a'(u; \cdot, \cdot)$ is continuous and V-elliptic (Lax-Milgram theorem). Under our assumptions, this Ritz projection is known to be almost $W^{1,\infty}$-stable, i. e., there holds the estimate

$$|||R_h v|||_\infty \le cL(h)|||v|||_\infty \quad v \in V \cap W^{1,\infty}(\Omega). \tag{3.1.12}$$

This stability estimate will be proven in the next section. In view of the usual approximation estimate for the nodal interpolation $I_h : V \cap C(\overline{\Omega}) \to V_h$,

$$|||u - I_h u|||_\infty \le ch^{1-2/p}\|u\|_{2,p}, \quad 2 < p \le \infty, \tag{3.1.13}$$

the stability estimate (3.1.12) for R_h implies the error estimate

$$\begin{aligned} |||u - R_h u|||_\infty &\le |||u - I_h u|||_\infty + |||R_h(I_h u - u)|||_\infty \\ &\le cL(h)|||u - I_h u|||_\infty \le c_0 L(h)h^{1-2/p}\|u\|_{2,p}, \quad 2 < p \le \infty. \end{aligned} \tag{3.1.14}$$

Further, the following error estimates are known for the Ritz projection:

$$\|u - R_h u\|_\infty \le ch^{2-2/p}L(h)\|u\|_{2,p}, \quad 2 < p \le \infty, \tag{3.1.15}$$

$$\|u - R_h u\|_2 \le ch^2\|u\|_{2,2}. \tag{3.1.16}$$

These can be deduced similarly as in (3.1.14) from the stability estimates

$$\|R_h v\|_\infty \le c\|v\|_\infty + chL(h)\|\nabla v\|_\infty, \quad v \in V \cap W^{1,\infty}(\Omega), \tag{3.1.17}$$

$$\|R_h v\| \le c\|v\| + ch\|\nabla v\|, \quad v \in V \cap H^2(\Omega), \tag{3.1.18}$$

which are among the estimates proven in the next section. The Ritz projection $R_h : V \to V_h$ can be extended to an operator (using the same notation) $\hat{R}_h : V_h^* \to V_h$ by defining

$$a'(u; \hat{R}_h v_h^*, \varphi_h) = v_h^*(\varphi_h) \quad \forall \varphi_h \in V_h, \;\; v_h^* \in V_h^*.$$

Notice that in the usual sense $V_h^* \subset V^* = V$. For this extended Ritz projection there holds the stability estimate

$$|||\hat{R}_h v_h^*|||_\infty \le cL(h)\|v_h^*\|_{\infty;h}, \tag{3.1.19}$$

where the dual norm $\|\cdot\|_{\infty;h}$ is defined by

$$\|v_h^*\|_{\infty;h} := \sup\big\{v_h^*(v_h) \,|\, v_h \in V_h, \; \|\nabla v_h\|_1 = 1\big\}.$$

This stability estimate will also be proven in the next section. The logarithmic factor $L(h)$ in the stability estimate (3.1.19) cannot be avoided, which can be shown by counter examples (exercise). In this analysis, we do not make use of the uniform size property of the mesh family $\{\mathbb{T}_h\}_{h>0}$, only the uniform shape property is required.. Therefore all results in this chapter hold on such mesh families, which particularly allows for local mesh refinement. It can be shown that for quasi-uniform meshes the extra logarithmic factor $L(h)$ in the $W^{1,\infty}$-stability estimate (3.1.12) and also in the $W^{1,\infty}$-error estimates (3.1.14) and consequently in the corresponding estimate (3.1.8) in the main Theorem 3.1 does not occur (see [49]).

(ii) In order to carry the estimate (3.1.14) over to the nonlinear problem (3.1.3), we use a homotopy argument. For a homotopy parameter $t \in [0,1]$, we introduce the semilinear forms

$$a_t(v; w) := ta(v; w) + (1-t)a'(u; v-u, w)$$

and the corresponding auxiliary problems

(P^t) Find $u^t \in V \cap W^{1,\infty}(\Omega)$ such that

$$a_t(u^t; v) = t(f, v) \quad \forall v \in V, \tag{3.1.20}$$

and their discrete analogues

(P_h^t) Find $u_h^t \in V_h$ such that

$$a_t(u_h^t; v_h) = t(f, v_h) \quad \forall v_h \in V_h. \tag{3.1.21}$$

Clearly, $u_h^1 = u_h$ and $u_h^0 = R_h u$. Further, for all $t \in [0,1]$, the function $u^t := u$ is a solution of problem (3.1.20) and there holds $a_0'(u; \cdot, \cdot) = a'(u; \cdot, \cdot)$. We define the set

$$\Theta_h := \big\{t \in [0,1] \,|\, \text{Problem } (P_h^t) \text{ has a locally unique solution } u_h^t \in V_h \cap B_R(u), \\ \text{for which there holds } |||u - u_h^t|||_\infty < 2c_0 L(h) h^{1-2/p}.\big\},$$

where c_0 is the constant in the "linear" error estimate (3.1.14). We want to show that, for sufficiently small h, $h \le h_0$, the set Θ_h is nonempty, open and closed with respect to $[0,1]$ and therefore coincides with $[0,1]$. Then, $1 \in \Theta_h$ implies the asserted error estimate.

(iii) By definition, we have $0 \in \Theta_h$ so that $\Theta_h \neq \emptyset$.

(iv) Next, we establish the closedness of Θ_h for any fixed h, which is the harder part of the proof. Consider a convergent sequence $(t_k)_{k\in\mathbb{N}} \subset \Theta_h$, with $t := \lim_{k\to\infty} t_k$, and the corresponding sequence $(u_h^{t_k})_{k\in\mathbb{N}}$ of discrete solutions $u_h^{t_k} \in V_h \cap B_R(u)$. Since the sequence $(u_h^{t_k})_{k\in\mathbb{N}}$ is bounded in the finite dimensional space V_h, there exists a convergent subsequence with limit $u_h^t \in V_h \cap B_R(u)$. By continuity u_h^t is a solution of the corresponding variational problem (P_h^t), for which the error estimate holds

$$|||u - u_h^t|||_\infty \le 2c_0 L(h) h^{1-2/p}.$$

On V_h the tangent form of $a_t(\cdot;\cdot)$ is given by

$$a_t'(v_h; w_h, \varphi_h) := ta'(v_h; w_h, \varphi_h) + (1-t)a'(u; w_h, \varphi_h), \quad v_h, w_h, \varphi_h \in V_h,$$

which clearly is V-elliptic for $v_h \in B_R(u)$. This implies that the discrete solution u_h^t is locally unique. Since for any other solution $\tilde{u}_h^t \in V_h \cap B_R(u)$ the relation

$$0 = a_t(u_h^t; v_h) - a_t(\tilde{u}_h^t; v_h) = \int_0^1 a_t'(\tilde{u}_h^t + t(u_h^t - \tilde{u}_h^t); u_h^t - \tilde{u}_h^t, v_h)\, dt,$$

observing that $\tilde{u}_h^t + t(u_h^t - \tilde{u}_h^t) \in B_R(u)$ and setting $v_h = u_h^t - \tilde{u}_h^t$ implies $\tilde{u}_h^t = u_h^t$. Further, observing that u_h^t solves

$$a_t(u_h^t; v_h) = ta(u_h^t; v_h) + (1-t)a'(u; u_h^t - u, v_h) = t(f, v_h) \quad \forall v_h \in V_h,$$

we can write

$$\begin{aligned} a'(u; u - u_h^t, v_h) + ta(u; v_h) - ta(u_h^t; v_h) &= a'(u; u - u_h^t, v_h) + t(f; v_h) - ta(u_h^t; v_h) \\ &= a'(u; u - u_h^t, v_h) + (1-t)a'(u; u_h^t - u, v_h) \\ &= ta'(u; u - u_h^t, v_h) \end{aligned}$$

and therefore,

$$\begin{aligned} a'(u; R_h u - u_h^t, v_h) &= a'(u; u - u_h^t, v_h) + t\big\{a(u; v_h) - a(u_h^t; v_h)\big\} - t\big\{a(u; v_h) - a(u_h^t; v_h)\big\} \\ &= ta'(u; u - u_h^t, v_h) - t\int_0^1 a'(u_h^t + s(u - u_h^t); u - u_h^t, v_h)\, ds \\ &= ta'(u; u - u_h^t, v_h) - t\int_0^1 a'(u_h^t + s(u - u_h^t); u - u_h^t, v_h)\, ds \\ &= t\int_0^1 \big\{a'(u; u - u_h^t, v_h) - a'(u_h^t + s(u - u_h^t); u - u_h^t, v_h)\big\}\, ds. \end{aligned}$$

The right-hand side of this equation can be viewed as a functional $v_h^* \in V_h^*$ acting on $v_h \in V_h$, i. e., $R_h u - u_h^t = \hat{R}_h v_h^*$. By the assumed Lipschitz continuity of $a'(\cdot; u - u_h^t, v_h)$ for bounded

arguments, the dual norm of this functional is estimate by

$$\|v_h^*\|_{\infty;h} \le ct|||u - u_h^t|||_\infty^2.$$

Consequently, by the stability estimate (3.1.19),

$$|||R_h u - u_h^t|||_\infty \le cL(h)\|v_h^*\|_{\infty;h} \le ctL(h)|||u - u_h^t|||_\infty^2.$$

Hence, by the error estimate (3.1.14), we obtain

$$\begin{aligned} |||u - u_h^t|||_\infty &\le |||u - R_h u|||_\infty + |||R_h u - u_h^t|||_\infty \\ &\le c_0 L(h) h^{1-2/p} + c_1 t L(h)|||u - u_h^t|||_\infty^2, \end{aligned} \tag{3.1.22}$$

with constants c_0 and c_1 independent of h and θ. This implies that for $h \le h_0$ sufficiently small

$$|||u - u_h^t|||_\infty < 2c_0 L(h) h^{1-2/p},$$

i. e., $t \in \Theta_h$.

(iv) For proving the openness of Θ_h, we employ the implicit function theorem. For any $t \in \Theta_h$ we define a function $H(t,\varphi) : [0,1] \times (V_h \cap B_R(u)) \to V_h^*$ (dual space of V_h) by setting

$$H(t,\varphi)(\cdot) := ta(\varphi;\cdot) + (1-t)a'(u;\varphi - u,\cdot) - t(f,\cdot).$$

Obviously, $H(t,\varphi)$ is continuous and continuously differentiable with respect to φ on $[0,1] \times (V_h \cap B_R(u))$ with derivative

$$H'(t,\varphi)(\psi,\cdot) = ta'(\varphi;\psi,\cdot) + (1-t)a'(u;\psi,\cdot), \quad \psi \in V_h.$$

Then, for any $\tau \in \Theta_h$, we have by definition,

$$u_h^\tau \in V_h \cap B_R(u), \quad H(\tau, u_h^\tau)(\cdot) = 0.$$

and

$$H'(\tau, u_h^\tau)(\psi,\cdot) = \tau a'(u_h^\tau;\psi,\cdot) + (1-\tau)a'u;\psi,\cdot), \quad \psi \in V_h.$$

By the V-ellipticity the relation

$$H'(\tau, u_h^\tau)(\psi,\chi) = 0, \quad \chi \in V_h,$$

implies that necessarily $\psi = 0$. Hence the inverse $H'(\tau,\psi)^{-1}$ exists since V_h is finite dimensional. Then, the implicit function theorem, applied to $H(t,\varphi)$, yields the existence of a neighborhood $N(\tau) \subset [0,1]$ such that there are functions $\varphi^{(t)} \in V_h \cap B_R(u)$, which satisfy

$$H(t,\varphi^{(t)})(\cdot) = 0, \quad t \in N(\tau),$$

and are continuous with respect to $t \in N(\tau)$. This obviously shows that $t \in \Theta_h$ for $|t-\tau|$ sufficiently small. Hence, Θ_h is open and the proof is complete. Q.E.D.

3.1.2 Auxiliary L^∞-stability estimates for the linearized problems

In the following, we give the proofs of the stability estimates (3.1.12) and (3.1.19) for the Ritz projections R_h and $\hat{R}_h$ introduced above in the proof of Theorem 3.1. Since this has expository character, we restrict the presentation to the model problem of the Poisson equation:

$$-\Delta u = f \quad \text{in } \Omega, \quad u = 0 \quad \text{on } \partial\Omega, \tag{3.1.23}$$

on a convex polygonal domain $\Omega \subset \mathbb{R}^2$. The argument for the more general elliptic operator $A'(u)$ is similar but technically more involved. Details can be found in the literature [37], [49], and [**?**]. Further, we only consider low-order piecewise linear finite elements on a mesh familily $\{\mathbb{T}_h\}_{h>0}$ assumed to be strongly shape-regular but not in all cases necessarily strongly size-regular. We shall try to avoid as much as possible the use of the "uniform-size property" in order to allow for local mesh refinement. Accordingly, we set $h := \max_{T\in\mathbb{T}_h} h_T$ (h_T the radius of the minimal circumscribed circle of T) and $\rho := \min_{T\in\mathbb{T}_h} \rho_T$ (ρ_T radius of the maximal inscribed circle of T).

Definition 3.1: *The family of triangulations $\{\mathbb{T}_h\}_{h>0}$ is said to be "weakly size-regular", if there exist some $\alpha \geq 1$, such that uniformly for all meshes:*

$$h_{\min} \geq c h_{\max}^{\alpha}, \quad \mathbb{T}_h \in \{\mathbb{T}_h\}_{h>0}. \tag{3.1.24}$$

Remark 3.3: In practice the case $1 \leq \alpha \leq 2$ is most relevant in mesh adaptation. Faster mesh refinement is rarely necessary in standard situations. The main feature of "weak size regularity" is that in this case the logarithmic factor $\log(1/h_{\min})$ frequently occurring in error estimates behaves like $\alpha \log(1/h_{\max})$. In the extreme case of refinement towards a point by m-times local bisection, leading to $h_{\min} \sim 2^{-m} h_{\max}$, there would hold $\log(2^m/h_{\max}) \sim m\log(2)+\log(1/h_{\max})$, meaning a significant loss in accuracy.

First, we consider the standard Ritz projection $R_h : V \to V_h$, which in the present situation is defined by

$$(\nabla R_h u, \nabla \varphi_h) = (\nabla u, \nabla \varphi_h) \quad \forall \varphi_h \in V_h. \tag{3.1.25}$$

We note the following L^2-stability estimate

$$\|R_h u\| \leq \|u\| + ch\|\nabla u\|, \quad u \in V, \tag{3.1.26}$$

which is easily obtained using a standard duality argument (exercise). The following lemma provides the corresponding L^∞-stability estimate.

Theorem 3.2 (L^∞ stability): *Suppose that the domain $\Omega \subset \mathbb{R}^2$ is convex polygonal and that the regular mesh family $\{T_h\}_{h>0}$ is strongly shape- but possibly only weakly size-regular with an exponents $\alpha \geq 1$. Then, there holds*

$$\|R_h u\|_\infty \leq c\|u\|_\infty + c\alpha h L(h)\|\nabla u\|_\infty, \tag{3.1.27}$$

for functions $u \in H_0^1(\Omega) \cap W^{1,\infty}(\Omega)$, with a constant c independent of u, h, and α.

Proof: We set $u_h := R_h u$. Let T_* be an arbitrary cell of a mesh $\mathbb{T}_h$ with diameter $h_* = h_{T_*}$ and maximal inscribed circle $B_* \subset T_*$ with center point x_* and radius $\rho_* = \rho_{T_*}$. By a standard scaling argument and the equivalence of norms on the finite dimensional space P_1 there holds

$$\|u_h\|_{\infty;B_*} \leq c|B_*|^{-1} \int_{B_*} |u_h|\, dx. \tag{3.1.28}$$

The integral on the right of (3.1.28) can be rewritten as

$$|B_*|^{-1} \int_{B_*} |u_h|\, dx = (u_h, \delta),$$

where the regularized Dirac function $\delta := \delta^h \in L^2(\Omega)$ is defined by

$$\delta := |B_*|^{-1}\mathrm{sign}(u_h) \text{ in } B_*, \quad \delta := 0 \text{ else.}$$

To the function δ, we associate a regularized Green's function $g = g^h \in V$ as the solution of the dual problem

$$(\nabla\varphi, \nabla g) = (\varphi, \delta) \quad \forall \varphi_{\in} V, \tag{3.1.29}$$

and its Ritz projection $g_h := R_h g \in V_h$. Then, by definition of u_h and by Galerkin orthogonality, there holds

$$\begin{aligned}(u_h, \delta) &= (\nabla u_h, \nabla g) = (\nabla u_h, \nabla g_h) = (\nabla u, \nabla g_h)\\ &= (\nabla u, \nabla(g_h - g)) + (\nabla u, \nabla g) = (\nabla u, \nabla(g_h - g)) + (u, \delta),\end{aligned}$$

and consequently,

$$\|u_h\|_{\infty;B_*} \leq \|\nabla u\|_\infty \|\nabla(g - g_h)\|_1 + \|u\|_{\infty;B_*}. \tag{3.1.30}$$

Hence the proof is completed by proving the following Lemma 3.1 and observing that (exercise),

$$\|u_h\|_{\infty;T_*} \leq c\|u_h\|_{\infty;B_*},$$

and, by the assumed strong shape- and weak size-regularity,

$$L(\rho_*) \leq L(ch_*) \leq L(ch_{\max}^\alpha) \leq c\alpha L(h_{\max}) = c\alpha L(h).$$

Q.E.D.

Lemma 3.1: *For the regularized Green's function there holds the $H^{1,1}$-error estimate*

$$\|\nabla(g - g_h)\|_1 \leq chL(\rho_*). \tag{3.1.31}$$

The proof of this lemma requires some preparation. We define the weight function

$$\sigma(x) := (|x - x_*|^2 + \kappa^2\rho_*^2)^{1/2},$$

with some parameter $\kappa \geq 1$, which will be appropropriately fixed below. By elementary calculation, we see that

$$\sigma \geq \kappa\rho_*, \quad |\nabla\sigma| \leq 1, \quad \|\sigma^{-1}\| \leq c\, L(\rho_*)^{1/2}, \tag{3.1.32}$$

and for each mesh cell $T \in \mathbb{T}_h$,

$$\max_T \sigma \leq \min_T \sigma + h_T \max_T |\nabla\sigma| \leq \min_T \sigma + h_T. \tag{3.1.33}$$

The assertion of the following auxiliary lemma is crucial for the proof of Lemma 3.1.

Lemma 3.2: *For fixed $\kappa \geq 1$ there hold the following a priori estimates:*

$$\|g\|_\infty \leq cL(\rho_*), \tag{3.1.34}$$

$$\|\nabla g\| + \|\sigma\nabla^2 g\| \leq cL(\rho_*)^{1/2}, \tag{3.1.35}$$

$$\|\nabla^2 g\| \leq c\rho_*^{-1}. \tag{3.1.36}$$

Proof: (i) The true Green's function G_x on Ω corresponding to an arbitrary point $x \in \Omega$ admits the well-known estimate

$$|G_x(y)| \leq c\{|\ln(|y-x|)| + 1\},$$

which may be derived by using the maximum principle. Then, from the estimate (exercise)

$$|g(x)| = |(\nabla g, \nabla G_x)| = |(\delta, G_x)| \leq |B_*|^{-1} \int_{B_*} |G_x|\, dy \leq cL(\rho_*),$$

we obtain (3.1.34).
(ii) Observing that

$$\|\nabla g\|^2 = (\delta, g) \leq \|g\|_\infty \leq cL(\rho_*),$$

we obtain the first part of (3.1.35). Further, by the usual H^2 a priori estimate, we obtain (3.1.36),

$$\|\nabla^2 g\| \leq c\|\Delta g\| = c\|\delta\| \leq c\rho_*^{-1}.$$

(iii) Next, we set $\xi := x - x_*$ and find

$$|\xi_i \nabla^2 g| \leq |\nabla^2(\xi_i g)| + |\nabla g|,$$

and consequently,

$$\begin{aligned}\|\sigma\nabla^2 g\|^2 &= \sum_{i=1}^2 \|\xi_i\nabla^2 g\|^2 + \kappa^2\rho_*^2\|\nabla^2 g\|^2 \\ &\leq c\sum_{i=1}^2 \left\{\|\nabla^2(\xi_i g)\|^2 + \|\nabla g\|^2\right\} + \kappa^2\rho_*^2\|\nabla^2 g\|^2.\end{aligned}$$

By the usual H^2 a priori estimate,

$$\begin{aligned}\|\nabla^2(\xi_i g)\| &\le \|\Delta(\xi_i g)\| \le \|\xi_i \Delta g\| + \|\nabla g\| \\ &= \|\xi_i \delta\| + \|\nabla g\| \le c + cL(\rho_*)^{1/2}.\end{aligned}$$

Combining the foregoing estimates, we obtain the second part of (3.1.35),

$$\|\sigma \nabla^2 g\|_2 \le cL(\rho_*)^{1/2},$$

which completes the proof. Q.E.D.

Proof of Lemma 3.1: Let $\eta := g - g_h$. From the L^2-stability estimate (3.1.26), which also holds on the general meshes considered, or alternatively by directly using the V-stability of the Ritz projection and applying a duality argument, we obtain the well-known L^2-error estimate

$$\|v - R_h v\| + h\|\nabla(v - R_h v)\| \le ch\|\nabla v\|, \quad v \in V. \tag{3.1.37}$$

Further, we recall the usual cellwise estimate for the standard nodal interpolation:

$$\|v - I_h v\|_T + h_T\|\nabla(v - I_h v)\|_T \le ch_T^2\|\nabla^2 v\|_T, \quad T \in \mathbb{T}_h,\ v \in H^2(T). \tag{3.1.38}$$

(i) We fix $\kappa = 1$. Combining the L^2-error estimate (3.1.37) for $v := g$ with the a priori estimate (3.1.35), we have

$$\|\eta\| + h\|\nabla\eta\| \le ch\|\nabla g\| \le chL(\rho_*)^{1/2}. \tag{3.1.39}$$

Further there holds

$$\|\nabla\eta\|_1 \le \|\sigma^{-1}\|\,\|\sigma\nabla\eta\| \le cL(\rho_*)^{1/2}\|\sigma\nabla\eta\|. \tag{3.1.40}$$

For the term on the right, we have

$$\|\sigma\nabla\eta\|^2 = (\nabla\eta, \nabla(\sigma^2\eta) - (\nabla\eta, \eta\nabla\sigma^2) =: E_1 - E_2.$$

The terms E_1 and E_2 will be estimated separately. First, using Galerkin orthogonality, we get

$$E_1 = (\nabla\eta, \nabla(\sigma^2\eta - \psi_h))$$

with the nodal interpolation $\psi_h := I_h(\sigma^2\eta) \in V_h$. This term is estimated further using the cellwise interpolation estimate (3.1.38),

$$E_1 \le \sum_{T\in\mathbb{T}_h} \|\nabla\eta\|_T\|\nabla(\sigma^2\eta - \psi_h)\|_T \le c\sum_{T\in\mathbb{T}_h} h_T\|\nabla\eta\|_T\|\nabla^2(\sigma^2\eta)\|_T.$$

For the second factors on the right, we have

$$\|\nabla^2(\sigma^2\eta)\|_T \le c\big\{\|\eta\|_T + \|\sigma\nabla\eta)\|_T + \|\sigma^2\nabla^2 g\|_T\big\},$$

and consequently,

$$E_1 \le c\sum_{T\in\mathbb{T}_h} h_T\Big\{\|\nabla\eta\|_T\big\{\|\eta\|_T + \|\sigma\nabla\eta\|_T\big\} + \|\nabla\eta\|_T\|\sigma^2\nabla^2 g\|_T\Big\}.$$

In view of the relation $\max_T \sigma \le \min_T \sigma + h_T$, we have

$$\begin{aligned}\|\nabla\eta\|_T\|\sigma^2\nabla^2 g\|_T &\le \max_T\sigma\,\|\nabla\eta\|_T\|\sigma\nabla^2 g\|_T\\ &\le \big\{\|\sigma\nabla\eta\|_T + h_T\|\nabla\eta\|_T\big\}\|\sigma\nabla^2 g\|_T.\end{aligned}$$

Hence, by Schwarz's inequality,

$$E_1 \le ch\|\nabla\eta\|\big\{\|\eta\| + \|\sigma\nabla\eta\|\big\} + ch\big\{\|\sigma\nabla\eta\| + h\|\nabla\eta\|\big\}\|\sigma\nabla^2 g\|.$$

In view of the L^2-error estimate (3.1.39) and the a priori estimate (3.1.35), we conclude

$$E_1 \le \tfrac{1}{4}\|\sigma\nabla\eta\|^2 + ch^2 L(\rho_*).$$

For the second term E_2, we analogously obtain

$$E_2 \le c\|\sigma\nabla\eta\|\|\eta\| \le \tfrac{1}{4}\|\sigma\nabla\eta\|^2 + ch^2 L(\rho_*).$$

Finally, combining the estimates for E_1 and E_2, we obtain

$$\|\sigma\nabla\eta\|^2 \le \tfrac{1}{2}\|\sigma\nabla\eta\|^2 + ch^2 L(\rho_*),$$

which in view of (3.1.40) completes the proof of Lemma 3.1. Q:E:D.

Remark 3.4: We note that in the foregoing argument, we did not use the "strong shape regularity" of the mesh family $\{\mathbb{T}_h\}_{h\in\mathbb{R}_+}$, i. e., in 2D the L^1 error estimate (3.1.31) holds true also under the assumption of only "weak shape regularity", which is defined similarly as "weak size regularity".

Remark 3.5: The logarithmic factor in the stability estimate (3.1.27) is unavoidable in general. This can be demonstrated by analytical arguments for special situations and is also confirmed by numerical experiments.

Next, we derive $W^{1,\infty}$-stability estimates. We note that in case that the mesh family is quasi-uniform, a $W^{1,\infty}$-stability estimate could be deduced (exercise) from the estimate (3.1.27) of Theorem 3.2 by using the inverse relation

$$\|\nabla v_h\|_\infty \le ch_{\min}^{-1}\|v_h\|_\infty, \quad v_h \in V_h,$$

together with the interpolation estimate

$$\|v - I_h v\|_\infty + h_{\max}\|\nabla(v - I_h v)\|_\infty \le ch_{\max}\|\nabla v\|_\infty, \quad v \in W^{1,\infty}(\Omega).$$

However, this argument cannot be used for the extended Ritz projection $\hat{R}_h$, for which the proof of $W^{1,\infty}$-stability requires more work, even on quasi-uniform mesh families. In the following argument, we will make this regularity assumption for technical simplicity. The extension to only weakly size-uniform meshes, which would be very desirable in the context of adaptive methods, has not been fully accomplished yet. We will comment on this aspect below in the proof of the following theorem.

Theorem 3.3 ($W^{1,\infty}$-stability): *Suppose that the domain $\Omega \subset \mathbb{R}^2$ is convex polygonal and that the regular mesh family $\{T_h\}_{h>0}$ is quasi-uniform. Then, there hold the $W^{1,\infty}$-stability estimates*

$$\|\nabla R_h u\|_\infty \le cL(h)\|\nabla u\|_\infty, \tag{3.1.41}$$

for $u \in H_0^1(\Omega) \cap W^{1,\infty}(\Omega)$, and

$$\|\nabla \hat{R}_h v_h^*\|_\infty \le cL(h)\|v_h^*\|_{\infty;h}, \tag{3.1.42}$$

for $v_h^* \in V_h^*$, with constants c independent of u and h.

Proof: We continue using the notation of the proof of Theorem 3.2 and particularly that of Lemma 3.1. We set again $u_h := R_h u$ or $u_h := \hat{R}_h v_h^*$.

(i) For any fixed $h > 0$ let $T_* \in \mathbb{T}_h$ be an arbitrary cell with maximal inscribed circle $B_* := B_{\rho_*}(x_*)$, with radius ρ_* and center point x_*, and $h_* := h_{T_*}$. We will again use the weight function $\sigma(x) := \left(|x - x_*|^2 + \kappa^2\rho_*^2\right)^{1/2}$, which satisfies

$$\sigma \ge \kappa\rho_*, \quad |\nabla\sigma| \le 1, \quad \|\sigma^{-1}\| \le cL(\rho_*)^{1/2}. \tag{3.1.43}$$

Now, there exists a Dirac-like function $\delta \in C_0^\infty(B_*)$ with the properties

$$0 \le \delta \le c\rho_*^{-2}, \quad |\nabla\delta| \le c\rho_*^{-3}, \quad \int_{B_*} \delta\, dx = 1.$$

Since ∇u_h is constant on T_*, we have

$$\|\partial_i u_h\|_{\infty;T_*} = |(\partial_i u_h, \delta)|, \quad i = 1, 2.$$

To the function δ, we associate regularized ("derivative") Green's functions $g_i' \in V$, $i = 1, 2$, as the solution of the "dual problems"

$$(\nabla\varphi, \nabla g_i') = (\partial_i\varphi, \delta) \quad \forall\varphi \in V, \tag{3.1.44}$$

and their Ritz projections $R_h g_i' \in V_h$. Then, by definition of u_h and Galerkin orthogonality, there holds in the case $u_h = R_h u$,

$$\begin{aligned}(\partial_i u_h, \delta) &= (\nabla u_h, \nabla g_i') = (\nabla u_h, \nabla R_h g_i') = (\nabla u, \nabla R_h g_i') \\ &= (\nabla u, \nabla(R_h g_i' - g_i')) + (\nabla u, \nabla g_i') = (\nabla u, \nabla(R_h g_i' - g_i')) + (\partial_i u, \delta),\end{aligned}$$

and, consequently,

$$\|\partial_i u_h\|_{\infty;T_*} \le \|\nabla u\|_\infty \max_{i=1,2} \|\nabla(R_h g_i' - g_i')\|_1 + \|\nabla u\|_{\infty;T_*}. \tag{3.1.45}$$

In the case $u_h = \hat{R}_h v_h^*$, there holds

$$(\partial_i \hat{R}_h v_h^*, \delta) = (\nabla \hat{R}_h v_h^*, \nabla g_i') = (\nabla \hat{R}_h v_h^*, \nabla R_h g_i') = v_h^*(R_h g_i'),$$

and, consequently,

$$\|\partial_i \hat{R}_h v_h^*\|_{B_*;\infty} \le \|v_h^*\|_{\infty;h} \|\nabla R_h g_i'\|_1.$$

Hence using the estimate in the following Lemma 3.3 and observing again that, by the assumed properties of the considered meshes,

$$L(\rho_*) \le L(ch_*) \le L(ch_{\max}) \le cL(h), \tag{3.1.46}$$

the proof is completed. Q.E.D.

In the following, we drop the index $i \in \{1,2\}$ in the derivative Green's functions and set $g' := g'_i$ and $g'_h := R_h g'$.

Lemma 3.3: *For the regularized derivative Green's functions g' there holds the $H^{1,1}$ estimate*

$$\|\nabla(g' - g'_h)\|_1 + \|\nabla g'_h\|_1 \le cL(\rho_*). \tag{3.1.47}$$

Proof: For abbreviation we set $\eta' := g' - g'_h$. The proof is given in a sequence of steps.

(i) There holds

$$\|\nabla\eta'\|_1 \le \|\sigma^{-1}\|\|\sigma\nabla\eta'\| \le cL(\rho_*)^{1/2}\|\sigma\nabla\eta'\|. \tag{3.1.48}$$

For the term on the right, we have

$$\|\sigma\nabla\eta'\|^2 = (\nabla\eta', \nabla(\sigma^2\eta')) - (\nabla\eta', \eta'\nabla\sigma^2) =: E'_1 - E'_2.$$

The terms E'_1 and E'_2 will be estimated separately. First, using Galerkin orthogonality, we get

$$E'_1 = (\nabla\eta', \nabla(\sigma^2\eta' - \psi_h))$$

with the nodal interpolant $\psi_h := I_h(\sigma^2\eta') \in V_h$. This term is estimated further, using the usual cellwise L^2-interpolation estimate and the quasi-uniformity of the meshes,

$$\begin{aligned}
E'_1 &\le \sum_{T\in\mathbb{T}_h} \|\nabla\eta'\|_T\|\nabla(\sigma^2\eta' - \psi_h)\|_T \\
&\le c\sum_{T\in\mathbb{T}_h} h_T\|\nabla\eta'\|_T\|\nabla^2(\sigma^2\eta')\|_T \\
&\le c\kappa^{-1}\sum_{T\in\mathbb{T}_h} \|\sigma\nabla\eta'\|_T\big\{\|\eta'\|_T + \|\sigma\nabla\eta'\|_T + \|\sigma^2\nabla^2 g'\|_T\big\} \\
&\le c\kappa^{-1}\|\sigma\nabla\eta'\|\big\{\|\eta'\| + \|\sigma\nabla\eta'\| + \|\sigma^2\nabla^2 g'\|\big\}.
\end{aligned}$$

Remark 3.6: We note that the quasi-uniformity of the mesh family $\{\mathbb{T}_h\}_{h>0}$ is used for the estimate

$$h_T\|\nabla\eta'\|_T \le c\kappa^{-1}\|\sigma\nabla\eta'\|_T, \tag{3.1.49}$$

which requires that $\kappa h_T \le c\min_T\sigma$, $T \in \mathbb{T}_h$. We conjecture that this is remains true for only weakly size-regular meshes, which would then imply the assertion of the theorem also under thia weaker assumption.

For the next step, we need to estimate $\|\eta'\|$ by $\|\sigma\nabla\eta'\|$. For this, we use a duality argument. Let $z \in V$ be the solution of the auxiliary problem

$$(\nabla\varphi, \nabla z) = (\varphi, \eta')\|\eta'\|^{-1} \quad \forall\varphi \in V,$$

satisfying $z \in H^2(\Omega)$ and $\|z\|_{2,2} \le c$. Then, there holds

$$\|\eta'\| = (\nabla\eta', \nabla z) = (\nabla\eta', \nabla(z - I_h z)),$$

and further

$$\|\eta'\| \le c \sum_{T\in\mathbb{T}_h} h_T \|\nabla\eta'\|_T \|\nabla^2 z\|_T \le c\kappa^{-1}\|\sigma\nabla\eta'\|.$$

Now, using the just proven estimate $\|\eta'\| \le c\kappa^{-1}\|\sigma\nabla\eta'\|$, we conclude that

$$E_1 \le c\kappa^{-1}\|\sigma\nabla\eta'\|^2 + c\|\sigma^2\nabla^2 g'\|^2.$$

For the second term E_2, we have

$$E_2 \le c\|\sigma\nabla\eta'\|\|\eta'\| \le c\kappa^{-1}\|\sigma\nabla\eta'\|^2.$$

Combining this with the above estimate for E_1', we obtain

$$\|\sigma\nabla\eta'\|^2 \le c\kappa^{-1}\|\sigma\nabla\eta'\|^2 + c\|\sigma^2\nabla^2 g'\|.$$

Hence, in view of the estimate (3.4) in Lemma 3.4, below, chosing κ sufficiently large, we prove the main assertion of the lemma. Then, the complete assertion follows by using the estimate

$$\|\nabla g_h'\|_1 \le \|\nabla(g_h' - g')\|_1 + \|\nabla g'\|_1 \le cL(h)^{1/2}\big\{\|\sigma\nabla(g_h' - g')\| + \|\sigma\nabla g'\|\big\},$$

again together with the a priori bound (3.1.50). Q.E.D.

Lemma 3.4: *There holds the a priori estimate*

$$\|\sigma\nabla g'\| + \|\sigma^2\nabla^2 g'\| \le cL(\rho_*)^{1/2}. \tag{3.1.50}$$

Proof: Using the usual H^2-a priori estimate on Ω, we estimate as follows:

$$\begin{aligned}
\|\sigma^2\nabla^2 g'\| &\le \|\nabla^2(\sigma^2 g')\| + c\|\sigma\nabla g'\| + c\|g'\| \\
&\le c\|\Delta(\sigma^2 g')\| + c\|\sigma\nabla g'\| + c\|g'\| \\
&\le c\|\sigma^2\Delta g'\| + c\|\sigma\nabla g'\| + c\|g'\| \\
&\le c\|\sigma^2\partial_i\delta'\| + c\|\sigma\nabla g'\| + c\|g'\| \\
&\le c + c\|\sigma\nabla g'\| + c\|g'\|.
\end{aligned}$$

Further, there holds

$$\begin{aligned}\|\sigma\nabla g'\|^2 &= (\nabla(\sigma^2 g'),\nabla g') - (\nabla\sigma^2 g',\nabla g')\\ &= (\partial_i(\sigma^2 g'),\delta')_{B_*} + \tfrac{1}{2}(\Delta\sigma^2 g', g'),\\ &= -(\sigma^2 g',\partial_i\delta')_{B_*} + 2\|g'\|^2\\ &\le c\rho_*^{-1}\|g'\|_{1;B_*} + 2\|g'\|^2\\ &\le c(1+\|g'\|^2).\end{aligned}$$

Let $z \in V \cap H^2(\Omega)$ be the solution of the auxiliary problem

$$(\nabla z,\nabla\varphi) = (g',\varphi)\|g'\|^{-1},$$

satisfying $\|z\|_{2,2} \le c$. With this notation, there holds

$$\|g'\|^2 = (\nabla z,\nabla g') = (\nabla z,\delta') \le c\rho_*^{-2}\|\nabla z\|_{1;B_*} \le cL(\rho_*)\|z\|_{2,2} \le cL(\rho_*).$$

This finally implies the estimate (3.1.50). Q.E.D.

3.2 Solution of the discretized problems

In the following, we discuss techniques for solving the nonlinear algebraic system resulting from the finite element discretization of a quasi-linear elliptic problem as considered above,

$$a(u_h;\varphi_h) = (f,\varphi_h) \quad \forall\varphi_h \in V_h, \tag{3.2.51}$$

We concentrate on the special case that the nonlinear problem has the form

$$a(u;\varphi) = (F(\nabla u),\nabla\varphi) = (K(\nabla u)\nabla u,\nabla\varphi) \quad \forall\varphi \in V.$$

As particular example, we keep in mind the modified p-Laplace problem with

$$F(\nabla u) = K(\nabla u)\nabla u = \left(1+|\nabla u|^2\right)^{p/2-1}\nabla u, \quad 1 \le p < \infty.$$

3.2.1 A brief survey of iterative solution methods

Let $\{\varphi_h^i, i = 1,\dots,N_h = \dim V_h\}$ be the usual "nodal basis" of the finite element subspace $V_h \subset V = H_0^1(\Omega)$. In the case of "linear" finite elements these basis functiona are defined by the relation (so-called "Lagrange basis")

$$\varphi_h^i(a_j) = \delta_{ij}, \quad i,j = 1,\dots,N_h, \quad a_j \text{ interior nodal point of triangulation } \mathbb{T}_h.$$

Then, using the representation $u_h = \sum_{j=1}^{N_h} x_j\varphi_h^j$ the finite dimensional problem (3.2.51) can equivalently be written in the form

$$a(u_h;\varphi_h^i) = (f,\varphi_h^i) \quad i = 1,\dots,N_h. \tag{3.2.52}$$

This is a nonlinear algebraic system for the coefficient vector $x = (x_j)_{j=1}^{N_h}$ of the form

$$A(x)x = b, \tag{3.2.53}$$

where, analogously to the linear case, the (nonlinear) system matrix $A(x) = (a(x)_{ij})_{i,j=1}^{N_h}$ and the load vector $b = (b_i)_{i=1}^{N_h}$ are given by

$$a(x)_{ij} = (K(\nabla u_h)\nabla\varphi_h^j, \nabla\varphi_h^i),, \quad b_i = (f, \varphi_h^i)$$

For solving this problem, we consider the following standard iterative methods:

1. Fixed-point (defect correction) iteration
   The nonlinear system (3.2.53) is reformulated as a fixed point equation

   $$Cx = Cx + b - A(x)x, \quad C \text{ preconditioning matrix},$$

   which is then solved by the corresponding fixed point (defect correction) iteration

   $$C\delta x^t = d^t := b - A(x^{t-1})x^{t-1}, \quad x^t = x^{t-1} + \delta x^t. \tag{3.2.54}$$

   By the Banach fixed point theorem this iteration converges for appropriate starting values x^0 if the fixed point mapping

   $$G(x) := (I - C^{-1}A(x))x + C^{-1}b$$

   is a contraction on a closed subset $M \in \mathbb{R}^{N_h}$. From the linear situation with $A(x) = A$, we know that the speed of the convergence of such iterations depends essentially on the choice of the "preconditioner" C. This ranges from simple diagonal scaling such as in the Jacobi iteration with strong mesh-dependence up to sophisticated nonlinear multigrid schemes with sometimes almost mesh-independent behavior. In the nonlinear case considered, the behavior of these methods will not be better than in the linear case. Therefore, we have to expect a strong mesh-dependence of the convergence speed, i. e., the number of iterations required on a certain mesh $\mathbb{T}_h$ to reach the level of the discretization error grows rapidly with the dimension N_h (like h^{-2} in the simplest case). The defect correction iteration can also be formulated within the function space framework:

   $$c(\delta u_h^t, \varphi_h) = d^t(u_h^{t-1}; \varphi_h) := (f, \varphi_h) - a(u_h^{t-1}; \varphi_h), \quad \forall\varphi_h \in V_h, \quad u_h^t = u_h^{t-1} + \delta u_h^t,$$

   with an suitable (regular) bilinear form $c(\cdot,\cdot)$ (for preconditioning).

2. Functional iteration
   Problems of the form (3.2.53) can often be solved by a so-called "functional iteration", in which starting with a suitable initial guess x^0 a sequence of iterates is computed by successively solving the linear equations

   $$A(x^{t-1})x^t = b, \quad t \in \mathbb{N}. \tag{3.2.55}$$

   This iteration can also be formulated in function space. Starting from a suitable initial guess $u_h^0 \in V_h$ one computes iterates $u_h^t \in V_h$ by sucessively solving the linear problems

   $$(K(\nabla u_h^{t-1})\nabla u_h^t, \nabla\varphi_h) = (f, \varphi_h) \quad \forall\varphi_h \in V_h.$$

Functional iteration is particularly attractive if the problem is only mildly nonlinear such as the "semi-linear" diffusion equation

$$-\nabla \cdot (a(u)\nabla u) = f,$$

or the diffusion-transport equation

$$-\nu\Delta u + \beta(u) \cdot \nabla u = f.$$

Using the finite dimensionality of the problem one can often establish the convergence of the functional iteration, however without estimates for the speed of convergence.

3. Gradient method
If the nonlinear problem originates from a minimization problem, one may use a so-called "descent method" for its solution. The classical representative of this type of methods is the "gradient method". We formulate this method in the abstract Hilbert space setting. Here, starting from a suitable initial guess $u_h^0 \in V_h$ one determines a sequence of iterates by successively solving the one-dimensional minimization problems (so-called "line search")

$$u_h^{t+1} = u_h^t - \lambda_t g_h^t : \quad J(u_h^{t+1}) = \min_{\lambda \in \mathbb{R}_+} J(u_h^t - \lambda g_h^t), \tag{3.2.56}$$

where the descent direction (negative gradient direction) is determined by the linear equation

$$(g_h^t, \varphi_h) = J'(u_h^t)(\varphi_h) \quad \forall \varphi_h \in V_h.$$

Under certain conditions on the functional $J(\cdot)$ this method is known to converge globally with linear rate depending on the mesh size h.

In the linear case, which corresponds to the quadratic functional

$$J(u_h) = \tfrac{1}{2} a(u_h, u_h) - (f, u_h)$$

the gradient method looks particularly simple. The gradient at a point $u_h^t \in V_h$ is the defect functional

$$J'(u_h^t)(\varphi_h) = a(u_h^t, \varphi_h) - (f, \varphi_h), \quad \varphi_h \in V_h,$$

and the corresponding gradient direction $g_h^t \in V_h$ is given as the solution of the equation

$$(g_h^t, \varphi_h) = a(u_h^t, \varphi_h) - (f, \varphi_h), \quad \varphi_h \in V_h,$$

which is easy to solve. Then, the optimal step length λ_t in the line search is determined by

$$\lambda_t = \frac{a(u_h^t, g_h^t) - (f, g_h^t)}{a(g_h^t, g_h^t)} = \frac{\|g_h^t\|^2}{a(g_h^t, g_h^t)}.$$

The gradient method in its original form is as slow as the simplest fixed point iterations (Jacobi or Gauss-Seidel method) and its convergence depends strongly on the mesh size.

4. Newton iteration
The most popular method for solving nonlinear problems of type (3.2.53) is the Newton method and its variants. The classical Newton iteration for the system $A(x)x = b$ in

defect correction form reads as follows:

$$[x^{tT}A'(x^t) + A(x^t)]\delta x^t = b - A(x^t)x^t, \quad x^{t+1} = x^t + \delta x^t, \tag{3.2.57}$$

where $A'(x)$ denotes the derivative of the matrix $A(\cdot)$. Below, we shall consider the Newton method at first in $\mathbb{R}^n$. Here, the classical theorem of Newton-Kantorovich ensures the local quadratic convergence of the Newton method under certain structural assumptions on the problem considered. However, in this formulation the dependence of the convergence on the dimension n, i. e., the mesh size in the present situation, does not become clear. Therefore, subsequently, we will consider the Newton method in a general function space setting,

$$a'(u_h^t; u_h^{t+1}, v_h) = a'(u_h^t; u_h^{t-1}, v_h) + (f, v_h) - a(u_h^t; v_h), \quad \forall v_h \in V_h, \tag{3.2.58}$$

or in form of a defect correction process,

$$a'(u_h^t; \delta u_h^t, v_h) = (f, v_h) - a(u_h^t; v_h) \quad \forall v_h \in V_h, \quad u_h^{t+1} = u_h^t + \delta u_h^7. \tag{3.2.59}$$

which allows us to prove convergence almost independent of the mesh size (so-called "mesh independence principle").

3.2.2 The Newton method in $\mathbb{R}^n$

We give a brief analysis of the classical Newton method in $\mathbb{R}^n$ for the approximation of roots of differentiable functions. Here, the notation $\|\cdot\|$ is used for any vector norm and the associated natural matrix norm (e. g., the Euclidean norm). The Euclidean scalar product is denoted by $\langle\cdot,\cdot\rangle$.

Let $D \subset \mathbb{R}^n$ be an open non-empty set and $f : D \to \mathbb{R}^n$ a differentiable function for which a root $z \in D$ is to be computed. The Jacobi matrix $f'(\cdot)$ is assumed to be regular on the level set

$$D_* := \big\{x \in D \,|\, \|f(x)\| \le \|f(x^*)\|\big\}, \quad x^* \in D \text{ arbitrarily fixed},$$

with uniformly bounded inverse,

$$\|f'(x)^{-1}\| \le \beta, \quad x \in D_*.$$

Further let $f'(\cdot)$ be uniformly Lipschitz continuous on D_*,

$$\|f'(x) - f'(y)\| \le \gamma\|x - y\|, \quad x, y \in D_*.$$

For this finite dimensional setting, we have the following Theorem of Newton-Kantorovich.

Theorem 3.4 (Newton-Kantorovich): *Under the above assumptions let for the initial point $x^0 \in D_*$ with $\alpha := \|f'(x^0)^{-1}f(x^0)\|$ the following condition be satisfies:*

$$q := \tfrac{1}{2}\alpha\beta\gamma < 1. \tag{3.2.60}$$

Then, the Newton iteration

$$f'(x^t)x^{t+1} = f'(x^t)x^t - f(x^t), \quad t \ge 1, \tag{3.2.61}$$

generates a sequence $(x^t)_{t\in\mathbb{N}} \subset D_*$, *which converges quadratically to a root* $z \in D_*$ *of* f. *Further, there holds the a priori error estimate*

$$\|x^t - z\| \le \frac{\alpha}{1-q}\, q^{(2^t-1)}, \quad t \ge 1. \tag{3.2.62}$$

Proof: To the starting point $x^0 \in D_*$ there corresponds the non-empty, closed level set

$$D_0 := \big\{x \in D \,|\, \|f(x)\| \le \|f(x^0)\|\big\} \subset D_*.$$

We consider the continuous mapping $g : D_0 \to \mathbb{R}^d$ defined by

$$g(x) := x - f'(x)^{-1} f(x), \quad x \in D_0.$$

(i) First, we derive some auxiliary results. For $x \in D_0$, we set

$$x_r := x - r f'(x)^{-1} f(x), \;\; 0 \le r \le 1,$$

and

$$R := \max\big\{r \in [0,1] \,|\, x_s \in D_0, 0 \le s \le r\big\} = \max\big\{r \in [0,1] \,|\, \|f(x_s)\| \le \|f(x^0)\|,\; 0 \le s \le r\big\}.$$

For the vector function $h(r) := f(x_r)$ there holds

$$h'(r) = -f'(x_r) f'(x)^{-1} f(x), \qquad h'(0) = -h(0).$$

For $0 \le r \le R$ this yields

$$\begin{aligned}
\|f(x_r)\| - (1-r)\|f(x)\| &\le \|f(x_r) - (1-r)f(x)\| = \|h(r) - (1-r)h(0)\| \\
&= \Big\| \int_0^r h'(s)\, ds + r h(0) \Big\| = \Big\| \int_0^r \{h'(s) - h'(0)\}\, ds \Big\| \\
&\le \int_0^r \|h'(s) - h'(0)\|\, ds,
\end{aligned}$$

and further obserbing $x_s - x = -s f'(x)^{-1} f(x)$:

$$\begin{aligned}
\|h'(s) - h'(0)\| &= \|\{f'(x_s) - f'(x)\} f'(x)^{-1} f(x)\| \\
&\le \gamma \|x_s - x\| \|f'(x)^{-1} f(x)\| \le \gamma s \|f'(x)^{-1} f(x)\|^2.
\end{aligned}$$

This yields

$$\|f(x_r)\| - (1-r)\|f(x)\| \;\le\; \tfrac{1}{2} r^2 \gamma \|f'(x)^{-1} f(x)\|^2. \tag{3.2.63}$$

With the quantity $\alpha_x := \|f'(x)^{-1} f(x)\|$ and the assumption $\|f'(x)^{-1}\| \le \beta$ it follows that

$$\|f(x_r)\| \le (1 - r + \tfrac{1}{2} r^2 \alpha_x \beta \gamma) \|f(x)\|.$$

In case that $\alpha_x \le \alpha$, we then obtain in view of the assumption $\frac{1}{2}\alpha\beta\gamma < 1$:

$$\|f(x_r)\| \le (1 - r + r^2) \|f(x)\|.$$

Consequently $R = 1$ in this case, i. e., $g(x) \in D_0$. For such $x \in D_0$, it follows that

$$\|g(x) - g^2(x)\| = \|g(x) - g(x) + f'(g(x))^{-1} f(g(x))\| \leq \beta \|f(g(x))\|.$$

With the help of the estimate (3.2.63) for $r = 1$, we obtain:

$$\|g(x) - g^2(x)\| \leq \tfrac{1}{2}\beta\gamma \|f'(x)^{-1} f(x)\|^2 = \tfrac{1}{2}\beta\gamma \|x - g(x)\|^2. \tag{3.2.64}$$

(ii) Next, we show that the Newton iterates $(x^t)_{t\in\mathbb{N}}$ exist in D_0 and satisfy the inequality

$$\|x^t - g(x^t)\| = \|f'(x^t)^{-1} f(x^t)\| \leq \alpha.$$

This is done by an induction argument. For $t = 0$ the assertion is obviously valid. Particularly, since $\alpha_{x^0} = \alpha$ there holds $g(x^0) \in D_0$. Let now $x^t \in D_0$ be an iterate with $g(x^t) \in D_0$ and $\|x^t - g(x^t)\| \leq \alpha$. Then,

$$\|x^{t+1} - g(x^{t+1})\| = \|g(x^t) - g^2(x^t)\| \leq \tfrac{1}{2}\beta\gamma \|x^t - g(x^t)\|^2 \leq \tfrac{1}{2}\alpha^2\beta\gamma \leq \alpha$$

and consequently in virtue of the above results $g(x^{t+1}) \in D_0$. Therefore, $(x^t)_{t\in\mathbb{N}} \subset D_0$ exists. Next, we show that this sequence is a Cauchy sequence. With the help of (3.2.64), we obtain

$$\|x^{t+1} - x^t\| = \|g^2(x^{t-1}) - g(x^{t-1})\| \leq \tfrac{1}{2}\beta\gamma \|g(x^{t-1}) - x^{t-1}\|^2 = \tfrac{1}{2}\beta\gamma \|x^t - x^{t-1}\|^2,$$

and iterating this inequality,

$$\begin{aligned}\|x^{t+1} - x^t\| &\leq \tfrac{1}{2}\beta\gamma \big(\tfrac{1}{2}\beta\gamma \|x^{t-1} - x^{t-2}\|^2\big)^2 \leq (\tfrac{1}{2}\beta\gamma)^{(2^2-1)} \|x^{t-1} - x^{t-2}\|^{(2^2)} \\ &\leq (\tfrac{1}{2}\beta\gamma)^{(2^2-1)} \big(\tfrac{1}{2}\beta\gamma \|x^{t-2} - x^{t-3}\|^2\big)^{(2^2)} = (\tfrac{1}{2}\beta\gamma)^{(2^3-1)} \|x^{t-2} - x^{t-3}\|^{(2^3)}.\end{aligned}$$

Continuing this iteration down to $t = 0$ and recalling $q = \frac{1}{2}\alpha\beta\gamma < 1$ yields

$$\|x^{t+1} - x^t\| \leq (\tfrac{1}{2}\beta\gamma)^{(2^t-1)} \|x^1 - x^0\|^{(2^t)} \leq (\tfrac{1}{2}\beta\gamma)^{(2^t-1)} \alpha^{(2^t)} \leq \alpha q^{(2^t-1)}.$$

For arbitrary $m \in \mathbb{N}$ it follows that

$$\begin{aligned}\|x^{t+m} - x^t\| &\leq \|x^{t+m} - x^{t+m-1}\| + \cdots + \|x^{t+2} - x^{t+1}\| + \|x^{t+1} - x^t\| \\ &\leq \alpha q^{(2^{t+m-1}-1)} + \cdots + \alpha q^{(2^{t+1}-1)} + \alpha q^{(2^t-1)} \\ &\leq \alpha q^{(2^t-1)} \big\{ (q^{(2^t)})^{(2^{m-1}-1)} + \cdots + q^{(2^t)} + 1 \big\} \\ &\leq \alpha q^{(2^t-1)} \sum_{j=0}^{\infty} (q^{(2^t)})^j \;=\; \frac{\alpha q^{(2^t-1)}}{1 - q^{(2^t)}}.\end{aligned}$$

This shows that $(x^t)_{t\in\mathbb{N}} \subset D_0$ is actually a Cauchy sequence. Its limit $z \in D_0$ is necessarily a fixed point of g and a root of f,

$$z = \lim_{t\to\infty} x^t = \lim_{t\to\infty} g(x^{t-1}) = g(z).$$

Letting $m \to \infty$ yields the asserted a priori error estimate (3.2.62),

$$\|z - x^t\| \leq \frac{\alpha q^{(2^t-1)}}{1 - q^{(2^t)}} \leq \frac{\alpha}{1-q} q^{(2^t-1)}.$$

To prove the a posteriori error estimate (**??**, we note that

$$f(x^t) = f(x^t) - f(z) = f'(\xi_t)(x^t - z), \quad \xi_t \in \overline{(x^t, z)},$$

what completes the proof. Q.E.D.

Remark 3.7: In the theorem of Newton-Kantorovich as stated above the assumptions are designed in such a way that also the existence of a root can be guaranteed. If this existence is being made an additional assumption, the argument for proving the local quadratic convergence of the Newton iteration can be significantly simplified.

In the realization of the Newton method there one has to cope with two main difficulties:

(i) high computational cost in each iteration step,

(ii) sufficiently "good" starting value x^0 required.

For overcome the first one of these difficulties one may use the so-called "simplified Newton iteration",

$$f'(c)\delta x^t = d^t := -f(x^t), \quad x^{t+1} = x^t + \delta x^t, \tag{3.2.65}$$

with a suitable point $c \in \mathbb{R}^n$, e. g., $c = x^0$, lying sufficiently close to the root z. In this iteration all linar systems to be solved have the same coefficient matrix, which can be utilized to lower the cost of these substeps (e. g., by computing once an LR decomposition of $f'(c)$ and using this in all subsequent iteration steps). However, this modification reduces the Newton method to a simple fixed point iteration with an only linear speed of convergence. To lower the difficulty of generating an appropriate starting value, one may try to enlarge the region of convergence of the Newton method by introducing a "damping" $\lambda_t \in (0, 1]$, which is adaptively adjusted in the course of the iteration,

$$f'(x^t)\delta x^t = d^t := -f(x^t), \quad x^{t+1} = x^t + \lambda_t \delta x^t. \tag{3.2.66}$$

The following theorem contains a useful damping strategy.

Theorem 3.5 (Damped Newton method): *Let the assumptions of Theorem 3.4 be satisfied. Then following the rule*

$$\lambda_t := \min\Big\{1, \frac{1}{\alpha_t \beta \gamma}\Big\}, \qquad \alpha_t := \|f'(x^t)^{-1} f(x^t)\|, \tag{3.2.67}$$

the damped Newton iteration (3.2.66) generates for any starting value $x^0 \in D_*$ a sequence $(x^t)_{t\in\mathbb{N}}$, for which after t_* steps the condition $q_* := \frac{1}{2}\alpha_{t_*}\beta\gamma < 1$ is satisfied. Then, for $t \geq t_*$ the iterates x^t converge quadratically to a root z of $f(x)$ with the a priori error estimate

$$\|x^t - z\| \leq \frac{\alpha}{1 - q_*} q_*^{(2^t-1)}, \quad t \geq t_*. \tag{3.2.68}$$

Proof: We use the notation from the proof of Theorem 3.4. For some $x \in D_0$ there holds, setting again $x_r := x - rf'(x)^{-1}f(x)$, $0 \le r \le 1$, and $\alpha_x := \|f'(x)^{-1}f(x)\|$ the estimate

$$\|f(x_r)\| \le (1 - r + \tfrac{1}{2}r^2\alpha_x\beta\gamma)\|f(x)\|, \quad 0 \le r \le R = \max\{r \in [0,1] \,|\, x_s \in D_0, 0 \le s \le r\}.$$

The prefactor becoms minimal for

$$r_* = \min\Big\{1, \frac{1}{\alpha_x\beta\gamma}\Big\} > 0 : \quad 1 - r_* + \tfrac{1}{2}r_*^2\alpha_x\beta\gamma \;\le\; 1 - \frac{1}{2\alpha_x\beta\gamma} < 1.$$

For the choice of

$$r_t := \min\Big\{1, \frac{1}{\alpha_t\beta\gamma}\Big\}, \quad \alpha_t := \|f'(x^t)^{-1}f(x^t)\| \le \beta\|f(x^t)\|,$$

we have $(x^t)_{t\in\mathbb{N}} \subset D_0$, and the norm $\|f(x^t)\|$ is strictly monotonically decreasing, i. e.,

$$\|f(x^{t+1})\| \le \Big(1 - \frac{1}{2\alpha_t\beta\gamma}\Big)\|f(x^t)\|.$$

Therefore, after finitely many, $t_* \ge 1$, iteration steps, we have $\frac{1}{2}\alpha_{t_*}\beta\gamma < 1$, and the quadratic convergence of the subsequent iteration $(x^t)_{t\ge t_*}$ follows from Theorem 3.4. Q.E.D.

3.2.3 The Newton method in function space

Now. we consider the Newton method formulated in function space for computing the solutions $u_h \in V_h \cap B_R(u)$ of the discretized problems corresponding to a general quasi-linear elliptic problem.

$$a(u;\varphi) := (F(\nabla u), \nabla\varphi) = (f,\varphi) \quad \forall\varphi \in V = H_0^1(\Omega).$$

Suppose that the discretization uses a subspace $V_h \subset V$ of piecewise linear elements on a regular triangulation $\mathbb{T}_h$ of $\overline{\Omega}$ with the usual nodal basis $\{\varphi_h^i,\, i = 1,\ldots,N_h = \dim V_h\}$. Then, the discrete problem in the function space V_h

$$a(u_h;\varphi_h) = (f,\varphi_h) \quad \forall\varphi_h \in V_h,$$

is equivalent to a nonlinear algebraic system for the coefficient vector $x = (x_j)_{j=1}^{N_h} \in \mathbb{R}^{N_h}$ in the representation $u_h = \sum_{j=1}^{N_h} x_j\varphi_h^j$,

$$f(x) = 0,$$

where the mapping $f = (f_i)_{i=1}^{N_h} : D \subset \mathbb{R}^{N_h} \to \mathbb{R}^{N_h}$ is givem by

$$f_i(x) := a(\sum_{j=1}^{N_h} x_j\varphi_h^j;\varphi_h^i) - (f,\varphi_h^i) = (F(\sum_{j=1}^{N_h} x_j\nabla\varphi_h^j),\nabla\varphi_h^i) - (f,\varphi_h^i), \quad i = 1,\ldots,N_h.$$

As usual the vector function $F(\eta) = (F_i(\eta))_{i=1}^2$, is assumed to be continuously differentiable with, for bounded arguments, positive definite and Lipschitz continuous Jacobian matrix $F'(\eta) = (\partial_{\eta_j}F_i(\eta))_{i,j=1}^2$:

$$\langle F'(\eta)\xi,\xi\rangle \ge \alpha|\xi|^2, \quad \xi \in \mathbb{R}^2, \qquad |F'(\eta) - F'(\eta')| \le \gamma|\eta - \eta'|, \quad \eta,\eta' \in B_R.$$

We want to show that these properties carry over to the mapping $f(\cdot)$. The Jacobian matrix $f'(\cdot)$ of the mapping $f(\cdot)$ has the elements

$$f'_{ij}(x) = \frac{\partial}{\partial_{x_j}} f_i(x) = (F'(\sum_{j=1}^{N_h} x_j \nabla\varphi_h^j)\nabla\varphi_h^j, \nabla\varphi_h^i), \quad i,l = 1,\dots,N_h.$$

Hence, for any two finite element functions and their corresponding nodal basis representations, $v_h = \sum_{j=1}^{N_h} y_j\varphi_h^j$, $w_h = \sum_{j=1}^{N_h} z_j\varphi_h^j \in V_h$, there holds

$$\begin{aligned}\langle f'(x)y, z\rangle &= \sum_{i,k=1}^{N_h} \langle f'_{ik}(x)y_k, z_i\rangle = \sum_{i,k=1}^{N_h} (F'(\sum_{j=1}^{N_h} x_j\nabla\varphi_h^j)y_k\nabla\varphi_h^k, z_i\nabla\varphi_h^i)\\ &= (F'(u_h)\nabla v_h, \nabla w_h) = a'(u_h; v_h, w_h).\end{aligned}$$

The usual assumptions on $F(\cdot)$ imply that the tangent form $a'(u_h;\dot{;}\cdot)$ is uniformly V-elliptic and Lipschitz continuous, for arguments $u_h, u'_h \in V_h \cap B_R(u)$, in the following sense:

$$a'(u_h; v_v, v_h) \geq \alpha\|\nabla v_h\|^2,$$

$$|a'(u_h; v_h, w_h) - a'(u'_h; v_h, w_h)| \leq \gamma\|\nabla(u_h - u'_h)\|_\infty\|\nabla v_h\|\|\nabla w_h\|.$$

Obviously, these properties carry over to the Jacobian matrix $f'(\cdot)$ as follows:

$$(f'(x)y, y) = a'(u_h; v_h, v_h) \geq \alpha\|\nabla v_h\|^2 \geq \alpha\langle Ay, y\rangle = \alpha\|y\|_A^2, \quad y \in \mathbb{R}^{N_h},$$

$$\begin{aligned}|f'(x; y, z) - f'(x'; y, z)| &= a'(u_h; v_h, v_h) - a'(u'_h; v_h, v_h)\\ &\leq \gamma\|\nabla(u_h - u'_h)\|_\infty\|\nabla v_h\|\|\nabla w_h\|\\ &\leq \gamma\|\nabla(u_h - u'_h)\|_\infty\|y\|_A\|z\|_A\\ &\leq c\gamma h^{-1}\|x - x'\|_A\|y\|_A\|z\|_A,\end{aligned}$$

with the positive definite (not uniformly in h) "stiffness matrix" $A = ((\nabla\varphi_h^j, \nabla\varphi_h^i))_{i,j=1}^{N_h}$

In view of the foregoing results, it appears most natural in the present situation to use the theorem of Newton-Kantorovich for the special mapping $f(\cdot)$ with the norm $\|\cdot\| := \|\cdot\|_A = \langle A\cdot,\cdot\rangle^{1/2}$ since then all assumption of the theorem are satisfied. However, due to the use of the inverse relation

$$\|\nabla v_h\|_\infty \leq ch^{-1}\|\nabla v_h\|, \quad v_h \in V_h,$$

the Lipschitz continuity of the Jacobian $f'(\cdot)$ is strongly h-dependent with a constant behaving like $\mathcal{O}(h^{-1})$. On the other hand, the direct use of the norm $\|\cdot\| := \|\nabla\cdot\|_\infty$ leads to likewise inpractical bounds for the inverse of the Jacobian matrix $f'(x)^{-1}$. This incompatibility results in unrealistically strong requirements on the quality of the initial guess x^0 to guarantee convergence within the B_R-ball of the solution u. For this reason, in our further analysis of the Newton method, we prefer the function space setting, which is more appropriate for the structural properties of the nonlinear problems considered.

Starting with a suitable initial guess $u_h^0 \in V_h \cap B_R(u)$, one obtains a sequence of iterates $u_h^t \in V_h \cap B_R(u)$, $t \in \mathbb{N}$, by sucessively solving the linear problems

$$a'(u_h^t; u_h^{t+1}, v_h) = a'(u_h^t; u_h^{t-1}, v_h) - a(u_h^t; v_h) + (f, v_h), \quad \forall v_h \in V_h, \tag{3.2.69}$$

or in form of a defect correction process,

$$a'(u_h^t;\delta u_h^t,v_h)=(f,v_h)-a(u_h^t;v_h)\quad\forall v_h\in V_h,\quad u_h^{t+1}=u_h^t+\delta u_h^7. \tag{3.2.70}$$

In view of Theorem 3.1, the theorem of Newton-Kantorovich, and the foregoing discussion the local quadratic convergence of this iteration is guaranteed for any fixed h sufficiently small, say $h\le h_1\le h_0$, if the starting value u_h^0 is taken sufficiently close to the discrete solution $u_h\in V_h\cap B_R(u)$. But, since our assumptions do not ensure the convergence of the Newton iteration in function space for the "continuous" problem, it cannot be expected that the speed of convergence is uniform as $h\to 0$. However, even in the present rather general situation, we have the following almost mesh-independence result for the Newton method. The proof is less complicated than that of the theorem of Newton-Kantorovic since the existence of a root of the nonlinear mapping considered is a priori known and does not need to be established in the course of the argument.

Theorem 3.6 (Mesh-independence of Newton method): *For sufficiently small h, $h\le h_1$, there exist positive constants c_1,c_2,c_3 independent of h, such that for any starting value $u_h^0\in V_h\cap B_R(u)$ satisfying*

$$|||u-u_h^0|||_\infty\le c_1L(h)^{-1}, \tag{3.2.71}$$

the Newton iterates $u_h^t\in V_h\cap B_R(u)$, $t\in\mathbb{N}$, are well defined and the accuray of the discretization error is reached after at most $t_h\le c_2L(L(h))$ steps,

$$|||u-u_h^t|||_\infty\le c_3L(h)h^{1-2/p},\quad t\ge t_h. \tag{3.2.72}$$

Proof: The proof is similar to that of Theorem 3.1.

i) First, assuming the existence of the iterate $u_h^{t-1}\in V_h\cap B_R(u)$, for some $t\in\mathbb{N}$, we consider the next iterate $u_h^t\in V_h$, which is defined by

$$a'(u_h^{t-1};u_h^t,v_h)=a'(u_h^{t-1};u_h^{t-1},v_h)-a(u_h^{t-1};v_h)+(f,v_h),\quad\forall v_h\in V_h.$$

For this, we derive the identity

$$\begin{aligned}
a'(u;R_hu-u_h^t,v_h)&=a'(u;u-u_h^t,v_h)-a'(u_h^{t-1};u,v_h)+a'(u_h^{t-1};u,v_h)\\
&=a'(u;u-u_h^t,v_h)-a'(u_h^{t-1};u-u_h^t,v_h)+a'(u_h^{t-1};u-u_h^{t-1},v_h)\\
&\quad-a(u_h^{t-1};v_h)+(f,v_h)\\
&=a'(u;u-u_h^t,v_h)-a'(u_h^{t-1};u-u_h^t,v_h)+a'(u_h^{t-1};u-u_h^{t-1},v_h)\\
&\quad-a(u_h^{t-1};v_h)+a(u;v_h)\\
&=a'(u;u-u_h^t,v_h)-a'(u_h^{t-1};u-u_h^t,v_h)+a'(u_h^{t-1};u-u_h^{t-1},v_h)\\
&\quad-\int_0^1 a'(u_h^{t-1}+s(u-u_h^{t-1});u-u_h^{t-1},v_h)\,ds\\
&=a'(u;u-u_h^t,v_h)-a'(u_h^{t-1};u-u_h^t,v_h)\\
&\quad+\int_0^1\left\{a'(u_h^{t-1};u-u_h^{t-1},v_h)-a'(u_h^{t-1}+s(u-u_h^{t-1});u-u_h^{t-1},v_h)\right\}ds.
\end{aligned}$$

The right-hand side is again viewed as a functional $v_h^*\in V_h^*$ acting on the function $v_h\in V_h$,

with dual norm

$$\|v_h^*\|_{\infty;h} \le c(R)\big\{|||u-u_h^t|||_\infty + |||u-u_h^{t-1}|||_\infty\big\}|||u-u_h^{t-1}|||_\infty.$$

Then, analogously as in the proof of Theorem 3.1, we conclude the estimate

$$|||u-u_h^t|||_\infty \le c'L(h)h^{1-2/p} + c''L(h)\{|||u-u_h^t|||_\infty + |||u-u_h^{t-1}|||_\infty\}|||u-u_h^{t-1}|||_\infty, \quad (3.2.73)$$

with constants c', c'' independent of h and t. Setting

$$a_t := |||u-u_h^t|||_\infty, \quad \delta := c'L(h)h^{1-2/p}, \quad \beta := c''L(h),$$

this reads

$$a_t \le \delta + \beta(a_t + a_{t-1})a_{t-1}, \quad t \in \mathbb{N}.$$

ii) We want to show that, for sufficiently small h,

$$a_t \le 3\delta + \frac{1}{3\beta}(3\beta a_0)^{2^t}, \quad t \in \mathbb{N}. \quad (3.2.74)$$

From this, one easily obtains that, for a starting value u_h^0 satisfying (3.2.71), all iterates $u_h^t \in V_h \cap B_R(u)$ exist and satisfy (3.2.72) after $t \le t_h = c_2L(L(h))$ steps,

$$\begin{aligned} |||u-u_h^t|||_\infty = a_t &\le 3\delta + \frac{1}{3\beta}(3\beta a_0)^{2^t} \\ &\le 3c'L(h)h^{1-2/p} + \frac{1}{3\beta}(3c''L(h)c_1L(h)^{-1})^{2^t} \\ &\le 3c'L(h)h^{1-2/p} + \frac{1}{3\beta}(3c''c_1)^{2^t} \\ &\le 6c'L(h)h^{1-2/p}, \end{aligned}$$

for $3c''c_1 =: q < 1$ and $t \ge t_* \approx L(L(h))$ determined by

$$q^{2^t} \le 3c'L(h)h^{1-2/p}, \quad 2^t\log(q) \le \log(3c'L(h)h^{1-2/p}) \approx L(h), \quad t\log(2) \approx L(L(h)).$$

iii) To prove (3.2.74), we set $b_t := 3\beta a_t$ and require h to be sufficiently small, such that

$$c'c''L(h)^2h^{1-2/p} \le \tfrac{1}{9}.$$

Then, we have

$$b_t = 3\beta a_t \le 3\beta\delta + \tfrac{1}{3}(3\beta a_t + 3\beta a_{t-1})3\beta a_{t-1} = 3\beta\delta + \tfrac{1}{3}(b_t + b_{t-1})b_{t-1},$$

and further, in view of the above assumptions,

$$b_t \le 3c'c''L(h)^2h^{1-2/p} + \tfrac{1}{3}(b_t + b_{t-1})b_{t-1} \le \tfrac{1}{3} + \tfrac{1}{3}(b_t + b_{t-1})b_{t-1}.$$

By assumption, for $c_1 \le 1/3$, there holds $b_0 = 3\beta a_0 \le 3L(h)c_1L(h)^{-1} \le 1$. Then, by induction, we find that $b_t \le 1$, $t \in \mathbb{N}$, and, consequently,

$$b_t \le \tfrac{9}{2}\beta\delta + \tfrac{1}{2}b_{t-1}^2.$$

From this, we infer by induction that

$$b_t \le 9\beta\delta + b_0^{2^t}, \quad t \ge 1.$$

which implies (3.2.74). Obviously, this inequality holds true for $t = 1$. Suppose that it holds true for some $t - 1 \ge 1$. Then,

$$\begin{aligned} b_t &\le \tfrac{9}{2}\beta\delta + \tfrac{1}{2}b_{t-1}^2 \le \tfrac{9}{2}\beta\delta + \tfrac{1}{2}\big(9\beta\delta + b_0^{2^{t-1}}\big)^2 \\ &\le \tfrac{9}{2}\beta\delta + (9\beta\delta)^2 + b_0^{2^t} \le 9\beta\delta + b_0^{2^t}, \end{aligned}$$

for h sufficiently small such that $9\beta\delta \le 1/2$. This complestes the proof. Q.E.D.

3.2.4 The projective Newton method

Next, we consider a multi-level variant of the Newton method (3.2.69), which works on sucessively refined meshes and is sometimes referred to as "projective Newton method". Let again $V_t := V_{h_t} \subset V$ be the usual low-order finite element subspaces with decreasing mesh sizes $h_0 > h_1 > \cdots > h_t \to 0$ $(t \to \infty)$. Starting with an initial guess $u_0 \in V_0$ on the coarsest mesh, one determines increasingly accurate approximations $u_t \in V_t$, $t \in \mathbb{N}$, by sucessively solving the linear problems

$$a'(u_{t-1}; u_t, v) = a'(u_{t-1}; u_{t-1}, v) - a(u_{t-1}; v) + (f, v) \quad \forall v \in V_t. \tag{3.2.75}$$

These problems are of increasing complexity, i. e., $\dim V_t \to \infty$ $(t \to \infty)$. The question is now how rapidly the mesh size h_t may be decreased in this process without loosing the convergence properties of the Newton method. The following theorem gives an answer to this question.

Theorem 3.7 (Projective Newton method): *Suppose that the coarsest mesh size h_0 is sufficiently small and that the initial guess $u_0 \in V_0 \cap B_R(u)$ satisfies*

$$|||u - u_0|||_\infty \le c_0 h_0^{1-2/p}. \tag{3.2.76}$$

Further, let the mesh sizes h_t be chosen such that

$$h_t \ge \kappa h_{t-1}^2 L(h_{t-1})^{p/(p-2)}, \tag{3.2.77}$$

with a sufficiently large constant $\kappa > 0$. Then, the iterates $u_t \in V_t \cap B_R(u)$ are well defined and there holds

$$|||u - u_t|||_\infty \le c_1 h_t^{1-2/p}. \tag{3.2.78}$$

Proof: The proof follws the same line of argument as already used in the proof of Theorem 3.6. Assuming that the iterate $u_{t-1} \in V_{t-1} \cap B_R(u)$ exists, we obtain the following estimate for the next iterate $u_t \in V_t$:

$$|||u - u_t|||_\infty \le c' h_t^{1-2/p} + c'' L(h_t)\big\{|||u - u_t|||_\infty + |||u - u_{t-1}|||_\infty\big\}|||u - u_{t-1}|||_\infty .$$

Setting $a_t := |||u - u_t|||_\infty$, $\delta_t := c' h_t^{1-2/p}$, and $\beta_t := c'' L(h_t)$ this inequality reads

$$a_t \le \delta_t + \beta_t (a_t + a_{t-1}) a_{t-1}.$$

Next, we set $\sigma := \sup_{t\in\mathbb{N}} \beta_t/\beta_{t-1} \ge 1$ and fix a number $m \ge 2\sigma + 1$. Then for h_0 sufficiently small, such that $a_0 \le (m\beta_1)^{-1}$ and $\delta_t \le (m^2\beta_t)^{-1}$, an elemantary induction argument shows that $a_t \le (m\sigma\beta_t)^{-1}$. This implies that

$$a_t \le \gamma\delta_t + \gamma\beta_t a_{t-1}^2, \tag{3.2.79}$$

where $\gamma := m/(m-1)$. To show the error estimate (3.2.78), we note that it clearly holds for $t = 1$ with $c_1 := 2\gamma\sigma$. Suppose that it holds true for some $t - 1 \ge 2$. Then,

$$a_t \le \gamma\delta_t + \gamma\beta_t a_{t-1}^2 \le \gamma\delta_t + \kappa^{-2}\gamma c'' c_1^2 \big(L(h_t)^{p/(p-2)} h_{t-1}^2\big)^{1-2/p},$$

and, consequently, $a_t \le c_1\delta_t$, for sufficiently large κ. From this the asserted estimate follows again by induction. Q.E.D.

3.3 Exercises

Exercise 3.1: Let $\Omega \subset \mathbb{R}^2$ be a convex polygonal domain, $V := H_0^1(\Omega)$ and $V_h \subset V$ the usual approximating finite element spaces consisting of piecewise linear elements on a quasi-uniform family $\{\mathbb{T}_h\}_{h>0}$ of triangulations. Derive for the Ritz projection $R_h : V \to V_h$, defined by

$$(\nabla R_h v, \nabla\varphi) = (\nabla v, \nabla\varphi) \quad \forall\varphi \in V,$$

the L^2-stability estimate

$$\|R_h v\| \le c\|v\| + ch\|\nabla v\|, \quad v \in V \cap H^2(\Omega),$$

and use this to prove the optimal-order L^2-error estimate

$$\|u - R_h u\| \le ch^2 \|u\|_{2,2}, \quad v \in V \cap H^2(\Omega).$$

Do these results also hold true on meshes not necessarily satisfying the uniform size condition? (Hint: Use the standard "duality argument".)

Exercise 3.2: Show that the function

$$u(x) = (x_1 + x_2)\log(|x|), \quad x \in \Omega := \{y \in \mathbb{R}^2 \,|\, |y| < 1\},$$

is in $V = H_0^1(\Omega)$ and solves the variational equation

$$(\nabla u, \nabla\varphi)_\Omega = (g, \nabla\varphi)_\Omega \quad \forall\varphi \in V,$$

where the right-hand side vector $g = (g_1, g_2)$ is given as

$$g(x) = |x|^{-2}(x_1^2 + 2x_1x_2 - x_2^2, -x_1^2 + 2x_1x_2 + x_2^2)^T.$$

The right-hand side g is bounded while ∇u has a logarithmic singularity. This demonstrates

that the $\log(1/h)$ in the stability estimate

$$|||R_h g|||_\infty \le c \log(1/h) \|g\|_{\infty;h},$$

for the extended Ritz projection $R_h : V_h^* \to V_h$ introduced in class cannot be removed.

Exercise 3.3: With the notation introduced in class consider the standard low-order finite element approximation of the quasi-linear boundary value problem

$$-\nabla \cdot F(\cdot, \nabla u) = f \quad \text{in } \Omega, \quad u = g \text{ on } \partial\Omega,$$

on a convex polygonal domain $\Omega \subset \mathbb{R}^2$ with smooth boundary data $g \in H^{2,p}(\Omega)$ for some $p > 2$. Suppose that there exists a solution $u \in H^{2,p}(\Omega)$. Then, the main theorem presented in class for homogeneous boundary data also applies to this more general situation. Use this to show for the minimal surface problem the optimal-order L^2-error estimate

$$\|u - u_h\| \le c(u) h^2.$$

(Hint: Check that in the minimal surface problem the assumptions made for the main theorem are satisfied.)

Exercise 3.4: In the proof of the Main Theorem on the finite element approximation of general quasi-linear elliptic problems, we have made use of the fact that a subset Θ_h of the interval $[0,1]$, which is nonempty and open as well as closed with respect to $[0,1]$ necessarily coincides with the whole interval, i. e., $\Theta_h = [0,1]$. Give an argument for this fact.

Exercise 3.5: Extend the proof of the Main Theorem on the finite element approximation of general quasi-linear elliptic problems given in class, more precisely the homotopy argument, to cover also the H^1-error estimate

$$\|u - u_h\|_{1,2} \le c(u) h,$$

for solutions $u \in V = H_0^1(\Omega)$ satisfying $u \in V \cap H^{2,p}(\Omega)$ for some $p > 2$. To this end, one may use the following stability estimate for the Ritz projection $\hat{R}_h$:

$$\|\nabla \hat{R}_h v_h^*\| \le c \|v_h^*\|_{2;h}, \qquad \|v_h^*\|_{2;h} := \sup\{v_h^*(v_h),\ \|\nabla v_h\| = 1\}.$$

(Hint: Define an appropriate set $\Theta_h \subset [0,1]$ and show that it coincides with $[0,1]$. In proving the closedness of Θ_h, among others, the above stability estimate is needed.)

Exercise 3.6: Let $\{\mathbb{T}_h\}_{h>0}$ be a family of triangulations in $\mathbb{R}^d$, $d = 2, 3$, which is strongly shape uniform but needs to be only weakly size uniform. For any cell $T \in \mathbb{T}_h$ let $B_T \subset T$ denote the maximal inscribed circle. Derive for piecewise linear elements the estimate

$$\|u_h\|_{\infty;T} \le c \|u_h\|_{\infty;B_T}, \quad u_h|_T \in P_1(T),$$

with a constant c independent of T and h. Can this estimate also be guaranteed if the family of meshes is not strongly shape uniform? (Hint: One possible argument is based on Taylor expansion.)

Exercise 3.7: In class the "real" Green's function $G_x(y)$ has been used, for which in 2D, we know the estimate

$$|G_x(y)| \le c|\log(|x-y|)| + 1.$$

Use this to show that on any circle $B(x)$ with midpoint x and radius $\rho \le 1$, there holds the estimate

$$|B(x)|^{-1} \int_{B(x)} |G_x(y)|\, dy \le cL(\rho),$$

with a constant c independent of x and ρ. (Hint: Rewrite the integral using polar coordinates.)

Exercise 3.8: In class the L^∞-stability estimate

$$\|R_h u\|_\infty \le c\|u\|_\infty + hL(h)\|\nabla u\|_\infty, \quad u \in V \cap W^{1,\infty}(\Omega),$$

for the standard Ritz projection $R_h : V \to V_h$ has been proven for strongly shape uniform but only weakly size uniform mesh families. Use this result on strongly size and shape uniform (i. e., quasi-uniform) mesh families together with the inverse relation

$$\|\nabla v_h\|_\infty \le ch_{\min}^{-1}\|v_h\|_\infty, \quad v_h \in V_h,$$

and the interpolation estimate

$$\|v - I_h v\|_\infty + h_{\max}\|\nabla(v - I_h v)\|_\infty \le ch_{\max}\|\nabla v\|_\infty, \quad v \in W^{1,\infty}(\Omega),$$

to derive the $W^{1,\infty}$-stability estimate

$$\|\nabla R_h u\|_\infty \le cL(h)\|\nabla u\|_\infty, \quad u \in V \cap W^{1,\infty}.$$

Remark: Notice that this argument essentially depends on the strong size uniformity of the meshes and does not work if only weak size uniformity holds.

Exercise 3.9: Let $T \in \mathbb{T}_h$ be a triangle in a shape-regular family of triangulations of a polygonal domain $\Omega \subset \mathbb{R}^2$. Further, let B_T be a maximal inscribed circle of T with midpoint x_T and radius ρ_T. Prove for functions $v \in H^1(\Omega)$ the Sobolev-type inequality

$$|B_T|^{-1} \int_{B_T} |v|\, dx \le cL(\rho_T)\|v\|_{1,2},$$

where again $L(\rho) := \max\{\log(1/\rho_T), 1\}$. (Hint: You may use the triangle T containing B_T, polar coordinates, integration by parts and a suitable trace inequality.)

Exercise 3.10: The $W^{1,\infty}$-stability estimate for the finite element Ritz projection

$$\|\nabla R_h u\|_\infty \le cL(h)\|\nabla u\|_\infty,$$

proven in class involves the logarithmic term $L(h) = \max\{|\log(1/h)|, 1\}$. Develop an idea for avoiding the occurrence of this term in the proof in the case that the mesh family considered is quasi-uniform, particularly strongly size-uniform.
Remark: This idea cannot work in the proof of the L^∞-stability estimate

$$\|R_h u\|_\infty \le \|u\|_\infty + cL(h)\|\nabla u\|_\infty,$$

since here the logarithmic term is known to be present in general.

Exercise 3.11: State the finite dimensional problems resulting from the usual finite element discretization of the following nonlinear boundary value problems and formulate the Newton iteration in function space for their solution:

a) Minimal surface problem

$$\min \int_\Omega \sqrt{1+|\nabla u|^2}\, dx \quad \text{on} \quad V_g := \{v \in H^1(\Omega) \,|\, v = g \ \text{ on } \partial\Omega\},$$

with g the trace of a prescribed function in $H^1(\Omega)$, $\Omega \subset \mathbb{R}^2$,

b) Semi-linear diffusion problem

$$-\nabla \cdot (a(u)\nabla u) = f \quad \text{in } \Omega, \quad u = 0 \ \text{ on } \partial\Omega,$$

for a scalar function $u \in V := H_0^1(\Omega)$, $\Omega \in \mathbb{R}^2$, with a continuously differentiable, positive coefficient function $a(\cdot) > 0$,

c) Diffusion-transport problem ("vector Burgers equation")

$$-\nu\Delta u + u \cdot \nabla u = f \quad \text{in } \Omega, \quad u = 0 \ \text{ on } \partial\Omega,$$

for a vector function $u \in V := H_0^1(\Omega)^2$, $\Omega \subset \mathbb{R}^2$, and a prescribed $f \in L^2(\Omega)^2$.

Exercise 3.12: Consider the theorem of Newton-Kantorovich on the convergence of the Newton method in $\mathbb{R}^n$ as presented in class. Show that in the one-dimensional case (n=1), under the corresponding assumptions of this theorem, the following a posteriori error estimate holds:

$$|x^t - z| \le \beta |f(x^t)| \le \tfrac{1}{2}\beta\gamma |x^t - x^{t-1}|^2, \quad t \ge 1.$$

Does the proof of this estimate also work in the multidimensional case $n \ge 2$?

Exercise 3.13: Let $V_h \subset V = H_0^1(\Omega)$ be the usual finite element subspaces of piecewise linear elements on a quasi-uniform family of triangulations $\{\mathbb{T}_h\}_{h>0}$. These finite dimensional spaces may be equipped with various norms, which for any fixed h are all equivalent, but with h-dependent equivalence constants $c(h)$. Determine this h-dependence for the following norms:

$$\begin{aligned}
&a) && \|v_h\|_2 \le c_1(h)\|v_h\|_\infty \le c_2(h)\|v_h\|_2, \quad v_h \in V_h,\\
&b) && \|v_h\|_{1,2} \le c_1(h)\|v_h\|_2 \le c_2(h)\|v_h\|_{1,2}, \quad v_h \in V_h,\\
&c) && \|v_h\|_2 \le c_1(h)\|\nabla v_h\|_\infty \le c_2(h)\|v_h\|_2, \quad v_h \in V_h.
\end{aligned}$$

Exercise 3.14: Consider the finite element discretization of a quasi-linear elliptic boundary value problem on a convex polygonal domain $\Omega \subset \mathbb{R}^2$,

$$a(u;\varphi) := (F(\nabla u), \nabla\varphi) = (f, \varphi) \quad \forall \varphi \in V = H_0^1(\Omega).$$

The discretization uses subspaces $V_h \subset V$ of piecewise linear elements on quasi-uniform triangulation $\mathbb{T}_h$ of $\overline{\Omega}$ with the usual nodal bases $\{\varphi_h^i,\, i = 1, \ldots, N_h = \dim V_h\}$. Then, the discrete

problems in the function spaces V_h

$$a(u_h;\varphi_h) = (f,\varphi_h) \quad \forall \varphi_h \in V_h,$$

are equivalent to nonlinear algebraic systems

$$f(x) = 0,$$

for the coefficient vectors $x = (x_i)_{i=1}^{N_h} \in \mathbb{R}^{N_h}$ in the representations $u_h = \sum_{i=1}^{N_h} x_i \varphi_h^i$.

a) Let the vector function $F(\cdot)$ be given in the (linear)(form $F(\nabla u) = a\nabla u$ with a constant $a > 0$. With the given notation, state explicitly the corresponding equivalent discrete problems in function space and in $\mathbb{R}^n$.

b) Proceed as in (a) but for the nonlinear vector function

$$F(\nabla u) = \frac{\nabla u}{\sqrt{1+|\nabla u|^2}}.$$

Exercise 3.15: Consider the Newton method for solving the finite dimensional problems resulting from the usual finite element discretization of the quasi-linear elliptic boundary value problem

$$a(u;\varphi) := (F(\nabla u), \nabla\varphi) = (f,\varphi) \quad \forall \varphi \in V = H_0^1(\Omega).$$

It follows from the theorem of Newton-Kantorovich proven in class that, if the vector function $F(\cdot)$ is differentiable with positive definite and Lipschitz continuous Jacobian, for fixed h the Newton iteration defined by

$$a'(u_h^{t-1}; u_h^t, \varphi_h) = a'(u_h^{t-1}; u_h^{t-1}, \varphi_h) + (f,\varphi_h) - .a(u_h^{t-1};\varphi_h) \quad \forall \varphi_{\in} V_h,$$

converges in the norm $|||\cdot|||_\infty := \|\nabla\cdot\|_\infty$ (locally) quadratically to the discrete solution u_h. Use a variant of the argument given in class to show that, if the Jacobian of $F(\cdot)$ is only continuous (not necessarily Lipschitz continuous), the Newton iteration still converges with superlinear speed,

$$\frac{|||u_h^t - u_h|||_\infty}{|||u_h^{t-1} - u_h|||_\infty} \to 0 \quad (t \to \infty).$$

(Hint: One may base the argument on the proof of the Newton-Kantorovich theorem as given in class.)

4 The (stationary) Navier-Stokes System

In this chapter, we consider the finite element approximation of the basic problem in mathematical Fluid Mechanics, the so-called "Navier-Stokes equations", which describe the behavior of the flow of certain fluids or gases. These equations are of diffusion-transport type and in contrast to most models in Structural Mechanics do not originate from a minimization principle. Their nonlinearity is of very special kind, which allows for a rather complete mathematical treatment with respect to theory as well as numerical approximation. The material of this chapter and further details can largely be found in the lecture notes Rannacher [4], the books of Temam [19], Galdi [7], Girault & Raviart [27], Brenner & Scott [34], and the articles Heywood & Rannacher [39], Heywood, Rannacher & Turek [40], Rannacher & Turek [50], Rannacher [45, 46, 47].

4.1 The stationary Navier-Stokes equations

With the so-called "kinematic viscosity" parameter ν, we obtain the classical (incompressible) "Navier-Stokes equations" in $\mathbb{R}^d$ $(d = 2, 3)$:

$$\partial_t v + v \cdot \nabla v - \nu \Delta v + \nabla p = f, \quad \nabla \cdot v = 0, , \tag{4.1.1}$$

for the velocity vector v and the scalar pressure p in a viscous fluid under the action of an exterior body force (e. g., gravity) with density f. This system represents the conservation equations for mass and momentum for a homogeneous (uniform material properties), isothermal (constant tempereature), incompressible (constant density set to one) Newtonian (linear material behavior, i. e. stress-strain relation) fluid. We consider this model on bounded domains $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$, which are assumed to be "sufficiently" regular. The equations (4.1.1) are supplemented by the "no-slip" condition along rigid walls Γ_{rigid} (justified by experimental observation) and non homogeneous Dirichlet conditions at inlets Γ_{in} and occasionally Neumann-type conditions at outlets Γ_{out} ($\partial\Omega = \Gamma_{\text{rigid}} \cup \Gamma_{\text{in}} \cup \Gamma_{\text{out}}$):

$$v = 0 \ \text{ on } \Gamma_{\text{rigid}}, \quad v = v^{\text{in}} \ \text{ on } \Gamma_{\text{in}}, \quad \nu \partial_n v + pn = 0 \ \text{ on } \Gamma_{\text{out}}.$$

We always assume that the rigid part of the boundary is non-trivial, $\text{meas}(\Gamma_{\text{rigid}}) > 0$. There are no boundary conditions explicitly imposed on the pressure. This setting is appropriate for modeling, e. g., flow through a finite pipe around an obstacle, where the flow is driven by a left-hand inflow with prescribed profile along $\partial\Omega_{\text{in}}$ of a globally defined divergence-free velocity field (e. g. Poiseuille flow) and at the right-hand outlet far behind the obstacle the flow is assumed to be parallel. The determination of the "correct" boundary condition along the artificial part of the boundary, Γ_{out}, is a difficult problem depending on the particular configuration considered. In the following, we often assume for simplicity that $\partial\Omega = \Gamma_{\text{rigid}}$, i. e., the flow is confined to a closed box and is solely driven by an exterior body force f.

In the case of large viscosity ν (as typical in biological fluids) the nonlinear acelleeration term $v \cdot \nabla v$ can be neglected and the Navier-Stokes system reduces to the linear "Stokes system":

$$\partial_t v - \nu \Delta v + \nabla p = f, \quad \nabla \cdot v = 0, , \tag{4.1.2}$$

which looks like the vector-heat equation but considered on the manifold of "incompressible" ("solenoidal" or "divergence-free") vector-fields. Despite its relatively simple structure, the general nonstationary Navier-Stokes equations pose some still unsolved mathematical questions

concerning existence and uniqueness of physical meaningful (i. e., regular and stable) solutions. Another prototypical model situation is that of the "lid-driven cavity" (flow in a closed box driven by the movement of the upper part of the wall. In the following, we will restrict us to the stationary case, i. e., neglect the time derivative in (4.1.1):

$$-\nu\Delta v + v\cdot\nabla v + \nabla p = f, \quad \nabla\cdot v = 0\,. \tag{4.1.3}$$

For setting up a variational formulation of the (stationary) Navier-Stokes problem, we introduce the following function spaces:

$$H = H^1_0(\Omega;\Gamma_D)^d := \{v\in H^1(\Omega)^d |\ v = 0 \ \text{ in } \Gamma_D\}\,, \quad L = L^2_0(\Omega) := \{q\in L^2(\Omega)|\,(q,1)_\Omega = 0\},$$

with the notation $\Gamma_D := \Gamma_{\text{rigid}}\cup\Gamma_{\text{in}}$. Then, the variational formulation of (4.1.1) reads: Find $v\in v^{\text{in}} + H$ and $p\in L$, such that

$$\nu(\nabla v,\nabla\varphi) + (v\cdot\nabla v,\varphi) - (p,\nabla\cdot v) = (f,\varphi) \quad \forall\varphi\in H, \tag{4.1.4}$$

$$(\nabla\cdot v,\chi) = 0 \quad \forall\chi\in L. \tag{4.1.5}$$

Let's assume that this problem has a solution, which is sufficiently regular for being a "classical" solution. By integration by parts in (4.1.4) and use of the boundary condition $\varphi_{|\Gamma_D} = 0$ one obtains that

$$\int_\Omega\{-\nu\Delta v + v\cdot\nabla v + \nabla p - f\}\varphi\,dx + \int_{\Gamma_{\text{out}}}\{\nu\partial_n v - pn\}\varphi\,do \;=\; 0, \quad \varphi\in H.$$

Consequently, each sufficiently regular solution of the variational problem is a classical solution of the Navier-Stokes equation and vice versa and it is obviously divergence-free. Further, it satisfies the following "natural" boundary condition on Γ_{out}:

$$\nu\partial_n v - pn = 0 \quad \text{on } \Gamma_{\text{out}}. \tag{4.1.6}$$

This boundary condition of Neumann-type results automatically from the variational formulation of the problem by imposing no explicit condition on Γ_{out}, which suggests the name "do nothing" outflow boundary condition. This (artificial) boundary condition appears natural as long as the flow can be assumed to be "parallel" across Γ_{out} and no further information about the flow behavior beyond Γ_{out} is available. Especially, (4.1.6) is satisfied by Poiseuille flow, if the pressure p is set to zero on Γ_{out}. Actually, in virtue of the divergence condition, $\nabla\cdot v = 0$, it turns out that the boundary condition (4.1.6) additionally implies a Dirichlet-type boundary condition for the pressure separately on each outlet, i. e., on each component of the outflow boundary,

$$\int_{\Gamma_{\text{out}}} p\,do = 0. \tag{4.1.7}$$

In the Navier-Stokes equations the pressure p appears under the gradient so that an additional (scalar) condition is required for guaranteeing its uniqueness. In configurations with "free" outflow boundary $\Gamma_{\text{out}}\neq\emptyset$ this is accomplished by the additional implicit condition (4.1.7). In the case $\Gamma_{\text{out}} = \emptyset$, instead one requires that p has mean value zero, i. e., $p\in L^2_0(\Omega)$. Below, we will show that the variational Navier-Stokes problem in the case $\partial\Omega = \Gamma_D$ always possesses a (not necessarily unique) solution. This proof requires some preparations.

4.1.1 The Stokes operator

We consider the linear Stokes problem with normalized viscosity parameter $\nu = 1$ and homogeneous Dirichlet boundary conditions along the whole boundary ($v|_{\partial\Omega} = 0$). Correspondingly, we shall use the function spaces $H := H_0^1(\Omega)^d$ and $L := L_0^2(\Omega)$. Then, the Stokes problem reads: Find $\{v, p\} \in H \times L$,such that

$$(\nabla v, \nabla\varphi) - (p, \nabla\cdot\varphi) = (f, \varphi) \quad \forall\varphi \in H, \tag{4.1.8}$$

$$(\nabla\cdot v, \chi) = 0 \quad \forall\chi \in L. \tag{4.1.9}$$

We introduce the further spaces:

$$J_1(\Omega) := \{v \in H_0^1(\Omega)^d \mid \nabla\cdot v = 0 \text{ a. e. in } \Omega\},$$
$$J_0(\Omega) := \{v \in L^2(\Omega)^d \mid \nabla\cdot v = 0,\ n\cdot v_{|\partial\Omega} = 0 \text{ in the "weak" sense}\}.$$

The space $J_1(\Omega)$ is eqipped with the scalar product $(\nabla\cdot, \nabla\cdot)$ a Hilbert space, and therefore complerte. The subspace $J_0(\Omega) \subset L^2(\Omega)^d$ is closed. The orthogonal projection from $L^2(\Omega)^d$ onto $J_0(\Omega)$ is denoted by P. The space $\Phi := \{\varphi \in C_0^\infty(\Omega)^d \mid \nabla\cdot\varphi \equiv 0\}$ of solenoidal "test functions" is densely contained in $J_0(\Omega)$ as well as in $J_1(\Omega)$. With this notation problem (4.1.8) can be written in the following compact form: Find $v \in J_1(\Omega)$, such that

$$(\nabla v, \nabla\varphi) = (Pf, \varphi) \quad \forall\varphi \in J_1(\Omega). \tag{4.1.10}$$

For $f \in L^2(\Omega)^d$, by the Poincaré inequality the right-hand side in (4.1.10) defines a bounded linear functional on $J_1(\Omega)$. Hence, by the representation theorem of Riesz there exits a unique solution $v \in J_1(\Omega)$ of (4.1.10). Then, the equation (4.1.10) defines a linear operator

$$S : D(S) \subset J_1(\Omega) \subset J_0(\Omega) \to J_0(\Omega),$$

the so-called "Stokes operator". As operator in $J_0(\Omega)$ it has the representation $S = -P\Delta$ and is obviously symmetric and positive definite. Further it is onto and consequently "self-adjoint". Because of the compactness of the embeddings $H^1(\Omega) \hookrightarrow L^2(\Omega)$ and $J_1(\Omega) \hookrightarrow J_0(\Omega)$ the inverse $S^{-1} : J_0(\Omega) \to J_0(\Omega)$ is a compact operator. By the general spectral theory of (positive definite) self-adjoint operators with compact inverses in Hilbert spaces, we know that the spectrum (set of singular values) of S consists of real (positive) eigenvalues with finite multiplicities and no finite accumulstion point:

$$0 < \lambda_1 \le ... \le \lambda_k \le ... .$$

Further, there exists a corresponding system of L^2-orthonormal eigenfunctions $\{w_k\}_{k\in\mathbb{N}}$, which is complete in $J_0(\Omega)$ as well as in $J_1(\Omega)$, i. e.: Each $v \in J_0(\Omega)$ possesses an expansion of the form

$$v = \sum_{k=1}^{\infty} \alpha_k w_k, \quad \alpha_k = (v, w_k).$$

We will use these properties in the existence proof for solutions of the Navier-Stokes problem, below.

4.1.2 Existence result for the Navier-Stokes problem

We consider again the case of pure Dirichlet boundary conditions: $H := H^1_0(\Omega;\Gamma_D)^d$ and $L := L^2_0(\Omega)$. The Dirichlet data along inflow/outflow parts of the boundary are assumed to be given as traces of a solenoidal vector field $v^{\rm in} \in H^1(\Omega)^d$ satisfying the no-slip condition along $\Gamma_{\rm rigid}$. Using the function space $J_1(\Omega)$ the stationary Navier-Stokes problem (4.1.4) is written in compact form with pressure being eliminated: Find $v \in v^{\rm in} + J_1(\Omega)$, such that

$$\nu(\nabla v, \nabla\varphi) + (v\cdot\nabla v, \varphi) = (f,\varphi) \quad \forall\varphi \in J_1(\Omega). \tag{4.1.11}$$

This is the starting point for the proof of existence of solutions to the Navier-Stokes equations. For simplicity, from now on we will use the notation $V := J_1(\Omega)$ for the "solution space".

The special structure of the nonlinearity in the Navier-Stokes equations is decisive for their analysis. For functions $u, v, w \in V = J_1(\Omega)$ there holds

$$(u\cdot\nabla v, w) = (v, \nabla(vw)) - (u\cdot\nabla w, v) = -(\nabla\cdot u, vw) - (u\cdot\nabla w, v) = -(u\cdot\nabla w, v).$$

Setting here $v = w$, we obtain the important identity

$$(u\cdot\nabla v, v) = 0\,. \tag{4.1.12}$$

For the precise description of the natural regularity of the right-hand side, we use the "negative" Sobolev norm

$$\|f\|_{-1} := \sup_{\varphi\in H} \frac{\langle f,\varphi\rangle}{\|\nabla\varphi\|},$$

which is the natural norm of the dual space $H^{-1}(\Omega)$ of H. Clearly, for functions $f \in L^2(\Omega)^d$ there holds

$$\|f\|_{-1} \le \kappa\|f\|,$$

with the constant $\kappa > 0$ in the Poincaré inequality

$$\|\varphi\| \le \kappa\|\nabla\varphi\|, \quad \varphi \in H.$$

Theorem 4.1 (Existence theorem): *In the case $\partial\Omega = \Gamma_{\rm rigid}$ the stationary Navier-Stokes problem (4.1.4) possesses for any value of the viscosity parameter $\nu > 0$ at least one solution $\{v,p\} \in H \times L$. For sufficiently small data $c_*^2\nu^{-2}\|f\|_{-1} < 1$ this solution is unique.*

Proof: *i) Existenz:* We use the technic of "Galerkin approximation". With the eigenfunctions of the Stokes operator $\{w_i\}_{i\in\mathbb{N}}$, we define the finite dimensional subspaces $V_m := \text{span}\langle w_1, ..., w_m\rangle \subset V$. Problem (4.1.4) is then approximated by the following finite dimensional problems: Find $v_m \in V_m$, such that

$$\nu(\nabla v_m, \nabla\varphi) + (v_m\cdot\nabla v_m, \varphi) = (f,\varphi) \quad \forall\varphi \in V_m. \tag{4.1.13}$$

We want to show that these finite dimensional (nonlinear) problems possess solutions, which are uniformly bounded in V. Then, by a compactness argument, we conclude the existence of a solution of the infinite dimensional problem. To each $v \in V_m$, we associate an element $Q_m(v) \in V_m$ as the solution of the (finite dimensional) linear problem

$$\nu(\nabla Q_m(v), \nabla\varphi) + (v\cdot\nabla Q_m(v), \varphi) = (f,\varphi) \quad \forall\varphi \in V_m.$$

The solvability of this linear problem follows from the fact that corresponding homogeneous problem

$$\nu(\nabla w, \nabla\varphi) + (v\cdot\nabla w, \varphi) = 0 \quad \forall\varphi \in V_m$$

only has the trivial solution $w = 0$, what is easily seen by taking $\varphi = w$,

$$0 = \nu\|\nabla w\|^2 + (v\cdot\nabla w, w) = \nu\|\nabla w\|^2.$$

This defines a (nonlinear) mapping $Q_m : V_m \to V_m$. Setting $\varphi := Q_m(v)$ in the defining equation, we get

$$\nu\|\nabla Q_m(v)\|^2 = (f, Q_m(v)) \le \|f\|_{-1}\|\nabla Q_m(v)\|$$

and, consequently,

$$\|\nabla Q_m(v)\| \le \frac{\|f\|_{-1}}{\nu} =: R.$$

Hence, the mapping Q_m maps the (compact) ball $V_m \cap B_R := \{v \in V_m | \, |\, \|\nabla v\| \le R\}$ into itself. It is continuous (actually Lipschitz continuous), what can be deduced from the following relation for arbitrary $v, w \in V_m \cap B_R$:

$$\begin{aligned} 0 &= \nu(\nabla[Q_m(v) - Q_m(w)], \nabla\varphi) + (v\cdot\nabla Q_m(v) - w\cdot\nabla Q_m(w), \varphi) \\ &= \nu(\nabla[Q_m(v) - Q_m(w)], \nabla\varphi) + ((v-w)\cdot\nabla Q_m(v), \varphi) \\ &\quad + (w\cdot\nabla[Q_m(v) - Q_m(w)], \varphi) \quad \forall\varphi \in V_m. \end{aligned}$$

In $d = 2$ and $d = 3$ dimensions, we have the inequalities (with $\|\cdot\|_p := \|\cdot\|_{L^p}$)

$$\|w\|_3 \le c_*\|\nabla w\|, \quad \|\varphi\|_6 \le c_*\|\nabla\varphi\|, \quad w, \varphi \in V,$$

with a generic constant c_* only depending on Ω. Hence, chosing $\varphi := Q_m(v) - Q_m(w)$ yields

$$\begin{aligned} \nu\|\nabla[Q_m(v) - Q_m(w)]\|^2 &= -((v-w)\cdot\nabla Q_m(v), Q_m(v) - Q_m(w)) \\ &\le \|v-w\|_3\|\nabla Q_m(v)\|\|Q_m(v) - Q_m(w)\|_6 \\ &\le c_*^2\|\nabla(v-w)\|\|\nabla Q_m(v)\|\|\nabla(Q_m(v) - Q_m(w))\| \end{aligned}$$

and further

$$\|\nabla[Q_m(v) - Q_m(w)]\| \le c_*^2 R\|\nabla(v-w)\|.$$

Then, by the fixed point theorem of Brouwer the continuous mapping $Q_m : V_m\cap B_R \to V_m\cap B_R$ possesses (at least) one fixed point $v_m \in V_m$. By definition, this fixed point satisfies the equation

$$\nu(\nabla v_m, \nabla\varphi) + (v_m\cdot\nabla v_m, \varphi) = (f, \varphi) \quad \forall\varphi \in V_m,$$

and the uniform bound $\|\nabla v_m\| \le R$. The Hilbert space $V \subset H$ is compactly embedded into $J_0(\Omega) \subset L$. Hence, there exists a subsequence $(v_{m'})_{m'\in\mathbb{N}}$, which converges weakly in $J_1(\Omega)$ and strongly in $J_0(\Omega)$ to a function $v \in J_1(\Omega)$,

$$(\nabla(v_{m'} - v), \nabla\varphi) \to 0 \quad \forall\varphi \in J_1(\Omega), \qquad \|v_{m'} - v\| \to 0 \quad (m' \to \infty).$$

This limit $v \in J_1(\Omega)$ is then also solution of equation (4.1.4). To see this, we take an arbitrary $\varphi \in J_1(\Omega)$ und a sequence $\varphi_{m'} \in V_{m'}$, such that $\|\nabla(\varphi - \varphi_{m'})\| \to 0$ $(m' \to \infty)$. Then, for

$m' \to \infty$ (exercise),

$$(\nabla v_{m'}, \nabla \varphi_{m'}) \to (\nabla v, \nabla \varphi),$$
$$(v_{m'} \cdot \nabla v_{m'}, \varphi_{m'}) \to (v \cdot \nabla v, \varphi),$$
$$(f, \varphi_m) \to (f, \varphi),$$

and, we obtain in the limit that

$$\nu(\nabla v, \nabla \varphi) + (v \cdot \nabla v, \varphi) = (f, \varphi).$$

The existence of an associated pressure $p \in L$, which is uniquely determined in the space $L_0^2(\Omega)$, is supplied by Lemma 4.1, below.

ii) Uniqueness: For any solution $v \in V$, there holds

$$\nu \|\nabla v\|^2 = (f, v) - (v \cdot \nabla v, v) \leq \|f\|_{-1} \|\nabla v\|,$$

and, consequently,

$$\|\nabla v\| \leq \nu^{-1} \|f\|_{-1}.$$

Let now $v_1, v_2 \in V$ be two solutions. For the difference $w := v_1 - v_2$, there holds

$$\begin{aligned} \nu(\nabla w, \nabla \varphi) &= (v_2 \cdot \nabla v_2 - v_1 \cdot \nabla v_1, \varphi) \\ &= ((v_2 - v_1) \cdot \nabla v_2, \varphi) - ((v_2 - v_1) \cdot \nabla (v_2 - v_1), \varphi) + (v_1 \nabla (v_2 - v_1), \varphi) \\ &= (w \cdot \nabla v_1, \varphi) - (w \cdot \nabla w, \varphi) + (v_1 \cdot \nabla w, \varphi), \quad \varphi \in V, \end{aligned}$$

Setting $\varphi = w$, we find

$$\nu \|\nabla w\|^2 = (w \cdot \nabla v_1, w) \leq \|w\|_3 \|\nabla v_1\| \|w\|_6 \leq c_*^2 \|\nabla w\|^2 \nu^{-1} \|f\|_{-1}.$$

In case that $c_*^2 \nu^{-2} \|f\|_{-1} < 1$ this implies that necessarily $w = 0$. Q.E.D.

Once a solution $v \in V$ of the variational problem (4.1.11) is determined, it remains to show the existence of a correpondig pressure function $p \in L$, such that

$$(p, \nabla \cdot \varphi) = (f, \varphi) + \nu(\nabla v, \nabla \varphi) - (v \cdot \nabla v, \varphi) \quad \forall \varphi \in H. \tag{4.1.14}$$

The right-hand side of (4.1.14) represents a linear functional $l(\cdot)$ on H with the property $l(\varphi) = 0$, $\varphi \in V$. The existence of a corresponding pressure and its stability is ensured by the following result.

Lemma 4.1 ("inf-sup" inequality): *(i) For each linear functional $l(\cdot)$ on H with the property $l(\varphi) = 0$, $\varphi \in J_1(\Omega)$, there exists a uniquely determined $p \in L$, such that*

$$(p, \nabla \cdot \varphi) = l(\varphi) \quad \forall \varphi \in H, \qquad \beta \|p\| \leq \sup_{\varphi \in H} \frac{(p, \nabla \cdot \varphi)}{\|\nabla \varphi\|}. \tag{4.1.15}$$

(ii) To each function $p \in L$ there exists a function $v \in H$ with the property

$$p = \nabla \cdot v, \quad \|\nabla v\| \leq \beta \|p\|, \tag{4.1.16}$$

with a constant $\beta > 0$ *independent of* p. *Further there holds the stability inequality (continuous "inf-sup" inequality):*

$$\inf_{q\in L}\sup_{\varphi\in H}\frac{(q,\nabla\cdot\varphi)}{\|q\|\,\|\nabla\varphi\|}\geq\beta>0\,. \tag{4.1.17}$$

Proof: We use a functional analytic argument following Girault/Raviart [27] (Chapter 1); an alternative potential theoretical proof can be found in Galdi [7].
(i) We embed the present situation into an abstract functional analytic framework. Starting point are the Hilbert spaces $L^2(\Omega)$ and $H = H_0^1(\Omega)^d$ with the corresponding norms $\|\cdot\|$ and $\|\nabla\cdot\|$ and their dual spaces $L^2(\Omega)^* \cong L^2(\Omega)$ and $H^* = H^{-1}(\Omega)^d$ (consistimg of linear continuous functionals). By

$$\langle -\mathrm{grad}\, p, \varphi\rangle := (p, \mathrm{div}\,\varphi), \quad \varphi\in H,$$

the gradient (in distributional sense) is defined as linear operator $-\mathrm{grad} : L^2(\Omega)\to H^*$. The corresponding adjoint operator is the divergence operator $\mathrm{div} : H\to L^2(\Omega)^*\cong L^2(\Omega)$. For the image spaces (range) $R(\cdot)$ and null spaces $N(\cdot)$ of these operators there holds by general princioles:

$$\overline{R(\mathrm{grad})} = N(\mathrm{div})^0 = J_1(\Omega)^0, \qquad \overline{R(\mathrm{div})} = N(\mathrm{grad})^0 = J_0(\Omega).$$

Here, $N(\mathrm{div})^0 := \{\chi\in H^* :\ \langle\chi,\varphi\rangle = 0,\ \varphi\in N(\mathrm{div})\}$. A deep result from distribution theory (theorem of De Rham) implies that $R(\mathrm{grad})\subset H^*$ is closed. Hence, $R(\mathrm{grad}) = J_1(\Omega)^0$, what implies the first assertion (i).
(ii) Further there holds $N(\mathrm{grad}) = \mathrm{span}\{1\}$, such that the reduced operator

$$\widetilde{\mathrm{grad}} : L = L_0^2(\Omega)\to J_1(\Omega)^0$$

is an isomorphism. Since L and $J_1(\Omega)^0$ are Hilbert spaces, another general result implies that the operator $\widetilde{\mathrm{grad}}$ is also an isomorphism, i. e., the inverse operator $\widetilde{\mathrm{grad}}^{-1} : J_1(\Omega)^0\to L$ exists and is bounded:

$$\beta\|\widetilde{\mathrm{grad}}^{-1}(v)\|\leq\sup_{\varphi\in H}\frac{\langle v,\varphi\rangle}{\|\nabla\varphi\|}.$$

This shows that for arbitrary $p\in L$ there holds

$$\beta\|p\|\leq\sup_{\varphi\in H}\frac{\langle\widetilde{\mathrm{grad}}(p),\varphi\rangle}{\|\nabla\varphi\|} = \sup_{\varphi\in H}\frac{(p,\nabla\cdot\varphi)}{\|\nabla\varphi\|}\,.$$

With $\widetilde{\mathrm{grad}}$ also its adjoint operator $\widetilde{\mathrm{div}} : (J_1(\Omega)^0)^*\to L^*$ is an isomorphism. The space $J_1(\Omega)^0\subset H^*$ can be identified with the orthogonal complement of $J_1(\Omega)$ in H:

$$J_1(\Omega)^0\cong J_1(\Omega)^\perp := \{v\in H :\ (\nabla v,\nabla\varphi) = 0\ \ \forall\varphi\in J_1(\Omega)\}\,.$$

Consequently, for each $p\in L$ there is a $v\in J_1(\Omega)^\perp$ such that

$$p = \nabla\cdot v\,, \quad \beta\|\nabla v\|\leq\|p\|.$$

This completes the proof. Q.E.D.

Remark 4.1: For illustrating the important surjectiviy of the divergence operator, we give a

heuristic argument. Let $p \in L = L^2_0(\Omega)$ be give. The Neumann boundary value problem of the Laplacian,

$$\Delta z = p \text{ in } \Omega, \quad \partial_n v_{|\partial\Omega} = 0\,,$$

possesses a unique (weak) solution $z \in H^1(\Omega) \cap L^2_0(\Omega)$, which (on smoothly bounded or convex polygonal domains) possesses H^2-regularity and for which the following a priori estimate holds:

$$\|\nabla^2 z\| \le \beta^{-1}\|p\|.$$

The function $v := \nabla z$ has then obviously the following properties:

$$\nabla \cdot v = p, \qquad n \cdot v|_{\partial\Omega} = 0, \quad \beta\|\nabla v\| \le \|p\|.$$

Hence the assertion would be proven if additionally $\tau \cdot v_{|\partial\Omega} = 0$ with all tangential vectors τ at $\partial\Omega$.

Remark 4.2: :The case of inhomogeneous boundary data $v^{\text{in}} \not\equiv 0$ can be reduced to the situation considered in Theorem 4.1 by the following standard argument. Consider the flow field $u := v - v^{\text{in}} \in V$ which satisfies the equation

$$\begin{aligned}
\nu(\nabla u, \nabla\varphi) + (u \cdot \nabla u, \varphi) &= \nu(\nabla v, \nabla\varphi) - \nu(\nabla v^{\text{in}}, \nabla\varphi) + ((v - v^{\text{in}}) \cdot \nabla(v - v^{\text{in}}), \varphi) \\
&= \nu(\nabla v, \nabla\varphi) + (v \cdot \nabla v, \varphi) - (v^{\text{in}} \cdot \nabla v, \varphi) - (v \cdot \nabla v^{\text{in}}, \varphi) + (v^{\text{in}} \cdot \nabla v^{\text{in}}, \varphi) \\
&= (f, \varphi) - (v^{\text{in}} \cdot \nabla v, \varphi) - (v \cdot \nabla v^{\text{in}}, \varphi) + (v^{\text{in}} \cdot \nabla v^{\text{in}}, \varphi) \\
&= (f, \varphi) - (v^{\text{in}} \cdot \nabla u, \varphi) - (v^{\text{in}} \cdot \nabla v^{\text{in}}, \varphi) - (u \cdot \nabla v^{\text{in}}, \varphi)
\end{aligned}$$

and, consequently,

$$\nu(\nabla u, \nabla\varphi) + (u \cdot \nabla u, \varphi) + (v^{\text{in}} \cdot \nabla u, \varphi) + (u \cdot \nabla v^{\text{in}}, \varphi) = (f, \varphi) - (v^{\text{in}} \cdot \nabla v^{\text{in}}, \varphi) \quad \forall \varphi \in V.$$

In order to carry the argument from the proof of Theorem 4.1 over to this situation, we need to derive a bound for any (existing) solution $u \in V$ and the V-ellipticity of the bilinear form

$$a(u, \varphi) := \nu(\nabla u, \nabla\varphi) + (\hat{u} \cdot \nabla u, \varphi) + (v^{\text{in}} \cdot \nabla u, \varphi) + (u \cdot \nabla v^{\text{in}}, \varphi)$$

for any fixed $\hat{u} \in V$. For both purposes, we take $\varphi = u$ to obtain in the latter case

$$\begin{aligned}
a(u, u) &= \nu\|\nabla u\|^2 + (u \cdot \nabla v^{\text{in}}, u) = \nu\|\nabla u\|^2 - (u \cdot \nabla u, v^{\text{in}}) \\
&\ge \nu\|\nabla u\|^2 - \|u\|_6\|\nabla u\|\|v^{\text{in}}\|_3 \ge \big(\nu - c_*^2\|v^{\text{in}}\|_3\big)\|\nabla u\|^2.
\end{aligned}$$

Hence, we have the desired V-ellipticity, if the global representation of the boundary data is constructed such that

$$\|v^{\text{in}}\|_3 < \frac{\nu}{c_*^2}.$$

The remaining argument is left as an exercise.

Remark 4.3: It is remarkable that the (stationary) Navier-Stokes problem, despite its nonlinearity, is solvable for all values of the viscosity parameter $\nu > 0$. These solutions do not need to be unique in general, only for sufficiently small data. Actually there are many flow configurations for which there are multiple solutions for the same set of data. An example is the

flow in the gap between two concentric spheres (so-called "Taylor problem"). For rotating inner sphere and fixed outer sphere there appear different flow patterns. For slow rotational speed ("small-data" case) there is a unique solution. Under increasing the rotational speed this base flow turns unstable and other more complex stationary flow patterns occur (so-called "Taylor roles"), which exist simultaneously to the base flow for the same set of (stationary) data. If the rotational speed is further increased beyond some "critical" value these stationary states break down and the flow turns nonstationary (oscillating "Taylor roles") which persist for still stationary data.

The weak solution $\{v, p\} \in (H + v^{\text{in}}) \times L$ of the stationary Navier-Stokes problem obtained by Theorem 4.1 and Lemma 4.1 possesses additional regularity depending on the data of the problem. On smoothly bounded or convex polygonal (polyhedral) domains, we have $v \in H^2(\Omega)^d$ and $p \in H^1(\Omega)$, and there holds the a priori extimate

$$\|\nabla^2 v\| + \|\nabla p\| \leq c_s \left\{\|f\| + \|\nabla^2 v^{\text{in}}\|\right\}. \tag{4.1.18}$$

The constant c_s depends linearly on the viscosity paramter, $c_s \sim 1/\nu$. For the non-trivial proof of this result, we refer to the relevant literature.

Remark 4.4: Finally, we consider the case of general boundary conditions with $\Gamma_{\text{out}} \neq \emptyset$ and ask for the existence of solutions of the corresponding variational formulation (4.1.4) in the function space $H = H_0^1(\Omega, \Gamma_D)$. The proof of the central existance Theorem 4.1 essentially used the identity (4.1.12), which now only holds in the form

$$(u \cdot \nabla v, v) = \tfrac{1}{2}(n \cdot u, |v|^2)_{\Gamma_{\text{out}}}.$$

The boundary integral on the right does not need to be positive, e. g., in the case of dominating inflow through Γ_{out}. On the basis of this relation the proof of Theorem 4.1 only works for sufficiently small data. Whether this is just a weakness of the argument of proof or really an inherent feature of the Navier-Stokes problem is not known. Particularly it is not known whether the trivial solution $v \equiv 0$ and $p \equiv 0$ are the only solutions of the homogeneous problem

$$-\nu\Delta v + v \cdot \nabla v + \nabla p = 0, \quad \nabla \cdot v = 0 \quad \text{in } \Omega,$$

with the boundary conditions

$$v_{|\Gamma_{\text{in}} \cup \Gamma_{\text{rigid}}} = 0, \quad \nu \partial_n v - p n_{|\Gamma_{\text{out}}} = 0.$$

In the case of pure Dirichlet boundary conditions this is clearly the case. Since the "do-nothing" boundary condition is of purely technical nature without solid physical basis and is supposed to model a Poiseuille-like flow pattern (parallel pipe flow), the possible existence of spurious secondary solutions would be very disturbing and would render the use of this artificial outflow boundary condition questionable. A satisfactory answer to this question is still open.

4.1.3 Iterative solution schemes

For solving the variational Navier-Stokes problem (4.1.4), at first, we consider the following functional iteration

$$\nu(\nabla v^t, \nabla\varphi) + (v^{t-1} \cdot \nabla v^t, \varphi) = (f, \varphi) \quad \forall\varphi \in V := J_1, \tag{4.1.19}$$

with a starting value $v^0 \in V$. This reduces the nonlinear Navier-Stokes problem to a sequence of linear Oseen-type problems.

Theorem 4.2: *Let the data of the Navier-Stokes problem satisfy the same smallness assumption* $q := c_*^2\nu^{-2}\|f\|_{-1} < 1$ *as in Theorem 4.1, which guarantees the existence of a unique solution. Then, for any starting value* $v^0 \in V$ *the functional iteration (4.1.19) generates a sequence* $(v^t)_{t\in\mathbb{N}} \subset V$, *which converges to this solution with linear rate:*

$$\|\nabla(v - v^t)\| \leq \frac{q^t}{1-q}\|\nabla(v - v^0)\|, \quad t \geq 1. \tag{4.1.20}$$

Proof: The functional iteration (4.1.19) is well defined, since in virtue of the theorem of Lax-Milgram for $v^{t-1} \in V$ the next iterate $v^t \in V$ exists. The iteration can be viewed as a fixed point iteration $v^t = G(v^{t-1})$. For the fixed point mapping $G(\cdot) : V \to V$, one derives the estimate

$$\|\nabla(G(v) - G(w))\| \leq \frac{c_*^2\|f\|_{-1}}{\nu^2}\|\nabla(v - w)\|, \quad v, w \in V.$$

Hence under the assumption of the theorem the mapping $G(\cdot)$ is a contraction on V and by the Banach fixed point theorem the corresponding fixed point iteration converges to the unique fixed point of $G(\cdot)$, which is the solution of (4.1.4). The details of this argument are posed as an exercise. Q.E.D.

Remark 4.5: Even simpler than the above functional iteration is the fully explicit treatment of the nonlinearity leading to the iteration scheme

$$\nu(\nabla v^t, \nabla\varphi) = (f, \varphi) - (v^{t-1} \cdot \nabla v^{t-1}, \varphi) \quad \forall\varphi \in V, \tag{4.1.21}$$

for an suitable starting value $v^0 \in V$. In each step of this iteration only Stokes problems have to be solved. However, the convergence of this simple iteration requires strong restrictions on the quality of the starting value v^0, which makes this scheme only feasible in the case of very large viscosity when the nonlinear term can be viewed as a small perturbation of the leading Stokes term.

Next, we consider the Newton method in function space for solving problem (4.1.4). The tangent form of the semi-linear form governing the Navier-Stokes problem

$$a(v; \varphi) := \nu(\nabla v, \nabla\varphi) + (v \cdot \nabla v, \varphi)$$

is given by

$$a'(v; \psi, \varphi) := \nu(\nabla\psi, \nabla\varphi) + (v \cdot \nabla\psi, \varphi) + (\psi \cdot \nabla v, \varphi).$$

Starting from some $v^0 \in V$ the Newton iteration formally reads as follows:

$$a'(v^{t-1}; v^t, \varphi) = a'(v^{t-1}; v^{t-1}, \varphi) + (f, \varphi) - a(v^{t-1}; \varphi) \quad \forall \varphi \in V. \tag{4.1.22}$$

In concrete terms this iteration reads:

$$\begin{aligned}\nu(\nabla v^t, \nabla\varphi) + (v^{t-1} \cdot \nabla v^t, \varphi) + (v^t \cdot \nabla v^{t-1}, \varphi) \\ &= \nu(\nabla v^{t-1}, \nabla\varphi) + (v^{t-1} \cdot \nabla v^{t-1}, \varphi) + (v^{t-1} \cdot \nabla v^{t-1}, \varphi) \\ &\quad + (f, \varphi) - \nu(\nabla v^{t-1}, \nabla\varphi) - (v^{t-1} \cdot \nabla v^{t-1}, \varphi) \\ &= (v^{t-1} \cdot \nabla v^{t-1}, \varphi) + (f, \varphi).\end{aligned}$$

In contrast to the functional iteration (4.1.19) here the solvability of the linear subproblems in each iteration step is not so clear, as the governing bilinear form

$$a'(v^{t-1}; \psi, \varphi) = \nu(\nabla\psi, \nabla\varphi) + (v^{t-1} \cdot \nabla\psi, \varphi) + (\psi \cdot \nabla v^{t-1}, \varphi)$$

is generally not V-elliptic,

$$\begin{aligned}a'(v^{t-1}; \varphi, \varphi) &= \nu(\nabla\varphi, \nabla\varphi) + (v^{t-1} \cdot \nabla\varphi, \varphi) + (\varphi \cdot \nabla v^{t-1}, \varphi) \\ &= \nu\|\nabla\varphi\|^2 + ((\nabla v^{t-1})^T\varphi, \varphi).\end{aligned}$$

The matrix $(\nabla v^{t-1})^T$ is not positive definite since its trace $\mathrm{tr} = \sum_{i=1}^d \partial_i v_i^{t-1} = \nabla \cdot v^{t-1} = 0$, i. e., there are positive as well as negative eigenvalues. Therefore, the existence of the Newton sequence $(v^t)_{t\in\mathbb{N}}$ generally requires some smallness assumption on the data of the problem. However, the linear subproblems in the Newton steps may be solvable even if the governing bilinear forms $a'(v^{t-1}; \cdot, \cdot)$ are not V-elliptic but regular in a more general sense.

Theorem 4.3: *Let the data of the Navier-Stokes problem satisfy the same smallness assumption $q := c_*^2\nu^{-2}\|f\|_{-1} < 1$ as in Theorem 4.1, which guarantees the existence of a unique solution. Then, for any starting value $v^0 \in V$ satisfying*

$$\|\nabla(v - v^0)\| \le \frac{1-q}{4\nu^{-2}c_*^2} =: \rho < 1,$$

the Newton iteration (4.1.22) generates a sequence $(v^t)_{t\in\mathbb{N}}$, which converges to this solution with quadratic rate:

$$\|\nabla(v - v^t)\| \le \rho^{2^t}, \quad t \ge 1. \tag{4.1.23}$$

Remark 4.6: A sufficiently good starting value v^0 for the Newton iteration may be generated by a finite number of steps of the functional iteration (4.1.19), which under the same conditions on the data converges with linear rate for any starting value $v^0 \in V$.

Proof: (i) By assumption, the data of the problem satisfy $q := c_*^2\nu^{-2}\|f\|_{-1} < 1$, such that there is a unique solution $v \in V$. This solution admits the estimate

$$\nu\|\nabla v\|^2 = (f, v) - (v \cdot \nabla v, v) \le \|f\|_{-1}\|\nabla v\|$$

and, consequently, $\|\nabla v\| \le \nu^{-1}\|f\|_{-1}$. The tangent form of $a(\cdot;\cdot)$ is uniformly Lipschitz continuous in the following sense:

$$\begin{aligned}|a'(v;\psi,\varphi) - a'(w;\psi,\varphi)| &= |\nu(\nabla\psi,\nabla\varphi) + (v\cdot\nabla\psi,\varphi) + (\psi\cdot\nabla v,\varphi)\\ &\quad - \nu(\nabla\psi,\nabla\varphi) - (w\cdot\nabla\psi,\varphi) - (\psi\cdot\nabla w,\varphi)|\\ &= |((v-w)\cdot\nabla\psi,\varphi) + (\psi\cdot\nabla(v-w),\varphi)|\\ &\le \|v-w\|_3\|\nabla\psi\|\|\varphi\|_6 + \|\psi\|_3\|\nabla(v-w)\|\|\varphi\|_6\\ &\le 2c_*^2\|\nabla(v-w)\|\|\nabla\psi\|\nabla\varphi\|.\end{aligned}$$

In virtue of the above assumptions there holds

$$\begin{aligned}\|\nabla v^0\| &\le \|\nabla(v^0 - v)\| + \|\nabla v\| \le \frac{1 - c_*^2\nu^{-2}\|f\|_{-1}}{4\nu^{-1}c_*^2} + \nu^{-1}\|f\|_{-1}\\ &= \frac{1 - c_*^2\nu^{-2}\|f\|_{-1} + 4c_*^2\nu^{-2}\|f\|_{-1}}{4\nu^{-1}c_*^2} = \frac{1 + 3c_*^2\nu^{-2}\|f\|_{-1}}{4\nu^{-1}c_*^2} < \frac{\nu}{c_*^2}.\end{aligned}$$

Then, for $\varphi \in V$, there holds

$$\begin{aligned}a'(v^0;\varphi,\varphi) &= \nu\|\nabla\varphi\|^2 + (\varphi\cdot\nabla v^0,\varphi)\\ &\ge \nu\|\nabla\varphi\|^2 - \|\varphi\|_6\|\nabla v^0\|\|\varphi\|_3\\ &\ge \big(\nu - c_*^2\|\nabla v^0\|\big)\|\nabla\varphi\|^2 \ge \alpha\|\nabla\varphi\|^2, \quad \alpha > 0,\end{aligned}$$

i. e., we have V-ellipticity such that the first iterate $v^1 \in V$ is well defined. Starting from this initial result, we prove the assertion by induction.

(ii) Suppose now that for some $t \ge 1$ the iterate $v^{t-1} \in V$ exists and satisfies

$$\|\nabla(v - v^{t-1})\| \le \rho := \frac{1-q}{4\nu^{-1}c_*^2} = \frac{1 - c_*^2\nu^{-2}\|f\|_{-1}}{4\nu^{-1}c_*^2}.$$

Then, as in the case $t = 1$, we conclude that

$$\|\nabla v^{t-1}\| \le \frac{\nu}{c_*^2},$$

and from this the V-ellipticity of $a'(v^{t-1};\cdot,\cdot)$, which ensures that the next iterate $v^t \in V$ is well defined. We have to show that $\|\nabla(v - v^t)\| \le \rho$. Then, by induction the whole sequence $(v^t)_{t\ge 0}$ of Newton iterates exists in V.

(iii) Using the equations determining the solution v and the Newton iterate v^t, we find:

$$\begin{aligned}
a'(v; v-v^t, \varphi) &= a'(v; v-v^t, \varphi) - a'(v^{t-1}; v, \varphi) + a'(v^{t-1}; v, \varphi) \\
&= a'(v; v-v^t, \varphi) - a'(v^{t-1}; v-v^t, \varphi) - a'(v^{t-1}; v^t, \varphi) + a'(v^{t-1}; v, \varphi) \\
&= a'(v; v-v^t, \varphi) - a'(v^{t-1}; v-v^t, \varphi) - a'(v^{t-1}; v^{t-1}, \varphi) \\
&\quad - (f, \varphi) + a(v^{t-1}; \varphi) + a'(v^{t-1}; v, \varphi) \\
&= a'(v; v-v^t, \varphi) - a'(v^{t-1}; v-v^t, \varphi) \\
&\quad - a(v; \varphi) + a(v^{t-1}; \varphi) + a'(v^{t-1}; v-v^{t-1}, \varphi) \\
&= a'(v; v-v^t, \varphi) - a'(v^{t-1}; v-v^t, \varphi) \\
&\quad - \int_0^1 \big\{ a'(v^{t-1} + s(v-v^{t-1}); v-v^{t-1}, \varphi) - a'(v^{t-1}; v-v^{t-1}, \varphi) \big\}\, ds,
\end{aligned}$$

Consequently, by the above Lipschitz continuity estimate,

$$|a'(v; v-v^t, \varphi)| \le 2c_*^2 \big(\|\nabla(v-v^{t-1})\| \|\nabla(v-v^t)\| + \|\nabla(v-v^{t-1})\|^2 \big) \|\nabla\varphi\|.$$

Now, taking $\varphi = v - v^t$ in the relation

$$a'(v; v-v^t, \varphi) = \nu(\nabla(v-v^t), \nabla\varphi) + (v \cdot \nabla(v-v^t), \varphi) + ((v-v^t) \cdot \nabla v, \varphi)$$

gives us

$$\begin{aligned}
a'(v; v-v^t, v-v^t) &\ge \nu \|\nabla(v-v^t)\|^2 - |((v-v^t) \cdot \nabla v, v-v^t)| \\
&\ge \nu \|\nabla(v-v^t)\|^2 - \|v-v^t\|_3 \|\nabla v\| \|v-v^t\|_6 \\
&\ge \nu \|\nabla(v-v^t)\|^2 - c_*^2 \|\nabla v\| \|\nabla(v-v^t)\|^2 \\
&\ge \big(\nu - c_*^2 \|\nabla v\| \big) \|\nabla(v-v^t)\|^2.
\end{aligned}$$

Combining the foregoing estimates and observing $\|\nabla v\| \le \nu^{-1} \|f\|_{-1}$ and $c_*^2 \nu^{-2} \|f\|_{-1} < 1$, we conclude

$$\|\nabla(v-v^t)\| \le \frac{2\nu^{-1} c_*^2}{1 - c_*^2 \nu^{-2} \|f\|_{-1}} \big(\|\nabla(v-v^{t-1})\| \|\nabla(v-v^t)\| + \|\nabla(v-v^{t-1})\|^2 \big)$$

To simplify this relation, we set $a_t := \|\nabla(v-v^t)\|$ and use the constant $\rho^{-1} = 4\nu^{-1} c_*^2 / (1 - c_*^2 \nu^{-2} \|f\|_{-1})$ from above to obtain

$$a_t \le \tfrac{1}{2} \rho^{-1} (a_t a_{t-1} + a_{t-1}^2).$$

By assumption, there holds $a_{t-1} \le \rho$ and therefore,

$$a_t \le \tfrac{1}{2} \rho^{-1} (a_t a_{t-1} + a_{t-1}^2) \le \tfrac{1}{2} a_t + \tfrac{1}{2} a_{t-1}^2,$$

and, consequently, $a_t \le a_{t-1}^2 \le \rho^2 < \rho$. Finally iterating this inequality, we obtain that

$$a_t \le \rho^{2^t}, \quad t \ge 1,$$

which completes the proof. Q.E.D.

4.2 Finite element discretization

In this section, we present standard finite element methods for solving the stationary Navier-Stokes equations. In contrast to the existence theory developed above the finite element discretization bases on the coupled variational formulation involving velocity v and pressure p as unknowns. The reason is the difficulty in constructing fully conforming finite element subspaces $V_h \subset V$. Let $H_h \subset H$ and $L_h \subset L$ be finite element subspaces on families of structural regular decompositions $\{\mathbb{T}_h\}_{h>0}$, which satisfy the following discrete "inf-sup stability" condition (motivated by its continuous counterpart in Lemma 4.1):

$$\min_{\chi_h \in L_h} \max_{\varphi_h \in H_h} \frac{(\chi_h, \nabla \cdot \varphi_h)}{\|\chi_h\| \|\nabla \varphi_h\|} \geq \beta_h > 0, \tag{4.2.24}$$

with possibly h-dependent constants $\beta_h > 0$. Then, the approximate problems read as follows: Find $\{v_h, p_h\} \in H_h \times L_h$, such that

$$\nu(\nabla v_h, \nabla \varphi_h) + \tilde{n}(v_h, v_h, \varphi_h) - (p_h, \nabla \cdot \varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in H_h, \tag{4.2.25}$$

$$(\chi_h, \nabla \cdot v_h) = 0 \quad \forall \chi_h \in L_h, \tag{4.2.26}$$

where the modified nonlinear form $\tilde{b}(\cdot,\cdot,\cdot))$ is defined by

$$\tilde{n}(v_h, v_h, \varphi_h) := \tfrac{1}{2} n(v_h, v_h, \varphi_h) - \tfrac{1}{2} n(v_h, \varphi_h, v_h), \quad n(v_h, v_h, \varphi_h) = (v_h \cdot \nabla v_h, \varphi_h)$$

This modification of the nonlinearity on the discrete level is used in order to carry the argument from the proof of Theorem 4.1 over to the discrete level, since there holds

$$\tilde{n}(v_h, \varphi_h, \varphi_h) = 0, \quad v_h,\, \varphi_h \in H_h.$$

Otherwise, we would have to make the assumotion that the mesh size h is taken sufficiently small. This modification is compatible with the continuous level as there holds (exercise)

$$\tilde{n}(v, \psi, \varphi) = \tfrac{1}{2}(v \cdot \nabla \psi, \varphi) + \tfrac{1}{2}(v \cdot \nabla \varphi, \psi) = (v \cdot \nabla \psi, \varphi), \quad v \in V,\ \ \psi, \varphi \in H.$$

4.2.1 General "Stokes elements"

First, we consider the approximation of the (linear) variational Stokes problem: Find $\{v, p\} \in H \times L$, such that

$$(\nabla v, \nabla \varphi) - (p, \nabla \cdot \varphi) = (f, \varphi) \quad \forall \varphi \in H, \tag{4.2.27}$$

$$(\chi, \nabla \cdot v) = 0 \quad \forall \chi \in L. \tag{4.2.28}$$

with normalized viscosity $\nu = 1$ on a convex polygonal (or polyhedral) domain $\Omega \subset \mathbb{R}^d$ ($d = 2, 3$). For simplicity, we assume purely homogeneous Dirichlet boundary conditions, i. e., $\partial\Omega = \Gamma_{\text{rigid}}$ and therefor $H := H_0^1(\Omega)^d$ and $L := L_0^2(\Omega)$. Under these conditions, the (unique) solution of this problem possesses the additional regularity $\{v, p\} \in H^2(\Omega)^d \times H^1(\Omega)$ and there holds the a priori estimate

$$\|\nabla^2 v\| + \|\nabla p\| \leq c_s \|f\|. \tag{4.2.29}$$

The variational Stokes problem has the structure of a saddle point problem and is also called a "mixed" variational formulation. It can be obtained by the Euler-Lagrange approach from the constrained minimization problem

$$\min J(v) := \tfrac{1}{2}\|\nabla v\|^2 - (f,v) \quad \text{on } V \subset H,$$

where the pressure p plays the role of a Lagrange multiplier.

Let $H_h \subset H := H_0^1(\Omega)^d$ and $L_h \subset L := L_0^2(\Omega)$ be finite element subspaces, which satisfy the discrete "inf-sup stability" condition (4.2.24). Then, the approximate problems read as follows: Find $\{v_h, p_h\} \in H_h \times L_h$, such that

$$(\nabla v_h, \nabla \varphi_h) - (p_h, \nabla \cdot \varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in H_h, \tag{4.2.30}$$

$$(\chi_h, \nabla \cdot v_h) = 0 \quad \forall \chi_h \in L_h. \tag{4.2.31}$$

In this general context, we do not make any assumption on the shape- or size-uniformity of the family of decompositions $\{\mathbb{T}_h\}_{h>0}$. The finite element spaces $H_h \times L_h$ are generally called "Stokes elements" and particularly "conforming" because $H_h \subset H$ and $L_h \subset L$. Later, we will also consider so-called "non-conforming" Stokes elements with $H_h \not\subset H$ but still $L_h \subset L$. We further introduce the spaces

$$V_h := \{v_h \in H_h \,|\, (\nabla \cdot v_h, \chi_h) = 0 \ \ \forall \chi_h \in L_h\},$$

of discrete "solenoidal" vector fields. Notice that $V_h \not\subset V$ in general. The discrete Stokes problems (4.2.27)-(4.2.28) are uniquely solvable. This follows from the fact that any solution $\{v_h, p_h\}$ of the corresponding homogeneous problem satisfies

$$\|\nabla v_h\| = 0, \qquad (p_h, \nabla \cdot \varphi_h) = 0 \ \ \forall \varphi_h \in H_h,$$

which in view of the stability relation (4.2.24) implies that $\{v, p\} = \{0, 0\}$. For the convergence $\{v_h, p_h\} \to \{v, p\}$ is is necessary that the approximating spaces V_h are sufficiently "large" for approximating the space V. To this end, we require the following "minimal" approximation property for the spaces $H_h \times L_h \subset H \times L$:

$$\min_{\varphi_h \in H_h} \|\nabla(v - \varphi_h)\| + \min_{\chi_h \in L_h} \|p - \chi_h\| \le ch\big\{\|\nabla^2 v\| + \|\nabla p\|\big\} \le ch\|f\|. \tag{4.2.32}$$

Lemma 4.2 (Special best-approximation property): *The conforming approximation (4.2.30)-(4.2.31) of the Stokes problem (4.2.27)-(4.2.28) possesses the "(quasi)-best-approximation property" with respect to the spaces V_h:*

$$\|\nabla(v - v_h)\| \le c\big\{ \min_{\varphi_h \in V_h} \|\nabla(v - \varphi_h)\| + \min_{\chi_h \in L_h} \|p - \chi_h\|\big\}. \tag{4.2.33}$$

Proof: Subtracting the discrete Stokes equations from their continuous counterparts, we obtain the following Galerkin orthogonality relations:

$$(\nabla(v - v_h), \nabla \varphi_h) - (p - p_h, \nabla \cdot \varphi_h) = 0, \quad \varphi_h \in H_h. \tag{4.2.34}$$

$$(\chi_h, \nabla \cdot (v - v_h)) = 0, \quad \chi_h \in L_h. \tag{4.2.35}$$

Since $v_h \in V_h$ it follows that for arbitrary $\varphi_h \in V_h$ and $\chi_h \in L_h$ there holds

$$\begin{aligned}\|\nabla(v - v_h)\|^2 &= (\nabla(v - v_h), \nabla(v - \varphi_h) + (\nabla(v - v_h), \nabla(\varphi_h - v_h)) \\ &= (\nabla(v - v_h), \nabla(v - \varphi_h) + (p - p_h, \nabla \cdot (\varphi_h - v_h)) \\ &= (\nabla(v - v_h), \nabla(v - \varphi_h) + (p - \chi_h, \nabla \cdot (\varphi_h - v)) + (p - \chi_h, \nabla \cdot (v - v_h)).\end{aligned}$$

Using the Young inequality $ab \le a^2 + \frac{1}{4}b^2$, we conclude that

$$\|\nabla(v - v_h)\|^2 \le c\big\{\|\nabla(v - \varphi_h\|^2 + \|p - \chi_h\|^2\big\},$$

which implies the assertion. Q.E.D.

The estimate (4.2.33) is unsatisfactory since it is not clear how to construct proper approximations $i_h v \in V_h$ for $v \in V$. This, however, can be accomplished if the pairs $H_h \times L_h \subset H \times L$ for $h > 0$ are *uniformly* "inf-sup stable":

$$\min_{\chi_h \in L_h} \max_{\varphi_h \in H_h} \frac{(\chi_h, \nabla \cdot \varphi_h)}{\|\chi_h\| \|\nabla \varphi_h\|} \ge \beta_*, \quad h > 0, \tag{4.2.36}$$

with a fixed constant $\beta_* > 0$. This property requires the spaces H_h to be sufficiently "large" (higher dimensional) compared to L_h.

Lemma 4.3: *If the pairs $H_h \times L_h \subset H \times L$, $h > 0$, satisfy the uniform "inf-sup" condition (4.2.36), then there also holds the "adjoint" inequality*

$$\min_{\varphi_h \in H_h} \max_{\chi_h \in L_h} \frac{(\chi_h, \nabla \cdot \varphi_h)}{\|\chi_h\| \|\nabla \varphi_h\|} \ge \beta_*. \tag{4.2.37}$$

Further, for $v \in V$ there holds:

$$\min_{\varphi_h \in V_h} \|\nabla(v - \varphi_h)\| \le (1 + c\beta_*^{-1}) \min_{\varphi_h \in H_h} \|\nabla(v - \varphi_h)\|. \tag{4.2.38}$$

Proof: (i) First, we prove (4.2.37). The relation

$$(B_h \varphi_h)(\chi_h) := (\chi_h, \nabla \cdot \varphi_h) \quad \forall \chi_h \in L_h,$$

defines a linear operator $B_h : H_h \to L_h^*$. The corresponding adjoint operator $B_h^* : L_h \to H_h^*$ is defined by

$$(B_h^* \chi_h)(\varphi_h) := (B_h \varphi_h)(\chi_h) = (\chi_h, \nabla \cdot \varphi_h) \quad \forall \varphi_h \in H.$$

Since the spaces H_h and L_h are finite dimensional, we have the following relations:

$$\begin{aligned} R(B_h) &= N(B_h^*)^\perp, \quad R(B_h)^\perp = N(B_h^*), \\ R(B_h^*) &= N(B_h)^\perp, \quad R(B_h^*)^\perp = N(B_h), \end{aligned}$$

where $R(\cdot) = \text{range}(\cdot)$ and $N(\cdot) = \text{kernel}(\cdot)$. The stability inequality (4.2.36) implies that

$$\|B_h^* \chi_h\|_{H^*} = \sup_{\varphi_h \in H_h} \frac{(B_h^* \chi_h)(\varphi_h)}{\|\nabla \varphi_h\|} = \sup_{\varphi_h \in H_h} \frac{(\chi_h, \nabla \cdot \varphi_h)}{\|\nabla \varphi_h\|} \ge \beta_* \|\chi_h\|, \quad \chi_h \in L_h.$$

Hence the operator B_h^* is injective, i. e., $R(B_h)^\perp = N(B_h^*) = \{0\}$ with bounded inverse B_h^{*-1} :

$N(B_h)^\perp \subset H^* \to L_h$:

$$\|B_h^{*-1}\| \le \beta_*^{-1}.$$

Then, the adjoint $B_h : H \to N(B_h^*)^\perp \subset L_h^*$ is also invertible with the same bound,

$$\|B_h^{-1}\| \le \beta_*^{-1}.$$

This in turn implies

$$\beta_*\|\nabla\varphi_h\| \le \|B_h\varphi_h\|_{L_h^*} = \sup_{\chi_h\in L_h} \frac{(B_h\varphi_h)(\chi_h)}{\|\chi_h\|} = \sup_{\chi_h\in L_h} \frac{(\chi_h, \nabla\cdot\varphi_h)}{\|\chi_h\|},$$

what was to be shown.

(ii) It remains to show (4.2.38). Let $v \in V$, $v_h \in V_h$, and $\varphi_h \in H_h$ arbitrary. Then, in virtue of (4.2.37) there holds

$$\begin{aligned}
\|\nabla(v - v_h)\| &\le \|\nabla(v - \varphi_h)\| + \|\nabla(\varphi_h - v_h)\| \\
&\le \|\nabla(v - \varphi_h)\| + \beta_*^{-1} \sup_{\chi_h\in L_h} \frac{(\chi_h, \nabla\cdot(\varphi_h - v_h))}{\|\chi_h\|} \\
&= \|\nabla(v - \varphi_h)\| + \beta_*^{-1} \sup_{\chi_h\in L_h} \frac{(\chi_h, \nabla\cdot(\varphi_h - v))}{\|\chi_h\|} \\
&\le (1 + c\beta_*^{-1})\|\nabla(v - \varphi_h)\|.
\end{aligned}$$

Because of the arbitrary choice of $v_h \in V_h$ and $\varphi_h \in H_h$ this implies (4.2.38). Q.E.D.

Theorem 4.4 (General best approximation property): *The conforming approximation (4.2.30)-(4.2.31) of the Stokes problem (4.2.27)-(4.2.28) possesses in case of uniform "inf-sup" stability (4.2.36) the following "(quasi)-best approximation" property:*

$$\|\nabla(v - v_h)\| + \|p - p_h\| \le c\Big\{ \min_{\varphi_h\in H_h} \|\nabla(v - \varphi_h)\| + \min_{\chi_h\in L_h} \|p - \chi_h\|\Big\}, \tag{4.2.39}$$

$$\|v - v_h\| \le ch\Big\{ \min_{\varphi_h\in H_h} \|\nabla(v - \varphi_h)\| + \min_{\chi_h\in L_h} \|p - \chi_h\|\Big\}. \tag{4.2.40}$$

Together with the approximation property (4.2.32) this implies the following error estimates:

$$\|\nabla(v - v_h)\| + \|p - p_h\| \le ch\|f\|, \tag{4.2.41}$$

$$\|v - v_h\| \le ch^2\|f\|. \tag{4.2.42}$$

Proof: Subtracting the discrete Stokes equations and their continuous counterpart, we again obtain the Galerkin orthogonality relations

$$(\nabla(v - v_h), \nabla\varphi_h) - (p - p_h, \nabla\cdot\varphi_h) = 0, \quad \varphi_h \in H_h. \tag{4.2.43}$$

$$(\chi_h, \nabla\cdot(v - v_h)) = 0, \quad \chi_h \in L_h. \tag{4.2.44}$$

(i) We begin with the estimate of $\|\nabla(v - v_h)\|$. For this, we will not use the result of Lemma 4.3. We rather prepare for the argument used below in the context of so-called "stabilized"

Stokes elements. For arbitrary $\varphi_h \in H_h$ we obtain with help of (4.2.43) that

$$\begin{aligned}\|\nabla(v-v_h)\|^2 &= (\nabla(v-v_h),\nabla(v-\varphi_h)+(\nabla(v-v_h),\nabla(\varphi_h-v_h))\\ &= (\nabla(v-v_h),\nabla(v-\varphi_h)+(p-p_h,\nabla\cdot(\varphi_h-v_h))\\ &= (\nabla(v-v_h),\nabla(v-\varphi_h)+(p-p_h,\nabla\cdot(\varphi_h-v))+(p-p_h,\nabla\cdot(v-v_h)),\end{aligned}$$

and further with help of (4.2.44) with arbitrary $\chi_h \in L_h$

$$\|\nabla(v-v_h)\|^2 = (\nabla(v-v_h),\nabla(v-\varphi_h)+(p-p_h,\nabla\cdot(\varphi_h-v))+(p-\chi_h,\nabla\cdot(v-v_h)).$$

Using Young's inequality $ab \le \varepsilon^2 a^2 + (4\varepsilon^2)^{-1}b^2$, we conclude

$$\|\nabla(v-v_h)\| \le c\big\{(1+\varepsilon^{-1})\|\nabla(v-\varphi_h)\| + \|p-\chi_h\|\big\} + \varepsilon\|p-p_h\|. \tag{4.2.45}$$

(ii) Next, we estimate $\|p-p_h\|$. With help of the "inf-sup stability" (4.2.24), we get

$$\begin{aligned}\|p-p_h\| &\le \|p-\chi_h\| + \|\chi_h-p_h\|\\ &\le \|p-\chi_h\| + \beta_*^{-1}\sup_{\psi_h\in H_h}\frac{(\chi_h-p_h,\nabla\cdot\psi_h)}{\|\nabla\psi_h\|}\\ &\le \|p-\chi_h\| + \beta_*^{-1}\sup_{\psi_h\in H_h}\frac{(\chi_h-p,\nabla\cdot\psi_h)}{\|\nabla\psi_h\|} + \beta_*^{-1}\sup_{\psi_h\in H_h}\frac{(p-p_h,\nabla\cdot\psi_h)}{\|\nabla\psi_h\|}\\ &\le c(1+\beta_*^{-1})\|p-\chi_h\| + \beta_*^{-1}\sup_{\psi_h\in H_h}\frac{(p-p_h,\nabla\cdot\psi_h)}{\|\nabla\psi_h\|}.\end{aligned}$$

Further, the Galerkin orthogonality relation (4.2.44) implies

$$\begin{aligned}\|p-p_h\| &\le c(1+\beta_*^{-1})\|p-\chi_h\| + \beta_*^{-1}\sup_{\psi_h\in H_h}\frac{(\nabla(v-v_h),\nabla\psi_h)}{\|\nabla\psi_h\|}\\ &\le c(1+\beta_*^{-1})\|p-\chi_h\| + c\|\nabla(v-v_h)\|.\end{aligned}$$

Combining this estimate with (4.2.45) gives us

$$\begin{aligned}\|\nabla(v-v_h)\| \le c\big\{&\|\nabla(v-\varphi_h)\| + \|p-\chi_h\|\big\}\\ &+\varepsilon\big\{c(1+\beta_*^{-1})\|p-\chi_h\| + c\|\nabla(v-v_h)\|\big\}.\end{aligned}$$

By sufficiently small choice of $\varepsilon > 0$,

$$\|\nabla(v-v_h)\| \le c\big\{\|\nabla(v-\varphi_h)\| + (1+\beta_*^{-1})\|p-\chi_h\|\big\}.$$

Since $\varphi_h \in H_h$ and $\chi_h \in L_h$ are arbitrary, we obtain the first estimate (4.2.39).

(iii) For estimating $\|v-v_h\|$, we again use a duality argument. Let $\{z,q\} \in H \times L$ be the solution of the auxiliary Stokes proble

$$(\nabla\varphi,\nabla z) - (q,\nabla\cdot\varphi) = (\varphi, v-v_h)\|v-v_h\|^{-1} \quad \forall\varphi\in H, \tag{4.2.46}$$

$$(\chi,\nabla\cdot z) = 0 \quad \forall\chi\in L. \tag{4.2.47}$$

That is in $H^2(\Omega)^d \times H^1(\Omega)$ and satisfies the a priori estimate

$$\|\nabla^2 z\| + \|\nabla q\| \le c\|v - v_h\|\|v - v_h\|^{-1} = c. \tag{4.2.48}$$

Setting $\varphi := v - v_h$ and using Galerkin orthogonality with the natural nodal interpolation $i_h z \in H_h$ and $j_h q \in L_h$, we find

$$\begin{aligned}\|v - v_h\| &= (\nabla(v - v_h), \nabla z) - (q, \nabla \cdot (v - v_h)) \\ &= (\nabla(v - v_h), \nabla(z - i_h z)) + (\nabla(v - v_h), \nabla i_h z) - (q - j_h q, \nabla \cdot (v - v_h)) \\ &= (\nabla(v - v_h), \nabla(z - i_h z)) + (p - p_h, \nabla \cdot i_h z) - (q - j_h q, \nabla \cdot (v - v_h)) \\ &= (\nabla(v - v_h), \nabla(z - i_h z)) + (p - p_h, \nabla \cdot (i_h z - z)) - (q - j_h q, \nabla \cdot (v - v_h)).\end{aligned}$$

Further, we estimate

$$\begin{aligned}\|v - v_h\| &\le \|\nabla(v - v_h)\|\|\nabla(z - i_h z)\| + \|p - p_h\|\|\nabla \cdot (i_h z - z)\| + \|q - j_h q\|\|\nabla \cdot (v - v_h)\| \\ &\le ch\|\nabla(v - v_h)\|\|z\|_{H^2} + ch\|p - p_h\|\|z\|_{H^2} + ch\|q\|_{H^1}\|\nabla(v - v_h)\| \\ &\le ch\big\{\|\nabla(v - v_h)\| + \|p - p_h\|\big\}.\end{aligned}$$

This completes the proof. Q.E.D.

(I) Examples of "conforming" Stokes elements

In the following the various Stokes elements are described by specifying its nodal values for velocity and pressure ansatz. We restrict the presentation to the 2d case. In most cases there are natural analogues in 3D. We use the notation $P_H(T)$ and $P_L(T)$ for the local polynomial ansatz for velocity (H) and pressure (L), respectively, and the superscripts "c" and "dc" for indicating whether the global ansatz is "continuous (c)" or "discontinuous (dc)". In the following the mesh families $\{\mathbb{T}_h\}_{h>0}$ are assumed to be "shape uniform". We will use the following technical result on local H^1-stable interpolation in 2D and 3D.

Lemma 4.4: *There exist generalized interpolation operators $I_h^{(1)} : H \to H_h^{(1)}$ into the space $H_h^{(1)} \subset H$ of conforming cellwise linear or bi/tri-linear finite elements on triangular/tedrahedral or quadrilateral/hexahedral meshes, such that for $v \in H$ there holds*

$$\|v - I_h^{(1)} v\|_T + h_T\|\nabla I_h^{(1)} v\|_T \le ch_T\|v\|_{H^1(\tilde{T})}, \quad T \in \mathbb{T}_h, \tag{4.2.49}$$

where $\tilde{T} := \cup\{T' \in \mathbb{T}_h \,|\, T' \cap T \neq \emptyset\}$.

Proof: The proof uses cell or edge averages of v in the interpolation rather than (not H^1-stable) point values. For the very technical details, we refer to the literature (e. g., Girault/Raviart [27] or Brenner/Scott [34]). Q.E.D.

a) Stokes elements with discontinuous pressure:

(i) The triangular $P_1^{\rm c}/P_0^{\rm dc}$ element (a) and the quadrilateral $Q_1^{\rm c}/P_0^{\rm dc}$ element (b):

a) $P_H(T) := P_1(T)^2$. $P_L(T) := P_0(T)$;

b) $P_H(T) := Q_1(T)^2$, $P_L(T) := P_0(T)$;

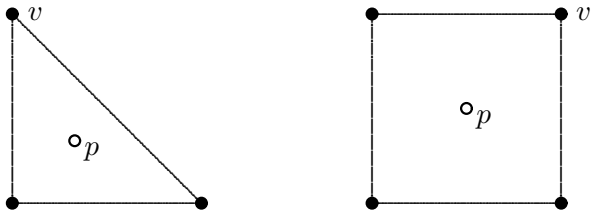


Figure 4.1: The conforming Stokes elements of type P_1^c/P_0^{dc} (left) and Q_1^c/P_0^{dc} (right).

The simplest P_1^c/P_0^{dc} element is not useful, as here the subspace V_h of "discrete solenoidal" velocities may be trivial. The Q_1^c/P_0^{dc} element is very popular in engineering because of its simplicity but turns out to be unstable in general, which may result in unphysical oscillations in the pressure approximations, so-calle "checkerboard" modes (exercise).

(ii) The triangular P_2^c/P_0^{dc} element (a), the extended triangular $\tilde{P}_2^c/P_1^{dc}$ element (b), and the quadrilateral Q_2^c/P_1^{dc} element (c):

a) $P_H(T) := P_2(T)^2, \quad P_L(T) := P_0(T)\,;$

b) $P_H(T) := \tilde{P}_2(T)^2 := P_2(T)^2 \oplus \mathrm{span}\{b_T^1, b_T^2\}, \quad P_L(T) := P_1(T)\,;$

c) $P_H(T) := Q_2(T)^2, \quad P_L(T) := P_1(T)\,.$

Here $b_T^1 = (b_T, 0)^T$ and $b_T^2 = (0, b_T)^T$ with the (unique) cubic "bulb function"

$$b_T \in P_3(T): \quad b_T|_{\partial K} = 0, \;\; |T|^{-1} \int_T b_T \, dx = 1.$$

Such a "bulb function" can be obtained by the following explicit construction: Let the three sides of a cell T be given by the linear equations $l_i(x) = a_i x_1 + b_i x_2 + c_i = 0$, $i = 1, 2, 3$. Then the cubic polynomial $b_T(x) := \gamma_T l_1(x) l_2(x) l_3(x)$ with $\gamma_T = |T|(\int_T l_1 l_2 l_3 \, dx)^{-1}$ has all the required properties.

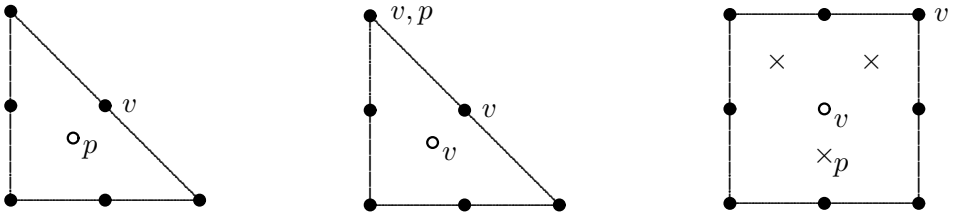


Figure 4.2: The conforming Stokes elements of type P_2^c/P_0^{dc} (left), $\tilde{P}_2^c/P_1^{dc}$ (middle) and Q_2^c/P_1^{dc} (right).

These Stokes elements satisfy the uniform "inf-sup" stability condition. Actually the proof of the stability motivates the use of the extended velocity ansatz $\tilde{P}_2$. This construction can be

generalized for higher polynomial degrees $r \geq 3$. In these discretizations mass conservation is realized at least locally on each cell,

$$\int_{\partial T} n \cdot v_h \, do = \int_T \nabla \cdot v_h \, dx = 0, \quad T \in \mathbb{T}_h. \tag{4.2.50}$$

Lemma 4.5 ("inf-sup" stability): *The conforming Stokes- elements of type* P_2^c/P_0^{dc}, $\tilde{P}_2^c/P_1^{dc}$ *and* Q_2^c/P_1^{dc} *are uniformly "inf-sup" stable.*

Proof: i) The proof bases on the continuous "inf-sup" stability estimate (4.1.17) and the local properties of the finite element functions used. The single steps of the argument can give hints for the construction of stable Stokes elements Let $q_h \in L_h$ be arbitrarily given. In virtue of the continuous stability estimate there holds

$$\beta \|q_h\| \leq \sup_{\varphi \in H} \frac{(q_h, \nabla \cdot \varphi)}{\|\nabla \varphi\|}. \tag{4.2.51}$$

This implies the relation

$$\beta \|q_h\| \leq \sup_{\varphi \in H} \left\{ \frac{(q_h, \nabla \cdot \varphi_h)}{\|\nabla \varphi_h\|} \frac{\|\nabla \varphi_h\|}{\|\nabla \varphi\|} \right\} + \sup_{\varphi \in H} \frac{(q_h, \nabla \cdot (\varphi - \varphi_h))}{\|\nabla \varphi\|}, \tag{4.2.52}$$

for arbitrary $\varphi_h \in H_h$. The way to the asserted discrete stability estimate is then the construction of an interpolation operator $\pi_h : H \to H_h$ with the following properties:

$$(a) \qquad (\chi_h, \nabla \cdot \pi_h v) = (\chi_h, \nabla \cdot v) \quad \forall \chi_h \in L_h, \tag{4.2.53}$$

$$(b) \qquad \|\nabla \pi_h v\| \leq c_1 \|\nabla v\|. \tag{4.2.54}$$

In view of (4.2.52) this implies

$$\beta \|q_h\| \leq \sup_{\varphi \in H} \left\{ \frac{(q_h, \nabla \cdot \pi_h \varphi)}{\|\nabla \pi_h \varphi\|} \frac{\|\nabla \pi_h \varphi\|}{\|\nabla \varphi\|} \right\} \leq c_1 \sup_{\varphi_h \in H_h} \frac{(q_h, \nabla \cdot \varphi_h)}{\|\nabla \varphi_h\|},$$

Hence, the discrete "inf-sup" stability estimate is satisfied with $\beta_* := \beta / c_1$.

(ii) The construction of the operator $\pi_h : H \to H_h$ is oriented by the critical properties (a) and (b). By integration by parts there holds

$$(q_h, \nabla \cdot v) = \sum_{T \in |T_h} (q_h, \nabla \cdot v)_T = \sum_{T \in \mathbb{T}_h} \left\{ (q_h, n \cdot v)_{\partial_T} - (\nabla q_h, v)_T \right\}, \quad q_h \in L_h.$$

A local construction of r_h has to satisfy the following conditions:

$$(q_h, n \cdot v)_\Gamma = (q_h, n \cdot \pi_h v)_\Gamma, \quad \Gamma \in \partial \mathbb{T}_h, \tag{4.2.55}$$

$$(\nabla q_h, v)_T = (\nabla q_h, \pi_h v)_T, \quad T \in \mathbb{T}_h. \tag{4.2.56}$$

Here and below, we use the notation $\partial \mathbb{T}_h$ and $\partial^2 \mathbb{T}_h$ for the set of edges and the set of vertices, respectively, of the mesh $\mathbb{T}_h$. In each concrete case the above conditions are to be supplemented by additional ones to generate an appropriate operator π_h. This involves also the proof of the stability property (b). We will sketch this argument for the simple Stokes elements considered.

(iii) The P_2^c/P_0^{dc} element: The corresponding operator π_h is defined by the following local conditions (observing $\nabla q_h|_T \equiv 0$):

$$\pi_h v(a) = I_h^{(1)} v(a), \quad a \in \partial^2 \mathbb{T}_h,$$
$$(\chi, \pi_h v)_\Gamma = (\chi, v)_\Gamma, \quad \chi \in P_0(\Gamma)^2, \ \Gamma \in \partial \mathbb{T}_h.$$

The set of 6 nodal functionals (per component) used in this definition is unisolvent with $P_2(T)$. By construction, there holds

$$(q_h, n \cdot v)_\Gamma = (q_h, n \cdot \pi_h v)_\Gamma, \quad \Gamma \in \partial \mathbb{T}_h.$$

For establishing the stability of π_h, we split $\pi_h v|_T$ into its cellwise linear part $I_h^{(1)} v|_T \in P_1(T)$ and a cellwise quadratic part $Q_h^{(2)} v|_T \in P_2(T)$, such that $\pi_h v = I_h^{(1)} v + Q_h^{(2)} v$, where

$$Q_h^{(2)} v(a) = 0, \quad a \in \partial^2 \mathbb{T}_h, \qquad \int_\Gamma Q_h^{(2)} v \, ds = \int_\Gamma (v - I_h^{(1)} v) \, ds, \quad \Gamma \in \partial \mathbb{T}_h.$$

Splitting $Q_h^{(2)} v$ into the contributions from the remaining basis functions corresponding to the edges Γ and employing the local trace inequality on the cells T, we deduce that (exercise)

$$\|\nabla Q_h^{(2)} v\|_T \le c \|\nabla (v - I_h^{(1)} v)\|_T.$$

From this, we conclude the desired bound

$$\|\nabla \pi_h v\| \le \|\nabla I_h^{(1)} v\| + \|\nabla Q_h^{(2)} v\| \le \|\nabla I_h^{(1)} v\| + c \|\nabla (v - I_h^{(1)} v)\| \le c \|\nabla v\|.$$

(iv) For the proof of the uniform "inf-sup" stability of the $\tilde{P}_2^c/P_1^{dc}$ element and the Q_2^c/P_1^{dc} element, we refer to the literature (Girault/Raviart [27]). Q.E.D.

Remark 4.7: The existence of a (linear) operator $\pi_h : H \to H_h$ with the properties (4.2.53)-(4.2.54) is also necessary for the *uniform* discrete "inf-sup" stability. To see this, let the stability be given. Then, by Lemma 4.3 there also holds

$$\inf_{v_h \in H_h} \sup_{\chi_h \in L_h} \frac{(\chi_h, \nabla \cdot v_h)}{\|\nabla v_h\| \|\chi_h\|} \ge \beta_*.$$

For any $v \in H$ this implies the existence of a unique element $\pi_h v \in V_h^T$, such that

$$(\chi_h, \nabla \cdot \pi_h v) = (\chi_h, \nabla \cdot v) \quad \forall \chi_h \in L_h,$$

and

$$\beta_* \|\nabla \pi_h v\| \le \sup_{\chi_h \in L_h} \frac{(\chi_h, \nabla \cdot \pi_h v)}{\|\chi_h\|} \le \sup_{\chi_h \in L_h} \frac{(\chi_h, \nabla \cdot v)}{\|\chi_h\|} \le c \|\nabla v\|.$$

This obviously defines as linear operator $\pi_h : H \to H_h$ with the desired properties. We emphasize that the explicit construction of such an operator π_h is generally more complicated than in the above examples.

b) Stokes elements with continuous pressure:

These elements are of relatively low dimension but do not possess the local mass conservation

property.

(i) The "$\tilde{P}_1^c/P_1^c$ (MINI) element" (a), the "P_2^c/P_1^c (Taylor-Hood) element" (b), and the "Q_2^c/Q_1^c (isoparametric) element" (c):

a) $P_H(T) := \tilde{P}_1(T) := P_1(T)^2 \oplus \mathrm{span}\{b_T^1, b_T^2\}, \quad P_L(T) := P_1(T)\,;$

b) $P_H(T) := P_2(T)^2, \quad P_L := P_1(T)\,;$

c) $P_H(T) := Q_2(T)^2, \quad P_L := Q_1(T)\,.$

Here, again $b_T^1 = (b_T, 0)^T$ and $b_T^2 = (0, b_T)^T$ with the cubic "bulb functions" b_T.

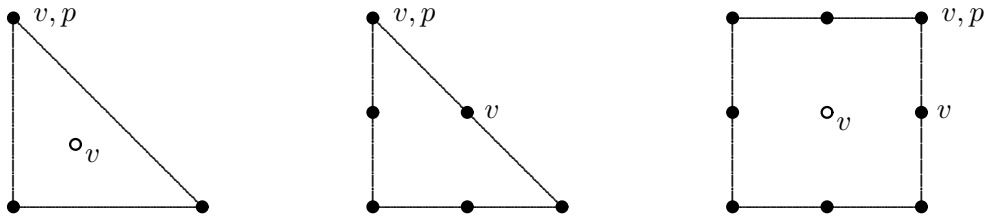


Figure 4.3: The conforming Stokes elements of type $P_1 \oplus \mathrm{span}\{b_K^1, b_K^2\}/P_1^c$ (left), P_2^c/P_1^c (middle) and Q_2^c/Q_1^c (right).

All three Stokes elements are uniformly "inf-sup" stable. In general, the continuous P_r^c/P_{r-2}^c ansatz (for $r \geq 3$) is "inf-sup" stable. The pairs P_r^c/P_{r-1}^c (generalized "Taylor-Hood" elements) is not stable and requires a local enhancement of the velocity ansatz.

Lemma 4.6 ("inf-sup" stability): *The "MINI element" of type $\tilde{P}_1^c/P_1^c$ as well as the "Taylor-Hood" elements of type P_2^c/P_1^c and Q_2^c/Q_1^c are uniformly "inf-sup" stable.*

Proof: The proof uses similar arguments as that of Lemma 4.5. Starting point is again the relation

$$\beta\|q_h\| \leq \sup_{v\in H} \left\{ \frac{(q_h, \nabla\cdot\varphi_h)}{\|\nabla v_h\|} \frac{\|\nabla v_h\|}{\|\nabla v\|} \right\} + \sup_{v\in H} \frac{(q_h, \nabla\cdot(v - v_h))}{\|\nabla v\|}, \tag{4.2.57}$$

for arbitrary $q_h \in L_h$ and $v_h \in H_h$. The goal is again the construction of an interpolation operator $\pi_h : H \to H_h$ with the properties (a) and (b). For the MINI element this can be accomplished similarly as above. Because of the continuity of the pressure ansatz, we have

$$(q_h, \nabla\cdot v) = -(\nabla q_h, v).$$

Hence, for the operator π_h has, besides the H-stability, only the property

$$(\nabla q_h, v) = (\nabla q_h, \pi_h v)_T = 0, \quad T \in \mathbb{T}_h,$$

to be realized. For that, the MINI element provides the two additional bulb components per cell in the velocity ansatz. Accordingly, we define the operator $\pi_h : H \to H_h$ by the conditions

$$\pi_h v(a) = I_h^{(1)} v(a), \;\; a \in \partial^2\mathbb{T}_h, \qquad (\pi_h v, 1)_T = (v, 1)_T, \;\; T \in \mathbb{T}_h.$$

The stability estimate is then obtained by using the splitting $\pi_h v|_T = I_h^{(1)} v|_T + \gamma_T b_T$ with $\gamma_T = |T|^{-1}(v - I_h^{(1)} v, 1)_T$ (to fulfill the mean value condition):

$$\begin{aligned}\|\nabla \pi_h v\|_T &\le \|\nabla I_h^{(1)} v\|_T + |T|^{-1} |(v - I_h^{(1)} v, 1)_T| \|\nabla b_T\|_T \\ &\le \|\nabla I_h^{(1)} v\|_T + |T|^{-1/2} \|v - I_h^{(1)} v\|_T \|\nabla b_T\|_T \\ &\le \|\nabla I_h^{(1)} v\|_T + c h_T^{-1} \|v - I_h^{(1)} v\|_T \\ &\le c \|v\|_{H^1(\tilde{T})},\end{aligned}$$

which yields

$$\|\nabla \pi_h v\| \le c \|v\|_{H^1(\Omega)} \le c \|\nabla v\|.$$

For the trianglar and quadrilateral Taylor-Hood elements the proof of stability requires a more sophisticated argument, for which we refer to the relevant literature. Q.E.D.

(II) Examples of "nonconforming" Stokes elements

We have seen that the stability condition (4.2.36) does not allow the use of the most natural lowest-order ansatz spaces $P_1^c / P_0^{\mathrm{dc}}$ and $Q_1^{\mathrm{c}} / P_0^{\mathrm{dc}}$. To enlarge the velocity space, one may turn to the corresponding "nonconforming", i. e., not fully continuous, finite elements for the velocity. The elements for the pressure is kept discontinuous in order to preserve local mass conservation. In the case of discontinuous velocity elements all spatial derivatives have to be defines cellwise:

$$(\nabla_h v_h)|_T := \nabla (v_h|_T), \quad T \in \mathbb{T}_h.$$

With this notation then a pair $\{v_h, p_h\} \in H_h \times L_h$ is to be determined, such that

$$(\nabla_h v_h, \nabla_h \varphi_h) - (p_h, \nabla_h \cdot \varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in H_h, \tag{4.2.58}$$

$$(\nabla_h \cdot v_h, \chi_h) = 0 \quad \forall \chi_h \in L_h. \tag{4.2.59}$$

For guaranteeing the existence of solutions of these finite dimensional problems and there stability as $h \to 0$, we use a discrete analogue of the Poincaré inequality,

$$\|v_h\| \le \gamma \|\nabla_h v_h\|, \quad v_h \in H_h. \tag{4.2.60}$$

This may be proven for the nonconforming Stokes elements considered by the same argument as used in the proof of Theorem 4.5, below, and is posed as an exercise. Observing that, for any solution $v_h \in H_h$, there holds

$$\begin{aligned}\|\nabla_h v_h\|^2 - (p_h, \nabla_h \cdot v_h) &= (f, v_h), \\ (p_h, \nabla_h \cdot v_h) &= 0,\end{aligned}$$

we obtain

$$\|\nabla_h v_h\| \le \|f\| \|v_h\| \|\nabla_h v_h\|^{-1} \le \gamma \|f\|. \tag{4.2.61}$$

This stability result, together with the "inf-sup" stability estimate for the pressure to be proven, particularly implies uniqueness and by that also existence of solutions.

Examples: The nonconforming $P_1^{\mathrm{nc}} / P_0^{\mathrm{dc}}$ element (a) and the $Q_1^{\mathrm{rot}} / P_0^{\mathrm{dc}}$ elements (b):

a) $P_H(T) := P_1^{\rm nc}(T)^2, \quad P_L(T) := P_0(T)\,;$

b) $P_H(T) := \tilde{Q}_1^{\rm nc}(T)^2, \quad P_L := P_0(T)\,.$

with the "rotated bilinear" ansatz on the square reference cell $\hat{T}$

$$\tilde{Q}_1(\hat{T}) := \mathrm{span}\{1, x_1, x_2, x_1^2 - x_2^2\}.$$

This modification of the usual *bilinear* ansatz $\mathrm{span}\{1, x_1, x_2, x_1x_2\}$ is necessary, since the latter is not "unisolvent" with respect to the edge-oriented nodal values (edge midpoint value or edge mean value). This is seen by considering the polynomial $p(x) := x_1x_2$, which vanishes at the 4 edge midpoints of the reference cell $\hat{T} = (-1,1)^2$. The modified ansatz $\tilde{Q}_1(\hat{T})$ results from $Q_1(\hat{T})$ by a coordinate rotation of 90^o (motivating the name "rotated bilinear element") and is therefore unisolvent. This Stokes element has a natural analogue in 3D:

$$\tilde{Q}_1(T) := \mathrm{span}\{1, x_1, x_2, x_3, x_1^2 - x_2^2, x_2^2 - x_3^2\}.$$

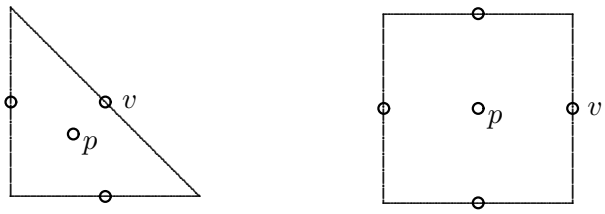


Figure 4.4: The nonconforming Stokes elementss of type $P_1^{\rm nc}/P_0$ (left) and $\tilde{Q}_1^{\rm nc}/P_0$ (right).

For proving the solvability of system (4.2.58)- (4.2.59) in the nonconforming case, we again consider a solution $\{v_h, p_\}\} \in H_h \times L_h$ of the corresponding homogeneous problem, i. e., for $f \equiv 0$. For that there holds

$$\|\nabla_h v_h\| = 0,$$

which implies $\nabla v_h|_T \equiv 0$ on each cell $T \in \mathbb{T}_h$ and consequently, in view of the continuity properties of $v_h \in H_h$ and the boundary condition, necessarily $v_h \equiv 0$.

Lemma 4.7 ("inf-sup" stability): *The nonconforming Stokes elements of type* $P_1^{\rm nc}/P_0^{\rm dc}$ *and* $\tilde{Q}_1^{\rm nc}/P_0^{\rm dc}$ *satisfy the uniform "inf-sup" stability condition:*

$$\inf_{q_h \in L_h} \Big(\sup_{\varphi_h \in H_h} \frac{(q_h, \nabla_h \cdot \varphi_h)}{\|q_h\| \|\nabla_h \varphi_h\|} \Big) \geq \beta_*. \tag{4.2.62}$$

Proof: We construct an interpolation operator $\pi_h : H \to H_h$ with the following properties:

$$\|\nabla_h \pi_h v\| \leq c_1 \|\nabla v\|, \quad v \in H, \tag{4.2.63}$$

$$(q_h, \nabla_h \cdot \pi_h v) = (q_h, \nabla \cdot v), \quad v \in H,\ q_h \in L_h. \tag{4.2.64}$$

With this construction, we can argue as follows:

$$\beta\|q_h\| \le \sup_{\varphi\in H}\frac{(q_h,\nabla\cdot\varphi)}{\|\nabla\varphi\|} = \sup_{\varphi\in H}\left\{\frac{(q_h,\nabla_h\cdot\pi_h\varphi)}{\|\nabla_h\pi_h\varphi_h\|}\frac{\|\nabla_h\pi_h\varphi_h\|}{\|\nabla\varphi\|}\right\} \le c_1\sup_{\varphi_h\in H_h}\frac{(q_h,\nabla_h\cdot\varphi_h)}{\|\nabla\varphi_h\|},$$

what implies the asserted stability estimate (4.2.62) with the constant $\beta_* := \beta c_1^{-1}$.

(i) We begin with the $P_1^{\rm nc}/P_0^{\rm dc}$ element. The reqirement

$$\int_\Gamma \pi_h v\,ds = \int_\Gamma v\,ds, \quad \Gamma\in\partial\mathbb{T}_h,$$

defines on each cell $T\in\mathbb{T}_h$ a local interpolation operator $\pi_h^T : H^1(T)^2 \to P_1(T)^2$. For $v\in H$ the pieces $\pi_h^T v$ can be composed to a global function $\pi_h v\in H_h$. The (even cellwise) H^1-stability of this interpolation operator can be concluded by observing that $\Delta\pi_h^T v_{|K}\equiv 0$ and $\partial_n\pi_h^T v|_\Gamma \equiv const.$ directly from the construction as follows:

$$\begin{aligned}\|\nabla\pi_h^T v\|_T^2 &= \int_{\partial T}\partial_n\pi_h^T v\pi_h^T v\,ds - (\Delta\pi_h^T v,\pi_h^T v)_T = \int_{\partial T}\partial_n\pi_h^T vv\,ds\\ &= (\nabla\pi_h^T v,\nabla v)_T + (\Delta\pi_h^T v, v)_T \le \|\nabla\pi_h^T v\|\|\nabla v\|_T.\end{aligned}$$

(ii) For the $\tilde{Q}_1^{\rm nc}/P_0$ element the argument is similar and left as an exercise. Q.E.D.

For the nonconforming Stokes elements considered the interpolation operator π_h is just the natural nodal interpolation $\pi_h = i_h : H\to H_h$. For that, we have the usual error estimate

$$\|v - i_h v\| + h\|\nabla_h(v-i_h v)\| \le c_i h^2\|\nabla^2 v\|. \tag{4.2.65}$$

The following theorem shows that the nonconforming Stokes elements considered allow for error estimates of optimal order.

Theorem 4.5 (A priori error estimate): *For the nonconforming Stokes elements of type $P_1^{\rm nc}/P_0^{\rm dc}$ and $\tilde{Q}_1^{\rm nc}/P_0^{\rm dc}$ (nonparametric) there hold the following a priori error estimates:*

$$\|\nabla_h(v-v_h)\| + \|p-p_h\| \le h\|f\|, \tag{4.2.66}$$

$$\|v-v_h\| \le ch^2\|f\|. \tag{4.2.67}$$

Proof: We give the proof only for the simpler $P_1^{\rm nc}/P_0^{\rm dc}$ element. The proof for the $\tilde{Q}_1^{\rm nc}/P_0^{\rm dc}$ element is posed as an exercise.

(i) Let again $e_h := v - v_h$ and $\eta_h := p - p_h$. With an arbitrary $\varphi_h\in H_h$ there holds

$$\|\nabla_h e_h\|^2 = (\nabla_h e_h,\nabla_h(v-\varphi_h)) + (\nabla_h e_h,\nabla_h(\varphi_h - v_h)),$$

and for an arbitrary $\psi_h\in H_h$:

$$\begin{aligned}(\nabla_h e_h,\nabla_h\psi_h) &= (\nabla v,\nabla_h\psi_h) - (\nabla v_h,\nabla\psi_h)\\ &= (\nabla v,\nabla_h\psi_h) - (f,\psi_h) - (p_h,\nabla_h\cdot\psi_h)\\ &= (\nabla v,\nabla_h\psi_h) - (p,\nabla_h\cdot\psi_h) - (f,\psi_h) + (\eta_h,\nabla_h\cdot\psi_h).\end{aligned}$$

Further, with an arbitrary $\chi_h \in L_h$:

$$\begin{aligned}(\eta_h, \nabla_h \cdot (\varphi_h - v_h)) &= (\eta_h, \nabla_h \cdot (\varphi_h - v)) + (\eta_h, \nabla_h \cdot e_h) \\ &= (\eta_h, \nabla_h \cdot (\varphi_h - v)) + (\eta_h, \nabla \cdot v) - (\eta_h, \nabla_h \cdot v_h) \\ &= (\eta_h, \nabla_h \cdot (\varphi_h - v)) + (p - \chi_h, \nabla_h \cdot e_h).\end{aligned}$$

Combining the foregoing relations, we obtain

$$\begin{aligned}\|\nabla_h e_h\|^2 \le \|\nabla_h e_h\| \|\nabla_h (v - \varphi_h)\| + \Delta_h(v,p)\big\{\|\nabla_h(\varphi_h - v)\| + \|\nabla_h e_h\|\big\} \\ + \|\eta_h\| \|\nabla(v - \varphi_h)\| + \|p - \chi_h\| \|\nabla_h e_h\|,\end{aligned}$$

with the "nonconformity term"

$$\Delta_h(v,p) := \max_{\psi_h \in H_h} \frac{(\nabla v, \nabla_h \psi_h) - (p, \nabla_h \cdot \psi_h) - (f, \psi_h)}{\|\nabla_h \psi_h\|}.$$

Using Young's inequality $ab \le \frac{1}{4}\varepsilon a^2 + \varepsilon^{-1} b^2$, for $a, b \in \mathbb{R}_+$ and arbitrary $\varepsilon > 0$, we conclude

$$\begin{aligned}\|\nabla_h e_h\|^2 \le \tfrac{3}{4}\varepsilon_1 \|\nabla_h e_h\|^2 + \varepsilon_1^{-1}\big\{\|\nabla_h(v - \varphi_h)\|^2 + \|p - \chi_h\|^2 + \Delta_h(v,p)^2\big\} \\ + \tfrac{1}{4}\varepsilon_2 \|\eta_h\|^2 + \big\{\varepsilon_2^{-1} + \tfrac{1}{2}\big\}\Delta_h(v,p)^2 + \tfrac{1}{2}\|\nabla_h(v - \varphi_h)\|^2,\end{aligned}$$

and setting $\varepsilon_1 = 1$:

$$\begin{aligned}\|\nabla_h e_h\|^2 \le 6 \min_{\varphi_h \in H_h} \|\nabla_h (v - \varphi_h)\|^2 + 4 \min_{\chi_h \in L_h} \|p - \chi_h\|^2 \\ + \{\varepsilon_2^{-1} + 6\}\Delta_h(v,p)^2 + \varepsilon_2 \|\eta_h\|^2.\end{aligned} \tag{4.2.68}$$

This intermediate result will be used later on.

(ii) For estimating the pressure error $\|\eta_h\|$, we use the "inf-sup" stability relation (4.2.62). With an arbitrary $\chi_h \in L_h$ it follows that

$$\begin{aligned}\|\eta_h\| &\le \|p - \chi_h\| + \|\chi_h - p_h\| \\ &\le \|p - \chi_h\| + \beta_*^{-1} \max_{\varphi_h \in H_h} \frac{(\chi_h - p_h, \nabla_h \cdot \varphi_h)}{\|\nabla_h \varphi_h\|} \\ &\le \|p - \chi_h\| + \beta_*^{-1} \max_{\varphi_h \in H_h} \frac{(\chi_h - p, \nabla_h \cdot \varphi_h)}{\|\nabla_h \varphi_h\|} + \beta_*^{-1} \max_{\varphi_h \in H_h} \frac{(\eta_h, \nabla_h \cdot \varphi_h)}{\|\nabla_h \varphi_h\|} \\ &\le (\beta_*^{-1} + 1)\|p - \chi_h\| + \beta_*^{-1} \max_{\varphi_h \in H_h} \frac{(\eta_h, \nabla_h \cdot \varphi_h)}{\|\nabla_h \varphi_h\|}.\end{aligned}$$

Further, there holds

$$\begin{aligned}(\eta_h, \nabla_h \cdot \varphi_h) &= (p, \nabla_h \cdot \varphi_h) - (p_h, \nabla_h \cdot \varphi_h) \\ &= (p, \nabla_h \cdot \varphi_h) - (\nabla v_h, \nabla_h \varphi_h) + (f, \varphi_h) \\ &= (p, \nabla_h \cdot \varphi_h) - (\nabla v, \nabla_h \varphi_h) + (f, \varphi_h) - (\nabla e_h, \nabla_h \varphi_h)\end{aligned}$$

and, consequently,

$$\max_{\varphi_h \in H_h} \frac{(\eta_h, \nabla_h \cdot \varphi_h)}{\|\nabla_h \varphi_h\|} \le \Delta_h(v,p) + \|\nabla_h e_h\|.$$

This implies

$$\|\eta_h\| \le (\beta_*^{-1}+1)\|p-\chi_h\| + \beta_*^{-1}\big\{\Delta_h(v,p) + \|\nabla_h e_h\|\big\}. \tag{4.2.69}$$

Combination of the intermediate results (4.2.68) and (4.2.69) and choice of $\varepsilon_2 := \frac{1}{2}\beta_*$ gives us

$$\|\nabla_h e_h\| + \|\eta_h\| \le c(\beta_*)\Big\{ \min_{\varphi_h\in H_h} \|\nabla_h(v-\varphi_h)\| + \min_{\chi\in L_h} \|p-\chi_h\| + \Delta_h(v,p)\Big\}, \tag{4.2.70}$$

with a generic constant $c(\beta_*) \approx \beta^{-1}+1 > 0$.

(iii) It remains to estimate the nonconformity term $\Delta_h(v,p)$. For that, we reformulate observing that $-\Delta v + \nabla p = f$ as follows:

$$\begin{aligned}(\nabla v, \nabla_h\psi_h) - (p, \nabla_h\cdot\psi_h) - (f,\psi_h) &= \sum_{T\in\mathbb{T}_h} \big\{(-\Delta v + \nabla p - f, \psi_h)_T + (\partial_n v - pn, \psi_h)_{\partial T}\big\} \\ &= \sum_{T\in\mathbb{T}_h} (\partial_n v - pn, \psi_h)_{\partial T} = \sum_{\Gamma\in\partial\mathbb{T}_h} (\partial_n v - pn, \psi_h)_\Gamma,\end{aligned}$$

where $\partial\mathbb{T}_h$ again denotes the set of all edges Γ of the cells $T\in\mathbb{T}_h$. Each side $\Gamma\in\partial\mathbb{T}_h$ is either common side of two cells $T, T'\in\mathbb{T}_h$ or part of $\partial\Omega$. The function $\partial_n v - pn$ on $\Gamma\subset T$ has in view of the continuity of ∇v and p opposite sign $\partial_n v - pn$ on $\Gamma\subset T'$. Consequently, we can write

$$\sum_{\Gamma\in\partial\mathbb{T}_h} (\partial_n v - pn, \psi_h)_\Gamma = \sum_{\Gamma\in\partial\mathbb{T}_h} (\partial_n v - pn, [\psi_h])_\Gamma,$$

where

$$[\psi_h]_\Gamma := \begin{cases} \frac{1}{2}(\psi_{h|\Gamma\cap T} - \psi_{h|\Gamma\cap T'}), & \Gamma = T\cap T', \\ \psi_{h|\Gamma}, & \Gamma\subset\partial\Omega. \end{cases}$$

On each side $\Gamma\in\partial\mathbb{T}_h$ by the special properties of $\psi_h\in H_h$ the jump $[\psi_h]$ has mean value zero. Consequently,

$$(\partial_n v - pn, [\psi_h])_\Gamma = (\partial_n v - pn, [\psi_h] - \overline{[\psi_h]}_\Gamma)_\Gamma = (\partial_n v - pn - (\overline{\partial_n v - pn})_\Gamma, [\psi_h] - \overline{[\psi_h]}_\Gamma)_\Gamma,$$

with the mean values

$$(\overline{\partial_n v - pn})_\Gamma := |\Gamma|^{-1}\int_\Gamma (\partial_n v - pn)\,ds, \qquad \overline{[\psi_h]}_\Gamma := |\Gamma|^{-1}\int_\Gamma [\psi_h]\,ds.$$

Splitting $[\psi_h]$ into the contributions of the two neighboring cells T, T' results in

$$(\partial_n v - pn, [\psi_h])_\Gamma = (\partial_n v - pn - (\overline{\partial_n v - pn})_\Gamma, \psi_{h|K} - \overline{\psi_h}_\Gamma)_\Gamma - (\overline{\partial_n v - pn})_\Gamma, \psi_h|_{T'} - \overline{\psi_h}_\Gamma)_\Gamma.$$

Further, by the usual transformation on the reference cell, we see that

$$\|\partial_n v - pn - (\overline{\partial_n v - pn})_\Gamma\|_\Gamma \le c_i h^{1/2}\{\|\nabla^2 v\|_T + \|\nabla p\|\}, \qquad \|\psi_h|_T - \overline{\psi_h}_\Gamma\|_\Gamma \le c_i h^{1/2}\|\nabla\psi_h\|_T,$$

and analogously for the cell T'. From the foregoing results, it now follows that for $\Gamma = T\cap T'$:

$$\begin{aligned}|(\partial_n v - pn, [\psi_h])_\Gamma| &\le \|\partial_n v - pn - (\overline{\partial_n v - pn})_\Gamma\|_\Gamma\big\{\|\psi_h|_T - \overline{\psi_h}_\Gamma\|_\Gamma + \|\psi_h|_{T'} - \overline{\psi_h}_\Gamma\|_\Gamma\big\} \\ &\le c^2 h\big\{\|\nabla^2 v\|_T + \|\nabla p\|_T\big\}\|\nabla_h\psi_h\|_{T\cup T'},\end{aligned}$$

and correspondingly for $\Gamma \subset \partial\Omega$. From this, we conclude

$$\Big| \sum_{\Gamma \in \partial \mathbb{T}_h} (\partial_n v - pn, \psi_h)_\Gamma \Big| \leq ch\{\|\nabla^2 v\| + \|\nabla p\|\}\|\nabla_h \psi_h\|,$$

and, consequently, witha constant $c > 0$ independent of β,

$$\Delta_h(v,p) \leq ch\{\|\nabla^2 v\| + \|\nabla p\|\} \leq ch\|f\|. \tag{4.2.71}$$

Using the interpolation estimates (proven by the usual transformation argument as in the Lemma of Bramble/Hilbert)

$$\min_{\varphi_h \in H_h} \|\nabla_h(v - \varphi_h)\| + \min_{\chi \in L_h} \|p - \chi_h\| \ \leq \ ch\{\|\nabla^2 v\| + \|\nabla p\|\} \leq ch\|f\|,$$

we finally obtain the error estimate (4.2.66).

(iv) For proving the L^2-error estimate (4.2.67), we again employ a duality argument. But the details are not given here. Q.E.D.

The analysis of the application of nonconforming Stokes elements in the approximation of the (nonlinear) Navier-Stokes problem involves estimation of the modified nonlinear form $\tilde{b}_h(v_h, \psi_h, \varphi_h) := \frac{1}{2}(v_h \cdot \nabla_h \psi, \varphi_h) - \frac{1}{2}(v_h \cdot \nabla_h \varphi_h, \psi)$. To this end, besides the "discrete" Poincaré inequality, we need discrete versions of some Sobolev inequalities.

Lemma 4.8: *For the $P_1^{\rm nc}/P_0^{\rm dc}$ and the $\tilde{Q}_1^{\rm nc}/P_0^{\rm dc}$ Stokes elements there holds the following discrete Sobolev inequality in two and three dimensions:*

$$\max\{\|v_h\|_3, \|v_h\|_6\} \leq c_*^{\rm nc} \|\nabla_h v_h\|, \quad v_h \in H_h. \tag{4.2.72}$$

Proof: The proof is posed as exercise. Q.E.D.

4.2.2 Stabilized Stokes elements

There are good reasons to use simple vertex-oriented finite elements in the discretization of the Stokes and Navier-Stokes problem in which the degrees of freedom of velocity and pressure live at the same nodal points. Simplest examples are the conforming $P_1^{\rm c}/P_1^{\rm c}$ and the $Q_1^{\rm c}/Q_1^{\rm c}$ elements, which however are not "inf-sup" stable by themselves. In the following, we discuss a technique for stabilizing these elements. We recall the following relation for $q_h \in L_h$:

$$\beta\|q_h\| \leq \sup_{\varphi \in H} \left\{ \frac{(q_h, \nabla \cdot \varphi_h)}{\|\nabla \varphi_h\|} \frac{\|\nabla \varphi_h\|}{\|\nabla \varphi\|} \right\} + \sup_{\varphi \in H} \frac{(q_h, \nabla \cdot (\varphi - \varphi_h))}{\|\nabla \varphi\|}, \tag{4.2.73}$$

where $\varphi_h := i_h^* \varphi \in H_h$ is an H^1-stable (modufied) nodal interpolant (such as $I_h^{(1)}$ given by Lemma 4.4) satisfying on each cell:

$$\|\varphi - i_h^* \varphi\|_T + h_T \|\nabla(\varphi - i_h^* \varphi)\|_T \leq \tilde{c}_i h_T \|\varphi\|_{H^1(\tilde{T})},$$

with the union $\tilde{T}$ of all cells neighboring T. Further, as for the standard nodal interpoalation, there holds for $\varphi \in H^2(\Omega)$:

$$\|\varphi - i_h^*\varphi\|_T + h_T\|\nabla(\varphi - i_h^*\varphi)\|_T \le c_i h_T^2 \|\varphi\|_{H^2(\tilde{T})}.$$

For the second term on the right in (4.2.73), we obtain after integration by parts (observe $\varphi - i_h^*\varphi \in H$):

$$\begin{aligned}
|(q_h, \nabla\cdot(\varphi - i_h^*\varphi))| = |(\nabla q_h, \varphi - i_h^*\varphi)| &\le \sum_{T\in\mathbb{T}_h} \|\nabla q_h\|_T \|\varphi - i_h^*\varphi\|_T \\
&\le c_i \sum_{K\in\mathbb{T}_h} h_T \|\nabla q_h\|_T \|\varphi\|_{H^1(\tilde{T})} \le c_i \Big(\sum_{T\in\mathbb{T}_h} h_T^2 \|\nabla q_h\|_T^2 \Big)^{1/2} \Big(\sum_{T\in\mathbb{T}_h} \|\varphi\|_{H^1(\tilde{T})}^2 \Big)^{1/2} \\
&\le c_0 c_i \Big(\sum_{T\in\mathbb{T}_h} h_T^2 \|\nabla q_h\|_T^2 \Big)^{1/2} \|\varphi\|_{H^1} \le c_0 c_i \Big(\sum_{T\in\mathbb{T}_h} h_T^2 \|\nabla q_h\|_T^2 \Big)^{1/2} \|\nabla\varphi\|,
\end{aligned}$$

with a for shape uniform decompositions $(\mathbb{T}_h)_{h>0}$ fixed constant $c_0 \ge 1$. For the first term on the right in (4.2.73), we have

$$\sup_{\varphi\in H} \left\{ \frac{(q_h, \nabla\cdot i_h^*\varphi)}{\|\nabla i_h^*\varphi\|} \frac{\|\nabla i_h^*\varphi\|}{\|\nabla\varphi\|} \right\} \le c_i \sup_{\varphi_h\in H_h} \frac{(q_h, \nabla\cdot\varphi_h)}{\|\nabla\varphi_h\|}.$$

Together the two last estimates imply

$$\beta_h \|q_h\| \le \sup_{\varphi_h\in H_h} \frac{(q_h, \nabla\cdot\varphi_h)}{\|\nabla\varphi_h\|} + \Big(\sum_{T\in\mathbb{T}_h} h_T^2 \|\nabla q_h\|_T^2 \Big)^{1/2}, \tag{4.2.74}$$

with the constant $\beta_h := (c_0 c_i)^{-1}\beta$. This result suggests the following modification of the approximation scheme (4.2.30 - 4.2.31):

$$(\nabla v_h, \nabla\varphi_h) - (p_h, \nabla\cdot\varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in H_h, \tag{4.2.75}$$

$$(\chi_h, \nabla\cdot v_h) + s_h(\chi_h, p_h) = 0 \quad \forall \chi_h \in L_h, \tag{4.2.76}$$

with the "stabilization form"

$$s_h(\chi_h, p_h) := \alpha \sum_{T\in\mathbb{T}_h} h_T^2 (\nabla\chi_h, \nabla p_h)_T,$$

where the constant $\alpha > 0$ has to be apprpriately chosen.

Lemma 4.9 (Stability): *The modified approximation scheme (4.2.75 - 4.2.76) possesses a unique solution $\{v_h, p_h\} \in H_h \times L_h$, and there holds the stability estimate*

$$\|\nabla v_h\| + \|p_h\| + s_h(p_h, p_h)^{1/2} \le c\|f\|. \tag{4.2.77}$$

Proof: For proving the existence of solutions it suffices to prove their uniqueness. For that, we set $\varphi_h := v_h$ in (4.2.75) and $\chi_h := p_h$ in (4.2.76) to otain

$$\begin{aligned}
\|\nabla v_h\|^2 - (p_h, \nabla\cdot v_h) &= (f, v_h), \\
(p_h, \nabla\cdot v_h) + s_h(p_h, p_h) &= 0.
\end{aligned}$$

Addition of these equations gives us

$$\|\nabla v_h\|^2 + s_h(p_h, p_h) = (f, v_h).$$

In case of homogeneous data, i. e., $f \equiv 0$, it follows that $v_h \equiv 0$ and $\nabla p_h \equiv 0$. This also implies $p_h \equiv 0$ for $.p_h \in L_h$. Further, since $(f, v_h) \leq \|f\|\|v_h\| \leq c\|f\|\|\nabla v_h\|$, we conclude the stability estimate

$$\|\nabla v_h\| + s_h(p_h, p_h)^{1/2} \leq c\|f\|.$$

In view of the "inf-sup" stability relation (4.2.74), we further obtain

$$\|p_h\| \leq c\|f\|,$$

what completes the proof. Q.E.D.

Theorem 4.6 (Stabilized Stokes elements): *For the solution* $\{v_h, p_h\} \in H_h \times L_h$ *of the modified approximation scheme (4.2.75 - 4.2.76) there hold the error estimates*

$$\|\nabla(v - v_h)\| + \|p - p_h\| + s_h(p - p_h, p - p_h)^{1/2} \leq ch\|f\|, \tag{4.2.78}$$

$$\|v - v_h\| \leq ch^2\|f\|. \tag{4.2.79}$$

Proof: Subtracting the discrete Stokes equations from the continuous ones results in

$$(\nabla(v - v_h), \nabla\varphi_h) - (p - p_h, \nabla \cdot \varphi_h) = 0, \quad \varphi_h \in H_h. \tag{4.2.80}$$

$$(\chi_h, \nabla \cdot (v - v_h)) + s_h(\chi_h, p - p_h) = s_h(\chi_h, p). \tag{4.2.81}$$

(i) We begin with the estimation of $\|\nabla(v - v_h)\|$ and $s_h(p - p_h, p - p_h)$. Because of (4.2.80) it follows, with arbitrary $\varphi_h \in H_h$:

$$\begin{aligned}
\|\nabla(v - v_h)\|^2 &= (\nabla(v - v_h), \nabla(v - \varphi_h) + (\nabla(v - v_h), \nabla(\varphi_h - v_h)) \\
&= (\nabla(v - v_h), \nabla(v - \varphi_h) + (p - p_h, \nabla \cdot (\varphi_h - v_h)) \\
&= (\nabla(v - v_h), \nabla(v - \varphi_h) + (p - p_h, \nabla \cdot (\varphi_h - v)) + (p - p_h, \nabla \cdot (v - v_h)),
\end{aligned}$$

and further because of (4.2.81), with arbitrary $\chi_h \in L_h$:

$$\begin{aligned}
\|\nabla(v - v_h)\|^2 &= (\nabla(v - v_h), \nabla(v - \varphi_h) + (p - p_h, \nabla \cdot (\varphi_h - v)) + (p - \chi_h, \nabla \cdot (v - v_h)) \\
&\quad + (\chi_h - p_h, \nabla \cdot (v - v_h)) \\
&= (\nabla(v - v_h), \nabla(v - \varphi_h) + (p - p_h, \nabla \cdot (\varphi_h - v)) + (p - \chi_h, \nabla \cdot (v - v_h)) \\
&\quad - s_h(\chi_h - p_h, p - p_h) + s_h(\chi_h - p_h, p) \\
&= (\nabla(v - v_h), \nabla(v - \varphi_h) + (p - p_h, \nabla \cdot (\varphi_h - v)) + (p - \chi_h, \nabla \cdot (v - v_h)) \\
&\quad - s_h(\chi_h - p, p - p_h) - s_h(p - p_h, p - p_h) + s_h(\chi_h - p_h, p).
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}\|\nabla(v-v_h)\|^2 + s_h(p-p_h,p-p_h) &= (\nabla(v-v_h),\nabla(v-\varphi_h)+(p-p_h,\nabla\cdot(\varphi_h-v))\\ &\quad+(p-\chi_h,\nabla\cdot(v-v_h))-s_h(\chi_h-p,p-p_h)+s_h(\chi_h-p_h,p)\\ &\le \|\nabla(v-v_h)\|\|\nabla(v-\varphi_h\|+\|p-p_h\|\|\nabla\cdot(\varphi_h-v)\|\\ &\quad+\|p-\chi_h\|\|\nabla\cdot(v-v_h)\|+s_h(\chi_h-p,\chi_h-p)^{1/2}s_h(p-p_h,p-p_h)^{1/2}\\ &\quad+s_h(\chi_h-p_h,\chi_h-p_h)^{1/2}s_h(p,p)^{1/2}\end{aligned}$$

Observing $\|\nabla p\| \le c\|f\|$ it follows that

$$s_h(p,p)^{1/2} = \Big(\alpha\sum_{T\in\mathbb{T}_h} h_T^2\|\nabla p\|_T^2\Big)^{1/2} \le ch\Big(\sum_{T\in\mathbb{T}_h}\|\nabla p\|_T^2\Big)^{1/2} = ch\|\nabla p\| \le ch\|f\|. \tag{4.2.82}$$

Further, using the local inverse relation for finite elements, we conclude

$$\begin{aligned}s_h(\chi_h-p_h,\chi_h-p_h)^{1/2} &= \Big(\alpha\sum_{T\in\mathbb{T}_h} h_T^2\|\nabla(\chi_h-p_h)\|_T^2\Big)^{1/2} \le c\Big(\sum_{T\in\mathbb{T}_h}\|\chi_h-p_h\|_T^2\Big)^{1/2}\\ &\le c\{\|p-p_h\|+\|p-\chi_h\|\}.\end{aligned}$$

From this, we conclude with help of the inequality $ab \le \varepsilon^2a^2 + (4\varepsilon^2)^{-1}b^2$:

$$\begin{aligned}\|\nabla(v-v_h)\|^2 + s_h(p-p_h,p-p_h) &\le c(1+\varepsilon^{-1})\big\{\|\nabla(v-\varphi_h)\|^2+\|p-\chi_h\|^2\\ &\quad+s_h(p-\chi_h,p-\chi_h)+h^2\|f\|^2\big\}+\varepsilon\|p-p_h\|^2.\end{aligned} \tag{4.2.83}$$

(ii) Next, we estimate $\|p-p_h\|$. By the "inf-sup" stability estimate (4.2.74) and the Galerkin orthogonality relation (4.2.80) it follows

$$\begin{aligned}\|p-p_h\| &\le \|p-\chi_h\|+\|\chi_h-p_h\|\\ &\le \|p-\chi_h\|+\beta_*^{-1}\sup_{\psi_h\in H_h}\frac{(\chi_h-p_h,\nabla\cdot\psi_h)}{\|\nabla\psi_h\|}+\beta_*^{-1}s_h(\chi_h-p_h,\chi_h-p_h)^{1/2}\\ &\le \|p-\chi_h\|+\beta^{-1}\sup_{\psi_h\in H_h}\frac{(\chi_h-p,\nabla\cdot\psi_h)}{\|\nabla\psi_h\|}+\beta^{-1}\sup_{\psi_h\in H_h}\frac{(p-p_h,\nabla\cdot\psi_h)}{\|\nabla\psi_h\|}\\ &\quad+\beta_*^{-1}s_h(\chi_h-p_h,\chi_h-p_h)^{1/2}\\ &= \|p-\chi_h\|+\beta_*^{-1}\sup_{\psi_h\in H_h}\frac{(\chi_h-p,\nabla\cdot\psi_h)}{\|\nabla\psi_h\|}+\beta_*^{-1}\sup_{\psi_h\in H_h}\frac{(\nabla(v-v_h),\nabla\psi_h)}{\|\nabla\psi_h\|}\\ &\quad+\beta_*^{-1}s_h(\chi_h-p_h,\chi_h-p_h)^{1/2}\end{aligned}$$

and, consequently,

$$\|p-p_h\| \le (1+\beta_*^{-1})\|p-\chi_h\|+\beta_*^{-1}\|\nabla(v-v_h)\|+\beta_*^{-1}s_h(\chi_h-p_h,\chi_h-p_h)^{1/2}. \tag{4.2.84}$$

Combining the foregoing estimates with (4.2.83) gives us

$$\begin{aligned}\|\nabla(v-v_h)\|^2 + s_h(p-p_h,p-p_h) &\le c(1+\varepsilon^{-1})\big\{\|\nabla(v-\varphi_h)\|^2+\|p-\chi_h\|^2\\ &\quad+s_h(p-\chi_h,p-\chi_h)+h^2\|f\|^2\big\}\\ &\quad+c\varepsilon\big\{(1+\beta_*^{-1})^2\|p-\chi_h\|^2+\beta_*^{-2}\|\nabla(v-v_h)\|^2+\beta_*^{-2}s_h(\chi_h-p_h,\chi_h-p_h)\big\}.\end{aligned}$$

Now, fixing $\varepsilon > 0$ sufficiently small, we conclude

$$\|\nabla(v - v_h)\| + s_h(p - p_h, p - p_h)^{1/2} \le c\big\{\|\nabla(v - \varphi_h)\| + \|p - \chi_h\| + s_h(p - \chi_h, p - \chi_h)^{1/2} + h\|f\|\big\}.$$

For $\varphi_h = i_h^* v$ and $\chi_h = i_h^* p$, we further obtain

$$\|\nabla(v - v_h)\| + s_h(p - p_h, p - p_h)^{1/2} \le ch\{\|\nabla^2 v\| + \|\nabla p\| + \|f\|\} \le ch\|f\|.$$

Therefore, with (4.2.84) it follows that

$$\|p - p_h\| \le ch\|f\|.$$

(iii) For estimating the L^2 error $\|v - v_h\|$, we again use a duality argument. Let $\{z, q\} \in H \times L$ the solution of the auxiliary "dual Stokes problem"

$$(\nabla\varphi, \nabla z) - (q, \nabla \cdot \varphi) = (\varphi, v - v_h)\|v - v_h\|^{-1} \quad \forall \varphi \in H, \tag{4.2.85}$$

$$(\chi, \nabla \cdot z) = 0 \quad \forall \chi \in L. \tag{4.2.86}$$

This is in $H^2(\Omega)^d \times H^1(\Omega)$ and there holds the a priori estimate

$$\|z\|_{H^2} + \|q\|_{H^1} \le c\|v - v_h\|\|v - v_h\|^{-1} = c. \tag{4.2.87}$$

With the test function $\varphi := v - v_h$ it follows using the Galerkin orthogonality relation with $i_h^* z \in H_h$ and $i_h^* q \in L_h$:

$$\begin{aligned}
\|v - v_h\| &= (\nabla(v - v_h), \nabla z) - (q, \nabla \cdot (v - v_h)) \\
&= (\nabla(v - v_h), \nabla(z - i_h^* z)) + (\nabla(v - v_h), \nabla i_h^* z) - (q - i_h^* q, \nabla \cdot (v - v_h)) \\
&\quad - (i_h^* q, \nabla \cdot (v - v_h)) \\
&= (\nabla(v - v_h), \nabla(z - i_h^* z)) + (p - p_h, \nabla \cdot i_h^* z) - (q - i_h^* q, \nabla \cdot (v - v_h)) \\
&\quad + s_h(i_h^* q, p - p_h) - s_h(i_h^* q, p) \\
&= (\nabla(v - v_h), \nabla(z - i_h^* z)) + (p - p_h, \nabla \cdot (i_h^* z - z)) - (q - i_h^* q, \nabla \cdot (v - v_h)) \\
&\quad + s_h(i_h^* q, p - p_h) - s_h(i_h^* q, p).
\end{aligned}$$

We further estimate using the estimates for i_h^* as follows:

$$\begin{aligned}
\|v - v_h\| &\le \|\nabla(v - v_h)\|\|\nabla(z - i_h^* z)\| + \|p - p_h\|\|\nabla \cdot (i_h^* z - z)\| + \|q - i_h^* q\|\|\nabla \cdot (v - v_h)\| \\
&\quad + s_h(i_h^* q, i_h^* q)^{1/2} s_h(p - p_h, p - p_h)^{1/2}) + s_h(i_h^* q, i_h^* q)^{1/2} s_h(p, p)^{1/2} \\
&\le ch\|\nabla(v - v_h)\|\|z\|_{H^2} + ch\|p - p_h\|\|z\|_{H^2} + ch\|q\|_{H^1}\|\nabla(v - v_h)\| \\
&\quad + ch\|\nabla q\| s_h(p - p_h, p - p_h)^{1/2} + ch^2\|\nabla q\|\|\nabla p\| \\
&\le ch\big\{\|\nabla(v - v_h)\| + \|p - p_h\| + s_h(p - p_h, p - p_h)^{1/2}\big\} + ch^2\|f\|.
\end{aligned}$$

This completes the proof. Q.E.D.

The error estimates (4.2.78) and (4.2.79) are optimal with respect to the order as well as the regularity requirements on the solution $\{v, p\}$. The same concept of pressure stabilization can also be used for the P_r^c/P_r^c and Q_r^c/Q_r^c Stokes elements with polynomial degree $r \ge 2$. In this case the achievable order of approximation is limited by the stabilization term in (4.2.76) to $\mathcal{O}(h^2)$. This order barrier can be overcome by choosing the following more consistent stabi-

lization:

$$(\nabla v_h, \nabla \varphi_h) - (p_h, \nabla \cdot \varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in H_h, \tag{4.2.88}$$

$$(\chi_h, \nabla \cdot v_h) + s_h(\chi_h, \nabla p_h) = g_h(\chi_h) \quad \forall \chi_h \in L_h, \tag{4.2.89}$$

with the correction term

$$g_h(\chi_h) := \alpha \sum_{T \in \mathbb{T}_h} h_T^2 \big\{ (\nabla \chi_h, f)_T + (\nabla \chi_h, \Delta v_h)_T \big\}.$$

In this case the stabilized discrete system is exactly fulfilled by the continuous solution $\{v, p\}$, such that full Galerkin orthogonality holds.

Remark 4.8: The stabilized P_1^c/P_1^c Stokes element has a close realtion to the "inf-sup" stable MINI element. Actuslly there is an algebraic equivalence for appropriately chosen stabilization parameters (exercise).

Remark 4.9: The stabilization technique described above is a particular case in a whole family of similar approaches. Though it is universally applicable is suffers from certain shortcomings:

- high computational costs for evaluating the system matrices and possibly additional coupling between velocity and pressure unknowns;
- triggering of unphysical flow behavior across "free" outflow boundary.

These defects can partially be lowered by other stabilization techniques, which use edge-oriented stabilisation

$$s_h(\chi_h, p_h) = \alpha \sum_{t \in \mathbb{T}_h} h_T([\partial_n \chi_h], [\partial_n p_h])_{\partial T},$$

where $[\cdot]$ denotes jump across inter-cell boundaries, or by "local projection stabilization" (LPS)

$$s_h(\chi_h, p_h) = \alpha \sum_{T \in \mathbb{T}_h} (\chi_h - \pi_{2h}\chi_h, p_h - \pi_{2h} p_h)_T,$$

where π_{2h} is the projection on a twice coarser mesh.

4.2.3 Navier-Stokes problem: the small-data case

Now, we come back to the finite element approximation of the original Navier-Stokes problem. Again, we restrict the consideration to polygonal/polyhedral domains $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ and to the case of homogeneous Dirichlet boundary conditions $v|_{\partial\Omega} = 0$. At first, we consider again the small-data case, $c_*^2 \nu^{-2} \|f\|_{-1} < 1$, in which the continuous problem is guaranteed to possess a unique solution $\{v, p\} \in H \times L$. This solution is in $H^2(\Omega)^d \times H^1(\Omega)$ and satisfies the a priori estimate

$$\|v\|_{H^2} + \|p\|_{H^1} \le c \|f\|. \tag{4.2.90}$$

Let $H_h \times L_h$ be any of the uniformly "inf-sup" stable Stokes elements introduced above. In the following discussion, we concentrate on the "conforming" case, $H_h \times L_h \subset H \times L$. But

all results obtained also hold true in modified form for the "nonconforming" and "stabilized" Stokes elements. Further, for the reasons discussed above, we use the symmetrized nonlinear form

$$\tilde{n}(v_h,\psi_h,\varphi_h) := \tfrac{1}{2}n(v_h,\psi_h,\varphi_h) - \tfrac{1}{2}n(v_h,\varphi_h,\psi_h), \quad n(v_h,\psi_h,\varphi_h) := (v_h\cdot\nabla\psi_h,\varphi_h).$$

With this notation the discrete problems read as follows: Find $\{v_h,p_h\}\in H_h\times L_h$, such that

$$\nu(\nabla v_h,\nabla\varphi_h)+\tilde{n}(v_h,v_h,\varphi_h)-(p_h,\nabla\cdot\varphi_h)=(f,\varphi_h)\quad\forall\varphi_h\in H_h, \tag{4.2.91}$$

$$(\chi_h,\nabla\cdot v_h)=0\quad\forall\chi_h\in L_h. \tag{4.2.92}$$

Due to the use of the symmetrized nonlinear form $\tilde{n}(\cdot,\cdot,\cdot)$ in the small-data case considered the existence of unique solutions can be shown by employing the Brouwer fixed point theorem and the "inf-sup" stability analogously as on the continuous level (exercise). For the following error analysis, we assume again that the discrete spaces $H_h\times L_L$ have the following minimal approximation properties: There exist interpolation operators $i_h : H\cap H^2(\Omega)^d\to H_h$ and $j_h : L\cap H^1(\Omega)\to L_h$, such that

$$\|w-i_hw\|+h\|\nabla(w-i_hw)\|\le c_ih^2\|w\|_{H^2},\quad w\in H\cap H^2(\Omega)^d, \tag{4.2.93}$$

$$\|q-j_hq\|\le c_ih\|q\|_{H^1},\quad q\in L\cap H^1(\Omega). \tag{4.2.94}$$

Theorem 4.7 (Error estimates): *Under the above assumptions (small data and "inf-sup" stable approximation), there hold the following error estimates:*

$$\|\nabla(v-v_h)\|+\|p-p_h\|\le c(\nu,\|f\|)h, \tag{4.2.95}$$

$$\|v-v_h\|\le c(\nu.\|f\|)h^2. \tag{4.2.96}$$

Proof: The proof uses similar arguments as that in the context of the linear Stokes problem. Additionally, we have to treat the nonlinearity. For that, we note the following relations for functions $w,\varphi,\psi\in H$:

$$|\tilde{n}(w,\psi,\varphi)|\le c_*^2\|\nabla w\|\|\nabla\varphi\|\|\nabla\psi\|, \tag{4.2.97}$$

$$\tilde{n}(w,\varphi,\varphi)=0. \tag{4.2.98}$$

Further, by assumption, we have $q:=c_*^2\nu^{-2}\|f\|_{-1}<1$. From the equations satisfied by v and v_h, using there the test functions $\varphi=v$ and $\varphi_h=v_h$, we conclude as usual the estimates

$$\|\nabla v\|\le\nu^{-1}\|f\|_{-1},\quad \|\nabla v_h\|\le\nu^{-1}\|f\|_{-1}.$$

(i) We split the error $e_h=v-v_h$ into two parts, $e_h=\xi_h+\eta_h$, where $\xi_h:=v-w_h$ is the error in the approximatetion of an auxiliary linearized Stokes problem and $\eta_h:=w_h-v_h$ represents the error caused by the presence of the nonlinearity. The auxiliary function $w_h\in V_h$ is defined as the solution of the Stokes problem

$$\nu(\nabla w_h,\nabla\varphi_h)=(f,\varphi_h)-\tilde{n}(v,v,\varphi_h)\quad\forall\varphi_h\in V_h. \tag{4.2.99}$$

Clearly, the corresponding "continuous" solution is v itself. Therefore, by the error estimates provided above for the approximation of the Stokes problem, we have the estimate

$$\|\xi_h\| + h\|\nabla\xi_h\| \le c(\nu, \|f\|)\, h^2, \tag{4.2.100}$$

where $c(\nu, \|f\|)$ originates from the constant in the estimate

$$|(f,\varphi) - \tilde{n}(v,v,\varphi)| \le \|f - v\cdot\nabla v\|\|\varphi\| \le c(\nu,\|f\|)\|\varphi\|, \quad \varphi \in V,$$

Next, we estimate η_h. Combining the equations satisfied by w_h and v_h, we obtain

$$\nu(\nabla\eta_h, \nabla\varphi_h) = \tilde{n}(v_h, v_h, \varphi_h) - \tilde{n}(v,v,\varphi_h), \quad \varphi_h \in V_h.$$

and setting $\varphi_h = \eta_h$:

$$\begin{aligned}
\nu\|\nabla\eta_h\|^2 &= \tilde{n}(v_h, v_h, \eta_h) - \tilde{n}(v,v,\eta_h)\\
&= \tilde{n}(v_h - v, v_h, \eta_h) + \tilde{n}(v, v_h - w_h + w_h - v, \eta_h)\\
&= -\tilde{n}(e_h, v_h, \eta_h) - \tilde{n}(v, \xi_h, \eta_h)\\
&\le c_*^2\big(\|\nabla e_h\|\|\nabla v_h\| + \|\nabla v\|\|\nabla\xi_h\|\big)\|\nabla\eta_h\|.
\end{aligned}$$

This implies

$$\|\nabla\eta_h\| \le c_*^2\nu^{-2}\|f\|_{-1}\|\nabla\eta_h\| + 2c_*^2\nu^{-2}\|f\|_{-1}\|\nabla\xi_h\|.$$

By the small-data assumption $q := c_*\nu^{-2}\|f\|_{-1} < 1$, we conclude that

$$\|\nabla\eta\| \le \frac{2q}{1-q}\|\nabla\xi_h\|.$$

Combining this with the estimate for $\|\nabla\xi_h\|$ derived above yields the desired estimate for $\|\nabla e_h\|$.

(ii) To estimate the pressure error $\|p - p_h\|$, we use the assumed "inf-sup" stability of the spaces $H_h \times L_h$. There holds

$$\begin{aligned}
\|p - p_h\| &\le \|p - j_h p\| + \|j_h p - p_h\|\\
&\le \|p - j_h p\| + \beta_*^{-1} \sup_{\varphi_h\in H_h} \frac{(j_h p - p_h, \nabla\cdot\varphi_h)}{\|\nabla\varphi_h\|}\\
&= \|p - j_h p\| + \beta_*^{-1} \sup_{\varphi_h\in H_h} \frac{(j_h p - p, \nabla\cdot\varphi_h)}{\|\nabla\varphi_h\|} + \beta_*^{-1} \sup_{\varphi_h\in H_h} \frac{(p - p_h, \nabla\cdot\varphi_h)}{\|\nabla\varphi_h\|}\\
&\le (1 + c\beta_*^{-1})\|p - j_h p\| + \beta_*^{-1} \sup_{\varphi_h\in H_h} \frac{(p - p_h, \nabla\cdot\varphi_h)}{\|\nabla\varphi_h\|}.
\end{aligned}$$

Combining the continuous and discrete Navier-Stokes equations, we obtain for $\varphi_h \in H_h$:

$$\begin{aligned}
(p - p_h, \nabla\cdot\varphi_h) &= \nu(\nabla(v - v_h), \nabla\varphi_h) - \tilde{n}(v,v,\varphi_h) + \tilde{n}(v_h, v_h, \varphi_h)\\
&= \nu(\nabla(v - v_h), \nabla\varphi_h) - \tilde{n}(v - v_h, v, \varphi_h) + \tilde{n}(v_h, v_h - v, \varphi_h),
\end{aligned}$$

and further

$$\begin{aligned}(p-p_h,\nabla\cdot\varphi_h) &\le \nu\|\nabla(v-v_h)\|\|\nabla\varphi_h\| + c_*^2\big(\|\nabla v\|+\|\nabla v_h\|\big)\|\nabla(v-v_h)\|\|\nabla\varphi_h\| \\ &\le \big(\nu + 2c_*^2\nu^{-1}\|f\|_{-1}\big)\|\nabla(v-v_h)\|\|\nabla\varphi_h\| \\ &= \nu(1+2q)\|\nabla(v-v_h)\|\|\nabla\varphi_h\|.\end{aligned}$$

Combining the foregoing results, we obtain

$$\begin{aligned}\|p-p_h\| &\le (1+c\beta_*^{-1})\|p-j_hp\| + \nu(1+2q)\|\nabla(v-v_h)\| \\ &\le ch\|p\|_{H^1} + \nu(1+2q)\|\nabla(v-v_h)\| \\ &\le c(\nu,\|f\|)\,h.\end{aligned}$$

(iii) Finally, for proving the L^2-error estimate, we use again a duality argument. By the small-data assumption the bilinear form (derivative form at the solution v)

$$a'(v;\psi,\varphi) := \nu(\nabla\psi,\nabla\varphi) + \tilde n(v,\psi,\varphi) + \tilde n(\psi,v,\varphi)$$

is V-elliptic:

$$\begin{aligned}a'(v;\varphi,\varphi) &= \nu\|\nabla\varphi\|^2 + \tilde n(v,\varphi,\varphi) + \tilde n(\varphi,v,\varphi) \ge \nu\|\nabla\varphi\|^2 - |\tilde n(\varphi,v,\varphi)| \\ &\ge \nu\|\nabla\varphi\|^2 - c_*^2\|\nabla v\|\|\nabla\varphi\|^2 \ge (\nu - c_*^2\nu^{-1}\|f\|_{-1})\|\nabla\varphi\|^2 > 0.\end{aligned}$$

Therefore, the dual problem

$$\nu(\nabla\varphi,\nabla z) + \tilde n(v,\varphi,z) + \tilde n(\varphi,v,z) = (\varphi,e_h)\quad\forall\varphi\in V.$$

possesses a unique solution $z\in V$, which is in $H^2(\Omega)^2$ and satisfies the a priori estimate $\|z\|_{H^2}\le c\|e_h\|$. Setting $\varphi = e_h$ yields, with arbitrary $z_h\in V_h$:

$$\begin{aligned}\|e_h\|^2 &= \nu(\nabla e_h,\nabla z) + \tilde n(v,e_h,z) + \tilde n(e_h,v,z) \\ &= \big[\nu(\nabla e_h,\nabla(z-z_h)) + \tilde n(v,e_h,z-z_h) + \tilde n(e_h,v,z-z_h)\big] \\ &\quad + \big[\nu(\nabla e_h,\nabla z_h) + \tilde n(v,e_h,z_h) + \tilde n(e_h,v,z_h)\big] \\ &=: A + B.\end{aligned}$$

For the first term, we get by the usual arguments

$$|A| \le (\nu + 2c_*^2\nu^{-1}\|f\|_{-1})\|\nabla e_h\|\|\nabla(z-z_h)\|.$$

For the second term there holds

$$\begin{aligned}B &= \nu(\nabla v,\nabla z_h) + \tilde n(v,v,z_h) + \tilde n(v,v,z_h) \\ &\quad - \nu(\nabla v_h,\nabla z_h) - \tilde n(v,v_h,z_h) - \tilde n(v_h,v,z_h) \\ &= (f,z_h) + \tilde n(v,v,z_h) - (f,z_h) + \tilde n(v_h,v_h,z_h) - \tilde n(v,v_h,z_h) - \tilde n(v_h,v,z_h) \\ &= \tilde n(v-v_h,v-v_h,z_h),\end{aligned}$$

and, consequently,

$$|B| \le c_*^2\|\nabla e_h\|^2\big(\|\nabla(z_h-z)\| + \|\nabla z\|\big).$$

Taking now $z_h = i_h z$ and using the above a priori estimate for $\|z\|_{H^2}$, we obtain

$$\|e_h\|^2 \le c(\nu, \|f\|_{-1}) h \|\nabla e_h\| \|z\|_{H^2} + c\|\nabla e_h\|^2 \|z\|_{H^2} \le c(\nu, \|f\|_{-1}) \big(h\|\nabla e_h\| + \|\nabla e_h\|^2\big) \|e_h\|.$$

Together with the already proven estimate for $\|\nabla e_h\|$ this finally implies the L^2-error estimate

$$\|e_h\| \le c(\nu, \|f\|_{-1})\, h^2,$$

which completes the proof. Q.E.D.

4.2.4 Transport stabilization for more general data

In the case of "larger" data the Navier-Stokes problem takes on the structure of a diffusion-transport problem with possibly dominant transport. In the usual finite element discretization the transport term contributes mainly to the off-diagonals in the system matrix, by which this loses its definiteness property. On coarser meshes this can result in unphysical oscillations of the discrete solution and in the failure of iterative solution methods. In order to guarantee stability of the discretization, we need extra stabilization of the transport term. This aspect is illustrated first in a one-dimensional situation.

4.2.5 A prototypical example in 1D

On the one-dimensional domain $\Omega = I := (0,1) \in \mathbb{R}^1$, we consider the linear *singularly perturbed* boundary value problem (so-called "Sturm-Liouville Problem")

$$-\varepsilon u''(x) + q(x)u'(x) = 0, \;\; x \in I, \quad u(0) = 1, \; u(1) = 0. \tag{4.2.101}$$

In the case $q \equiv 1$ the unique solution has the form (see Fig. 4.5)

$$u^\varepsilon(x) = \frac{e^{1/\varepsilon} - e^{x/\varepsilon}}{e^{1/\varepsilon} - 1}.$$

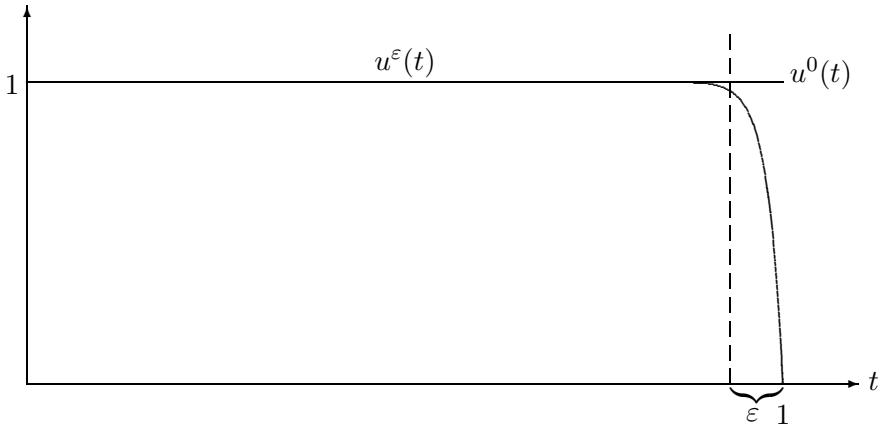


Figure 4.5: Solution of the singularly perturbed Sturm-Liouville problem for $\varepsilon = 0.1$.

In the case $\varepsilon \ll 1$, we have for $x = 1 - \delta$ and $\delta > \varepsilon$:

$$u^\varepsilon(1-\delta) = \frac{e^{1/\varepsilon}}{e^{1/\varepsilon}-1}\Big(1 - e^{-\delta/\varepsilon}\Big) \approx 1, \qquad \sup_{x\in I} |u^{\varepsilon\prime\prime}(x)| \approx \varepsilon^{-2},$$

what justifies the term "boundary layer solution". For $\varepsilon \to 0$ the limit solution is $u^0 \equiv 1$, which does not satisfy the boundary condition at $x = 1$. The approximation of this problem with the usual centered finite difference scheme (in 1D equivalent to piecewise linear finite elements) with equidistant mesh size $h = 1/(N+1)$ results in

$$-(\varepsilon + \tfrac{1}{2}h)y_{n-1} + 2\varepsilon y_n - (\varepsilon - \tfrac{1}{2}h)y_{n+1} = 0, \quad 1 \le n \le N, \quad y_0 = 1,\ y_{N+1} = 0.$$

The corresponding coefficiant matrix is diagonally dominant only under the restrictive condition

$$h \le 2\varepsilon. \tag{4.2.102}$$

For $h > 2\varepsilon$ the discrete solution shows an unphysical behavior. To see this, we make the ansatz $y_n = \lambda^n$. The possible values for λ are the roots $\lambda_\pm$ of the quadratic equation

$$\lambda^2 + \frac{2\varepsilon}{\frac{1}{2}h - \varepsilon}\lambda + \frac{\frac{1}{2}h + \varepsilon}{\frac{1}{2}h - \varepsilon} = 0.$$

Incorporating the boundary conditions $y_0 = 1$ and $y_{N+1} = 0$, we obtain

$$y_n = c_+\lambda_+^n + c_-\lambda_-^n,$$

and for the coefficients the realtions $c_+ + c_- = 1$ and $c_+\lambda_+^{N+1} + c_-\lambda_-^{N+1} = 0$, and, consequently,

$$c_- = \frac{\lambda_+^{N+1}}{\lambda_+^{N+1} - \lambda_-^{N+1}}, \qquad c_+ = 1 - \frac{\lambda_+^{N+1}}{\lambda_+^{N+1} - \lambda_-^{N+1}} = -\frac{\lambda_-^{N+1}}{\lambda_+^{N+1} - \lambda_-^{N+1}}.$$

Hence, the solution looks like

$$y_n = \frac{\lambda_+^{N+1}\lambda_-^n - \lambda_-^{N+1}\lambda_+^n}{\lambda_+^{N+1} - \lambda_-^{N+1}}, \quad n = 0, \dots, N+1. \tag{4.2.103}$$

In the present case these roots are given by

$$\lambda_{+,-} = \frac{-\varepsilon \pm \sqrt{\varepsilon^2 + (\frac{1}{2}h + \varepsilon)(\frac{1}{2}h - \varepsilon)}}{\frac{1}{2}h - \varepsilon} = \frac{\varepsilon \mp \frac{1}{2}h}{\varepsilon - \frac{1}{2}h}, \qquad \lambda_+ = 1,\ \ \lambda_- = \frac{\varepsilon + \frac{1}{2}h}{\varepsilon - \frac{1}{2}h}.$$

For $\varepsilon \ll \frac{1}{2}h$. we have $\lambda_- \approx -1$ and there results an oscillatory solution:

$$y_n = \frac{\lambda_-^n - \lambda_-^{N+1}}{1 - \lambda_-^{N+1}}, \quad n = 0, \dots, N+1,$$

which does not show the qualitatively correct form of the continuous solution. To suppress this defect different approaches are available.

(i) Upwind Discretization: The first-order term $u'(x)$ in the differential equation is discretized by one of the following one-sided difference quotients

$$\Delta_h^+ u(x) = \frac{u(x+h) - u(x)}{h}, \quad \Delta_h^- u(x) = \frac{u(x) - u(x-h)}{h}.$$

The choice of the *backward* difference quotient Δ_h^- observes the physical transport from the left to the right (see the form of the limit solution $u^0(x)$). This results in the difference equation

$$(-\varepsilon + h)y_{n-1} + (2\varepsilon + h)y_n - \varepsilon y_{n+1} = 0.$$

For arbitrary $h > 0$ the corresponding system matrix is diagonally dominant (even an M-matrix). In this case the ansatz $y_n \lambda^n$ leads to the quadratic equation

$$\lambda^2 - \frac{2\varepsilon + h}{\varepsilon}\lambda + \frac{\varepsilon + h}{\varepsilon} = 0,$$

with the roots

$$\lambda_{+,-} = \frac{2\varepsilon + h}{2\varepsilon} \pm \sqrt{(2\varepsilon + h)^2 - 4\varepsilon(\varepsilon + h)} = \frac{2\varepsilon + h \pm h}{2\varepsilon}, \qquad \lambda_+ = \frac{\varepsilon + h}{\varepsilon}, \ \ \lambda_- = 1.$$

The critical root λ_+ is now always positive, such that in the discrete solution (4.2.103),

$$y_n = \frac{\lambda_+^{N+1} - \lambda_+^n}{\lambda_+^{N+1} - 1},$$

no oscillations occur. For general transport coefficient $q(x)$ the "upwinding" has to be chosen locally in accordance to the sign of $q_n = q(x_n)$. This way of discretizing the transport term $u'(x)$ by one-sided difference quotients is called "backward differencing" or "upwinding". It limits the total accuracy of the discretization to first order only even in regions outside the boundary layer where the solution is smooth.

(ii) Artificial Diffusion: Maintaining the central difference approximation of the transport term $u'(x)$ the diffusion coefficient ε is set to a larger value $\varepsilon_h := \varepsilon + \delta h$. This results in the difference equation

$$-(\varepsilon_h + \tfrac{1}{2}h)y_{n-1} + 2\varepsilon_h y_n - (\varepsilon_h - \tfrac{1}{2}h)y_{n+1} = 0, \quad 1 \le n \le N.$$

For the corresponding discrete solution, we obtain again by the above ansatz the form

$$y_n = \frac{\lambda_+^{N+1} - \lambda_+^n}{\lambda_+^{N+1} - 1}, \quad \lambda_+ = \frac{\varepsilon_h + \frac{1}{2}h}{\varepsilon_h - \frac{1}{2}h}.$$

Obviously, in this case $\lambda_+ > 0$ for $\varepsilon + \delta h > \frac{1}{2}h$, i. e., for the choice $\delta \ge \frac{1}{2}$. By this modification, we obtain again a diagonally dominant (M-matrix) and, consequently, a stable discretization. However, this approach strongly smears out the boundary layer on the interval $[1 - \varepsilon_h, 1]$ and the total accuracy of approxination is also limited to first order.

Remark 4.10: The two described strategies for dealing with the lacking stability in case of dominant transport, simple "upwinding" and "artificial diffusion", are able to cure the problem but at the expense of limiting the approximation accuracy to first order. Other related

approaches of formally higher oder (such as higher-order one-sided finite differences or higher-order artificial diffusion) do not preserve the strong M-matrix property of the system matrix. This can only be achieved by low-order stabilization techniques.

(iii) Streamline Diffusion (SD-FEM): In the context of finite element discretization there is an approach which can be viewed as consistent, transport-oriented artificial diffusion. We consider the model problem

$$-\varepsilon u''(x) + qu'(x) + \alpha u(x) = f(x), \quad x \in I, \quad u(0) = u(1) = 0, \tag{4.2.104}$$

with $q \equiv 1$ and $\alpha \geq 0$. In the so-called "streamline diffusion method" the variational formulation of this boundary value problem,

$$\varepsilon(u', \varphi') + (u' + \alpha u, \varphi) = (f, \varphi), \quad \forall \varphi \in H := H_0^1(I), \tag{4.2.105}$$

is modified to

$$\varepsilon(u', \varphi') + (u' + \alpha u, \varphi + \delta\varphi') = (f, \varphi + \delta\varphi'), \quad \forall \varphi \in H, \tag{4.2.106}$$

with a parameter function δ, which is coupled to the local mesh size h like $0 < \delta \sim h \leq 1$ such that $\alpha\delta \leq 1$.

Lemma 4.10: *The (nonsymmetric) bilinear form*

$$a_\delta(u, v) := \varepsilon(u', v') + (u' + \alpha u, v + \delta v'), \quad u, v \in H,$$

is "elliptic",

$$a_\delta(v, v) \geq \tfrac{1}{2}\|v\|_\delta^2, \quad v \in H, \tag{4.2.107}$$

with respect to the "energy norm"

$$\|v\|_\delta := \big(\varepsilon\|v'\|^2 + \|\delta^{1/2}v'\|^2 + \alpha\|v\|^2\big)^{1/2},$$

Proof: We have

$$\begin{aligned} a_\delta(v, v) &= \varepsilon\|v'\|^2 + (v' + \alpha v, v + \delta v') = \varepsilon\|v'\|^2 + \|\delta^{1/2}v'\|^2 + \alpha\|v\|^2 + (v', v) + \alpha(v, \delta v') \\ &\geq \|v\|_\delta^2 - |(v', v)| - \alpha|(v, \delta v')| \\ &\geq \|v\|_\delta^2 - \tfrac{1}{2}\alpha\|v\|^2 - \tfrac{1}{2}\|\delta^{1/2}v'\|^2 \geq \tfrac{1}{2}\|v\|_\delta^2. \end{aligned}$$

Here, we have used the assumption $\alpha\delta \leq 1$ and the identity

$$\int_0^1 v'(x)v(x)\,dx = \frac{1}{2}\int_0^1 (v^2)'(x)\,dx = \frac{1}{2}\big\{v^2(1) - v^2(0)\big\} = 0.$$

This completes the proof. Q.E.D.

We emphasize that the parameter δ is a function of x (cellwise constant with respect to $0 = x_0 < \ldots < x_{N+1} = 1$) and, consequently, appears inside $\|\delta^{1/2}v'\|$. Analogously, we use the symbol $h = h(x)$ for a cellwise constant mesh-size function. The corresponding finite element

Galerkin method (with linear elements) in the subspaces $H_h \subset H = H_0^1(0,1)$ reads

$$u_h \in H_h : \quad a_\delta(u_h, \varphi_h) = l_\delta(\varphi_h) \quad \forall \varphi_h \in H_h, \tag{4.2.108}$$

with the modified functional $l_\delta(\varphi_h) := (f, \varphi_h + \delta\varphi_h')$.

Theorem 4.8 (Error estimate for SD-FEM): *Suppose that $\varepsilon \ll h_{\min}$, $\alpha = 1$, and that accordingly the stabilization parameter in the streamline diffusion in each subinterval I_n is chosen as $\delta_n \sim h_n$. Then, there holds the a priori error estimate*

$$\|u - u_h\|_\delta \le c\|h^{3/2}u''\|, \tag{4.2.109}$$

with a constant c independent of ε, h and δ.

Proof: We sketch the proof for the case $\alpha = 1$ and $\delta = h$. Combination of the variational equations for u and u_h yields the following perturbed Galerkin orthogonality relation for the error $e := u - u_h$:

$$\begin{aligned} a_\delta(e, \varphi_h) &= \varepsilon(u', \varphi_h') + (u' + u, \varphi_h + \delta\varphi_h') - \varepsilon(u_h', \varphi_h') - (u_h' + u_h, \varphi_h + \delta\varphi_h') \\ &= (f, \varphi_h) + (u' + u, \delta\varphi_h') - (f, \varphi_h + \delta\varphi_h') \\ &= (u' + u - f, \delta\varphi_h') \\ &= \varepsilon(u'', \delta\varphi_h'). \end{aligned} \tag{4.2.110}$$

With help of the ellipticity (4.2.107) and this orthogonality relation, we obtain:

$$\|e\|_\delta^2 \le a_\delta(e, u - \varphi_h) + \varepsilon(u'', \delta(\varphi_h - u_h)'), \tag{4.2.111}$$

with arbitrary $\varphi_h \in H_h$. The first term on the right is estimated by

$$\begin{aligned} |a_\delta(e, u - \varphi_h)| &\le \varepsilon|(e', (u - \varphi_h)')| + |(e' + e, u - \varphi_h)| + |(e' + e, \delta(u - \varphi_h)')| \\ &\le \varepsilon\|e'\|\|(u - \varphi_h)'\| + \{\|\delta^{1/2}e'\| + \|\delta^{1/2}e\|\}\|\delta^{-1/2}(u - \varphi_h)\| \\ &\quad + \{\|\delta^{1/2}e'\| + \|\delta^{1/2}e\|\}\|\delta^{1/2}(u - \varphi_h)'\|. \end{aligned}$$

For the choice $\varphi_h := i_h u$ it follows with help of the usual local interpolation estimate observing $\delta = h \le 1$ and $\varepsilon \le h_{\min}$:

$$\begin{aligned} |a_\delta(e, u - i_h u)| &\le c\varepsilon\|e'\|\|hu''\| + c\{\|\delta^{1/2}e'\| + \|e\|\}\|\delta^{-1/2}h^2u''\| \\ &\quad + c\{\|\delta^{1/2}e'\| + \|e\|\}\|\delta^{1/2}hu''\| \\ &\le c\big\{\varepsilon^{1/2}\|e'\| + \|\delta^{1/2}e'\| + \|e\|\big\}\|h^{3/2}u''\| \\ &\le \tfrac{1}{4}\|e\|_\delta^2 + c\|h^{3/2}u''\|^2. \end{aligned}$$

For the second term in (4.2.111), we get with $\varphi_h := i_h u$ by analogous arguments:

$$\begin{aligned} \varepsilon|(u'', \delta(i_h u - u_h)')| &\le \varepsilon\|\delta^{1/2}u''\|\big\{\|\delta^{1/2}(i_h u - u)'\| + \|\delta^{1/2}e'\|\big\} \\ &\le c\|h^{3/2}u''\|\big\{\|\delta^{1/2}hu''\| + \|\delta^{1/2}e'\|\big\} \\ &\le c\|h^{3/2}u''\|^2 + \tfrac{1}{4}\|e\|_\delta^2. \end{aligned}$$

Now, combination of the derived estimates implies the desired result. Q.E.D.

The error estimate (4.2.109) shows that for a smooth solution u (without boundary layer), or if the mesh is fine enough for resolving the boundary layer, the SD-FEM converges in the "energy norm" with order $\mathcal{O}(h^{3/2})$, which is higher than the order achievable by simple upwinding or artificial diffusion. However, the corresponding (nonsymmetriuc) system matrix is definite but neither diagonally dominant nor an M-matrix.

Transport stabilization for the Navier-Stokes equations

The methods for transport stabilization described above for the 1d case can also be applied in the context of the multi-dimensional nonlinear Navier-Stokes equations. We discuss here only the potentially higher-order SD-FEM and other approaches of this type, which are the most commonly used techniques within finite element discretizations.

(i) Streamline Diffusion (SDS): The idea of "streamline diffusion" is to add artificial diffusion only in transport (streamline) direction. This can be accomplishes in two different but essentially equivalent ways:

a) by augmenting the test functions by transport-oriented terms, what leads to a so-called "Petrov-Galerkin method", or

b) by adding certain "least-squares" terms in the variational formulation of the problem.

We describe a simple variant of this method for the stationary Navier-Stokes problem this time allowing nonhomogeneous inflow and outflow data $v|_{\Gamma_{\rm in}\cup\Gamma_{\rm out}} = v^{\rm in}$. The discretization may be by any of the conforming, nonconforming or stabilized Stokes elements discussed above. For that, we introduce the following notation for the main terms in the variational Navier-Stokes problem:

$$\begin{aligned} a_h(v_h,\varphi_h) &:= \nu(\nabla_h v_h, \nabla_h \varphi_h), \\ \tilde{n}_h(v_h,v_h,\varphi_h) &:= \tfrac{1}{2}\big\{(v_h\cdot\nabla_h v_h,\varphi_h) - (v_h\cdot\nabla_h\varphi_h, v_h)\big\}, \\ b(p_h,\varphi_h) &:= (p_h, \nabla_h\cdot\varphi_h), \end{aligned}$$

and for the additional stabilization terms:

$$\begin{aligned} s_h^p(p_h,\chi_h)) &:= \sum_{T\in\mathbb{T}_h} \delta_T^p(\nabla p_h, \nabla\chi_h)_T \\ s_h^v(v_h,\varphi_h) &:= \sum_{T\in\mathbb{T}_h} \delta_T^v\big\{(v_h\cdot\nabla v_h, \bar{v}_h\cdot\nabla\varphi_h)_T + (\nabla\cdot v_h, \nabla\cdot\varphi_h)_T\big\}, \\ r_h^p(v_h,\chi_h) &:= \sum_{T\in\mathbb{T}_h} \delta_T^p(f + [\nu\Delta v_h] - v_h\cdot\nabla v_h, \nabla\chi_h)_T, \\ r_h^v(v_h,p_h,\varphi_h) &:= \sum_{T\in\mathbb{T}_h} \delta_T^v(f + [\nu\Delta v_h + \nabla p_h], \bar{v}_h\cdot\nabla\varphi_h)_T, \end{aligned}$$

where $\bar{v}_h$ is a suitable reference velocity. The stabilization parameters are chosen as

$$\delta_T^p = \alpha_T^p \nu^{-1} h_T^2, \qquad \delta_T^v = \alpha_T^v \max_T |\bar{v}_h|^{-1} h_T,$$

with suitable damping constants α_T^p, α_T^v, usually chosen uniformly or adaptively in the range $[0.1, 1]$ for all mesh cells for balancing the effect of stabilization and approximation. Then, the discrete problems read as follows:

Find pairs $\{v_h, p_h\} \in (v_h^{\text{in}} + H_h) \times L_h$, such that

$$\begin{aligned} a_h(v_h,\varphi_h) + \tilde{n}_h(v_h.v_h,\varphi_h) + s_h^v(v_h,\varphi_h) + b_h(p_h,\varphi_h) & \\ = (f,\varphi_h) + r_h^v(v_h,p_h,\varphi_h) \quad \forall \varphi_h \in H_h, & \qquad (4.2.112) \\ b_h(\chi_h,v_h) + s_h^p(p_h,\chi_h) = r_h^p(v_h,\chi_h) \quad \forall \chi_h \in L_h. & \qquad (4.2.113) \end{aligned}$$

This discretization is fully consistent with the continuous Navier-Stokes problem since the combined stabilization and correction terms vanish for the continuous solution $\{v, p\}$. On the one hand, this ensures exact Galerkin orthogonality for the errors $\{v - v_h, p - p_h\}$, but, on the other hand, due to the many extra terms, it makes the computation of the system matrices very expensive and introduces additional couplings between velocity and pressure unknowns, what significantly increases the cost of the iterative solution of the resulting algebraic systems. The existence of solutions $\{v_h, p_h\} \in (v_h^{\text{in}} + H_h \times L_h$ of the system (4.2.112) - (4.2.113) follows analogously as in the non-stabilized case due to the definiteness and smallness of the stabilization terms. To summerize: The above stabilized discretization serves the following purposes:

– The term
$$\sum_{T\in\mathbb{T}_h} \delta_T^v (v_h \cdot \nabla v_h, v_h \cdot \nabla \varphi_h)_T$$
stabilizes the transport term in the sense of the directional "streamline diffusion".

– The term
$$\sum_{T\in\mathbb{T}_h} \delta_T^v (\nabla \cdot v_h, \nabla \cdot \varphi_h)_T$$
enhances the incompressibility condition.

– The term
$$\sum_{T\in\mathbb{T}_h} \delta_T^p (\nabla p_h, \nabla \chi_h)$$
stabilizes the pressure in case of an unstable "equal-order" Stokes element.

The correction terms only make the discretization consistent. A careful but very technical analysis shows that this discretization actually yields an improved error behavior compared to the simple low-order stabilization by "upwinding" or "artificial diffusion".

(ii) Local Projection Stabilization (LPS): An alternative approach to transport stabilization uses the concept of "local projection" as already employed above in the context of pressure stabilization. The method avoids some of the drawbacks of the streamline diffusion technique (unphysical outward bending of streamlines using the "do nothing" condition at free outflow boundary and high extra costs for the evaluation of the stabilization terms particularly in 3D). In the LPS the stabilization terms

$$\begin{aligned} s_h^p &:= \sum_{T\in\mathbb{T}_h} \delta_T^p (\nabla(p_h - \pi_{2h} p_h), \nabla(\chi_h - \pi_{2h}\chi_h))_T, \\ s_h^v &:= \sum_{T\in\mathbb{T}_h} \delta_T^v (v_h \cdot \nabla(v_h - \pi_{2h} v_h), v_h \cdot \nabla(\varphi - \pi_{2h}\varphi))_T, \end{aligned}$$

are used and there is no need for additional correction terms, i. e., $r_h^p = r_h^v = 0$. Here, π_{2h} is a local projection or interpolation into the spaces L_{2h} and H_{2h} on the coarser mesh $\mathbb{T}_{2h}$,

respectively. The stabilization parameters δ_T^p and δ_T^v are chosen similarly as in the streamline diffusion method. The resulting discretization is formally of second order accurate, the generation of the corresponding system matrices is relatively cheap, and the consistency deffect at outstream boundaries is avoided.

4.2.6 Treatment of nonlinearity

For dealing with the nonlinearity in the discretized Navier-Stokes problem there are several possible approaches following the concepts already described above on the continuous level. For notational symplicity, we describe these methods only for the case possibly with simple pressure stabilization (without correction terms) but without transport stabilization. In all these iterative methods one starts from a suitable field v_h^0 and generates a sequence of iterates $\{v_h^t, p_h^t\} \in H_h \times L_h$, which under certain conditions converges for $t \to \infty$ to the discrete solution $\{v_h, p_h\}$.

(i) Stokes linearization (explicit iteration)
In case of dominant diffusion $\nu \approx 1$ (e. g., for highly viscous liquids, small velocities, or small flow domains), one may use the simple "Stokes iteration"

$$\nu(\nabla_h v_h^t, \nabla_h \varphi_h) - (p_h^t, \nabla_h \cdot \varphi_h) = (f, \varphi_h) - \tilde{n}(v_h^{t-1}, v_h^{t-1}, \varphi_h) \quad \forall \varphi_h \in H_h, \tag{4.2.114}$$

$$(\chi_h, \nabla_h \cdot v_h^t) + s_h^p(\chi_h, p_h^t) = 0 \quad \forall \chi_h \in L_h, \tag{4.2.115}$$

in which the nonlinearity is treated fully explicitly. In each iteration step only a discrete symmetric and positive definite Stokes problem has to be solved. As on the continuous level this iteration converges under a small-data assumption ($q := c_*^2 \nu^{-2} \|f\|_{-1} < 1/2$ for conforming Stokes elements), if additionally the starting value is already sufficiently accurate, $\|\nabla_h(v_h^0 - v_h)\| \leq \nu(1-2q)/(2c_*^2) < 1$ (exercise). These convergence conditions are rather severe and usually not met in practice.

(ii) Oseen linearization (semi-implicit iteration)
The restrictive smallness condition for the starting value in the "Stokes iteration" can be avoided by using the semi-implicit "Oseen linearization" leading to the following functional iteration:

$$\nu(\nabla_h v_h^t, \nabla_h \varphi_h) + \tilde{n}(v_h^{t-1}, v_h^t, \varphi_h) - (p_h^t, \nabla_h \cdot \varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in H_h, \tag{4.2.116}$$

$$(\chi_h, \nabla_h \cdot v_h^t) + s_h^p(\chi_h, p_h^t) = 0 \quad \forall \chi_h \in L_h. \tag{4.2.117}$$

In the small-data case ($q = c_*^2 \nu^{-2} \|f\|_{-1} < 1$ for conforming, "inf-sup" stable Stokes elements), this iteration converges for any starting value v_h^0 with linear rate uniformly for h (exercise). In each iteration step a linear but nonsymmetric "Oseen problem" has to be solved. This usually requires additional transport stabilization, e. g., the LPS method with terms of the form

$$s_h^v(v_h^{t-1}, v_h^t, \varphi_h) := \sum_{T \in \mathbb{T}_h} \delta_h^v (v_h^{t-1} \cdot \nabla(v_h^t - \pi_{2h} v_h^t), v_h^{t-1} \cdot \nabla(\varphi - \pi_{2h}\varphi))_T$$

(iii) Newton linearization
The linearization by applying the Newton method leads to an iteration, which in the small-data case (again $q = c_*^2 \nu^{-2} \|f\|_{-1} < 1$ for conforming Stokes elements) converges quadratically uniformly in h provided the starting value is good enough. Suitable starting values may be obtained by the globally convergent functional iteration. However, in practice the Newton

method is observed to converge even in cases of more general "larger" data provided the starting value is chosen sufficiently close to the wanted (only locally unique) solution. In setting up the Newton iteration, one usually only consideres the derivative of the Navier-Stokes form and neglects the likewise nonlinear terms in the transport stabilization, since these contain the factors δ_T^v and are therefore considered as "small". The resulting iteration then reads as follows:

$$\begin{aligned} \nu(\nabla_h v_h^t, \nabla_h \varphi_h) + \tilde{n}(v_h^{t-1}, v_h^t, \varphi_h) + \tilde{n}(v_h^t, v_h^{t-1}, \varphi_h) - (p_h^t, \nabla_h \cdot \varphi_h) \\ = (f, \varphi_h) + \tilde{n}(v_h^{t-1}, v_h^{t-1}, \varphi_h) \quad \forall \varphi_h \in H_h, \end{aligned} \tag{4.2.118}$$

$$(\chi_h, \nabla_h \cdot v_h^t) + s_h^p(\chi_h, p_h^t) = 0 \quad \forall \chi_h \in L_h. \tag{4.2.119}$$

In each iteration step, one has to solve a linear but nonsymmetric and, in general, indefinite Oseen-like problem. The indefiniteness of the "reaction term" $\tilde{n}(v_h^t, v_h^{t-1}, \varphi_h)$ may cause the usual iterative algebraic methods to fail and can make the practical use of this method a very difficult task. However, dropping this indefinite term reduces the Newton method to the ordinary functional iteration, which converges only with linear rate and may not be capable to approximate solutions, which are only locally unique.

4.2.7 Solution of linear discrete problems

We now discuss the solution of the linear discrete problems occurring within the nonlinear iterations described above. To this end, we consider again the discretization by "inf-sup" stable or stabilized (conforming or nonconforming) Stokes elements as presented above. Further, we restrict us again to the case of pure homogeneous Dirichlet boundary conditions, $\partial\Omega = \Gamma_{\text{rigid}}$. Let

$$\{\psi_h^i,\, i = 1, \ldots, N_H := \dim H_h\}, \quad \{\chi_h^i,\, i = 1, \ldots, N_L := \dim L_h\},$$

be the usual nodal bases of the velocity space H_h and the pressure space L_h, respectively. Then, the linear problems to be solved take on the form of a block system:

$$\mathcal{A}\xi = \begin{bmatrix} A + N_1 + N_2 & B \\ -B^T & C \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} b + n_0 + n_2 \\ 0 \end{bmatrix} =: \beta, \tag{4.2.120}$$

for the nodal value vectors $\xi = \{x, y\}$ in the representations

$$v_h = \sum_{j=1}^{N_H} x_j \psi_h^j, \qquad p_h = \sum_{j=1}^{N_L} y_j \chi_h^j,$$

with the corresponding system matrices and right hand side vectors:

$$\begin{aligned} A &:= \big(\nu(\nabla_h \psi_h^i, \nabla_h \psi_h^j)\big)_{i,j}^{N_H}, \quad B := -\big((\chi_h^i, \nabla_h \cdot \psi_h^j)\big)_{i,j=1}^{N_L, N_H}, \\ N_1 &:= \big(\tilde{n}(v_h^{t-1}, \psi_h^i, \psi_h^j)\big)_{i,j}^{N_H}, \quad N_2 = \big(\tilde{n}(\psi_h^i, v_h^{t-1}, \psi_h^j)\big)_{i,j}^{N_H}, \\ C &:= \Big(\sum_{T \in \mathbb{T}_h} \delta_T^p \big(\nabla_h(\chi_h^i - \pi_{2h}\chi_h^i), \nabla_h(\chi_h^j - \pi_{2h}\chi_h^j)\big)_T\Big)_{i,j=1}^{N_L} \\ b &:= \big((f, \psi_h^i)\big)_{i=1}^{N_H}, \\ n_0 &:= -\big(\tilde{n}(v_h^{t-1}, v_h^{t-1}, \psi_h^i)\big)_{i=1}^{N_H}, \quad n_2 = \big(\tilde{n}(v_h^{t-1}, v_h^{t-1}, \psi_h^i)\big)_{i=1}^{N_H}. \end{aligned}$$

This block-system has the typical structure of a saddle point problem, in which the matrix $\mathcal{A}$ is strongly indefinite with positive as well as negative eigenvalues in the symmetric case. Here, the Stokes iteration corresponds to $N_1 = N_2 = 0$ and $n_2 = 0$, the Oseen (functional) iteration to $N_2 = 0$ and $n_1 = n_2 = 0$, and the Newton iteration to $n_1 = 0$. The mean value condition on the pressure, $(p_h, 1) = 0$, which ensures uniqueness, may be directly built into the nodal basis of L_h, leading to a global coupling of all pressure unknowns. Alternatively, it can be incorporated through an extra equation in the system $(p_h, 1) = \sum_{i=1}^{N_L} y_i(\chi_h^i, 1) = 0$, which however, spoils the symmetry of the matrix $\mathcal{A}$. Therefore, in the following discussion, we prefer the first option.

4.2.8 Schur complement methods

For simplicity, we use the same notation b for the right hand side vector in all three cases $b + n_0$ (Stokes iteration), b (Oseen functional iteration) and $b + n_2$ (Newton iteration). Then, the discretized Navier-Stokes problem reads

$$(A + N_1 + N_2)x + By = b, \tag{4.2.121}$$

$$-B^T x + Cy = 0. \tag{4.2.122}$$

The matrix $A + N_1 + N_2$ is nonsymmetric but usually regular. Consequently, the component x can be eliminated from the system as follows:

$$x = -(A + N_1 + N_2)^{-1}By + (A + N_1 + N_2)^{-1}b,$$
$$\Sigma y := (B^T(A + N_1 + N_2)^{-1}B + C)y = B^T(A + N_1 + N_2)^{-1}b.$$

The matrix $\Sigma := B^T(A + N_1 + N_2)^{-1}B + C$ is called "Schur complement" of $A + N_1 + N_2$ in the block matrix $\mathcal{A}$. It leads to the following block triangular decomposition:

$$\mathcal{A} = \begin{bmatrix} A + N_1 + N_2 & B \\ -B^T & C \end{bmatrix} = \begin{bmatrix} A + N_1 + N_2 & 0 \\ -B^T & \Sigma \end{bmatrix} \begin{bmatrix} I & (A + N_1 + N_2)^{-1}B \\ 0 & I \end{bmatrix}. \tag{4.2.123}$$

This suggests some iterative procedures, which are based on the fact that very efficient solution methods are available for the "inversion" of the main-diagonal block $A + N_1 + N_2$.

a) Uzawa Algorithm:

The "classical" method for iteratively solving the saddle point problems (4.2.120) is the "Uzawa iteration", which operates on the pressure variables y. Starting from some y^0, satisfying the mean value condition, one successively computes for $t \geq 1$:

$$(A + N_1 + N_2)x^t = b - By^{t-1}, \tag{4.2.124}$$

$$y^l = y^{t-1} + \theta(B^T x^l - Cy^{t-1}), \tag{4.2.125}$$

where $\theta > 0$ is a appropriately chosen relaxation parameter. In order to cope with possible irregularitied of the mesh $\mathbb{T}_h$ (cell anisotropies or local mesh refinements) the system is "preconditioned" by the mass matrix of the pressure space.

$$M = M_L = \big((\chi_h^i, \chi_h^j)_{i,j=1}^{N_L}.$$

The modified Uzawa iteration then reads as follows:

$$(A + N_1 + N_2)x^t = b - By^{t-1}, \tag{4.2.126}$$

$$My^l = My^{t-1} + \theta(B^T x^t - Cy^{t-1}), \tag{4.2.127}$$

Each iteration step requires essentially the "inversion" of the matrix $A + N_1 + N_2$ and of the mass matrix M, what can be achieved by variants of the CG-method for nonsymmetric or indefinite matrices or by a multigrid method. Eliminating the velocity variable x^l converts Uzawa algorithm into a fixed point iteration of the form:

$$\begin{aligned} y^t &= y^{t-1} + \theta M^{-1}\big(B^T(A + N_1 + N_2)^{-1}(b - By^{t-1}) - Cy^{t-1}\big) \\ &= (I - \theta M^{-1}\Sigma)y^{t-1} + \theta M^{-1}(B^T(A + N_1 + N_2)^{-1}b). \end{aligned}$$

Hence, the Uzawa algorithm can be interpreted as a damped Richardson iteration for solving the Schur complement equation. For this, we obtain in the simplest case $N_1 = N_2 = 0$ (i. e., solution of Stokes problem) by the Banach fixed point theorem the following result.

Theorem 4.9 (Uzawa Algorithm): *Let Σ be symmetric. For sufficiently small damping, $\theta < \lambda_{\max}(M^{-1}\Sigma)^{-1}$, the Uzawa algorithm converges to the solution $\{x, y\}$ of the saddle point problem (4.2.120). With*

$$0 < 1 - \frac{\lambda_{\min}(M^{-1}\Sigma)}{\lambda_{\max}(M^{-1}\Sigma)} = 1 - \frac{1}{\mathrm{cond}_2(M^{-1}\Sigma)} =: q < 1,$$

there holds the error estimate

$$|y^l - y| \le q^t |y^0 - y|, \quad t \ge 1. \tag{4.2.128}$$

Proof: The exact solution y satisfies the fixed point equation

$$y = (I - \theta M^{-1}\Sigma)y + \theta M^{-1}(B^T A^{-1} b).$$

For the iteration error $e^t := y - y^t$ there holds

$$e^t = (I - \theta M^{-1}\Sigma)e^{t-1}, \quad t \ge 1.$$

Consequently, we have convergence for arbitrary starting point y^0 if and only if the spectral radius of the iteration matrix satisfies $\mathrm{spr}(I - \theta M^{-1}\Sigma) < 1$. This is guaranteed by the choice of the relaxation parameter $\theta < \lambda_{\max}(M^{-1}\Sigma)^{-1}$. The representation

$$M^{-1}\Sigma = M^{-1/2}(M^{-1/2}\Sigma M^{-1/2})M^{1/2}$$

implies by a similarity transformation that the eigenvalues of the matrix $M^{-1}\Sigma$ equal those of the symmetric and positive matrix $M^{-1/2}\Sigma M^{-1/2}$. Consequently, all eigenvalues of $M^{-1}\Sigma$ are real and positive and, by the choice of θ there holds

$$\mathrm{spr}(I - \theta M^{-1}\Sigma) = \max\{1 - \theta\lambda(M^{-1}\Sigma)\} \le 1 - \frac{\lambda_{\min}(M^{-1}\Sigma)}{\lambda_{\max}(M^{-1}\Sigma)} < 1.$$

The assertion then follows from the general results for fixed point iterations. Q.E.D.

For the symmetric case $N_1 = N_2 = 0$, we shall see below that $\mathrm{cond}_2(M^{-1}\Sigma) \le 1$ uniformly for all h. Therefore the Uzawa algorithm converges linearly with mesh-independent rate provided that the relaxation parameter θ is chosen sufficiently small. By varible choice of $\theta = \theta_t$ the speed of convergence can be optimized. Then, the Uzawa algoithm corresponds to the ordinary "gradient method" applied to the Schur complement equation. We do not consider this in more detail since, next, we shall look at the much faster "conjugate gradient (CG) method". The Uzawa algorithm may also be applied in the case $N_1 \neq 0$ (Oseen functional iteration) as then the Schur complement $M^{-1}\Sigma$ is definite (though nonsymmetric). However, in the case $N_2 \neq 0$ (Newton iteration) it may fail to converge.

b) CG-type methods

We consider again the symmetric case $\Sigma = \Sigma^T$. As in the Uzawa algorithm the system matrix is preconditioned with the pressure mass matrix M_L, i. e., the CG method is used in form of a PCG method for the system

$$M^{-1}\Sigma y = M^{-1}(B^T A^{-1} b). \tag{4.2.129}$$

This is equivalent to applying the CG method to the symmetric and positive definite matrix (modulo the mean value zero condition) $M^{-1/2}\Sigma M^{-1/2}$. The speed of the CG method is determined by the spectral condition number

$$\kappa := \mathrm{cond}_2(M^{-1}\Sigma) = \frac{\lambda_{\max}(M^{-1}\Sigma)}{\lambda_{\min}(M^{-1}\Sigma)}$$

like

$$|y^t - y| \le \kappa \Big(\frac{1-\kappa^{-1/2}}{1+\kappa^{-1/2}}\Big)^t |y^0 - y|, \quad t \in \mathbb{N}, \tag{4.2.130}$$

where y^0 is the starting value of the iteration.

Theorem 4.10 (Schur complement): *For the Schur complement $\Sigma = B^T A^{-1} B$ there holds*

$$\mathrm{cond}_{\mathrm{nat}}(M^{-1}\Sigma) \le \frac{c_0}{\beta_h^2}, \tag{4.2.131}$$

with the constant $\beta_h > 0$ in the "inf-sup" stability inequality of the Stokes element H_h/L_h used and $c_0 \le 5$, assuming that in case of a stabilized Stokes element the stabilization parameters δ_T are chosen such that $\delta_T \|\nabla q_h\|_T^2 \le \|q_h\|_T^2$. In the case of conforming, "inf-sup" stable Stokes elements, we have $c_0 = 1$.

Proof: We note the identities

$$\begin{aligned}\lambda_{\min}(M^{-1}\Sigma) &= \min_{y\in\mathbb{R}^{N_L}} \frac{\langle \Sigma y, y\rangle}{\langle My, y\rangle} = \min_{y\in\mathbb{R}^{N_L}} \frac{\langle B^T A^{-1} By + Cy, y\rangle}{\langle My, y\rangle} \\ &= \min_{y\in\mathbb{R}^{N_L}} \Big\{ \frac{\langle AA^{-1}By, A^{-1}By\rangle}{\langle My, y\rangle} + \frac{\langle Cy, y\rangle}{\langle My, y\rangle} \Big\}.\end{aligned}$$

For any scalar product on $\mathbb{R}^{N_H}$ such as $\langle\cdot,\cdot\rangle_A := \langle A\cdot,\cdot\rangle$, there holds

$$|x|_A := \max_{z\in\mathbb{R}^{N_H}} \frac{\langle x,z\rangle_A}{|z|_A}, \quad x\in\mathbb{R}^{N_H}.$$

Using these relations for $x := A^{-1}By$ yields

$$\begin{aligned}\lambda_{\min}(M^{-1}\Sigma) &= \min_{y\in\mathbb{R}^{N_L}} \Big\{ \frac{|A^{-1}By|_A^2}{\langle My,y\rangle} + \frac{\langle Cy,y\rangle}{\langle My,y\rangle} \Big\} \\ &= \min_{y\in\mathbb{R}^{N_L}} \Big\{ \max_{z\in\mathbb{R}^{N_H}} \frac{\langle AA^{-1}By,z\rangle^2}{\langle Az,z\rangle\langle My,y\rangle} + \frac{\langle Cy,y\rangle}{\langle My,y\rangle} \Big\} \\ &= \min_{y\in\mathbb{R}^{N_L}} \Big\{ \max_{z\in\mathbb{R}^{N_H}} \frac{\langle By,z\rangle^2}{\langle Az,z\rangle\langle My,y\rangle} + \frac{\langle Cy,y\rangle}{\langle My,y\rangle} \Big\}.\end{aligned}$$

In view of the definition of the matrices A, B, C, and M we find via the assiciation $y\in\mathbb{R}^{N_L} \leftrightarrow p_h\in L_h$ and $z\in\mathbb{R}^{N_H}\leftrightarrow\psi_h\in H_h$ that

$$\lambda_{\min}(M^{-1}\Sigma) = \min_{y\in\mathbb{R}^{N_L}} \Big\{ \max_{z\in\mathbb{R}^{N_H}} \frac{(q_h,\nabla_h\cdot\psi_h)^2}{\|q_h\|^2\|\nabla_h\psi_h\|^2} + \frac{1}{\|q_h\|^2}\sum_{T\in\mathbb{T}_h}\delta_T\|\nabla q_h\|_T^2 \Big\} =: \beta_h^2.$$

Analogously,

$$\begin{aligned}\lambda_{\max}(M^{-1}\Sigma) &= \max_{y\in\mathbb{R}^{N_L}} \frac{\langle\Sigma y,y\rangle}{\langle My,y\rangle} \\ &= \max_{q_h\in L_h} \Big\{ \max_{\psi_h\in H_h} \frac{(q_h,\nabla_h\cdot\psi_h)^2}{\|q_h\|^2\|\nabla_h\psi_h\|^2} + \frac{1}{\|q_h\|^2}\sum_{T\in\mathbb{T}_h}\delta_T\|\nabla q_h\|_T^2 \Big\} \le 5,\end{aligned}$$

where, we use that in 3D:

$$\|\nabla_h\cdot\psi_h\|^2 \le c_1\|\nabla_h\psi_h\|^2,$$

with $c_1 = 4$ in the general case. For conforming. "imf-sup" stavle elements, $H_h\subset H$, this can be sharpened to $c_1 = 1$. This completes the proof. Q.E.D.

As byproduct of the above proof, we obtain the following norm bound for the Schur complement:

$$\|M^{-1}\Sigma\| \le c_1, \tag{4.2.132}$$

where $c_1\le 5$, in general, and $c_1 = 1$ for conforming, "inf-sup" stable Stokes elements.

In the practical realization of the CG methods for the matrix $M^{-1}\Sigma$ it is to be observed that each iteration step consists essentially of a matrix-vector multiplication with $M^{-1}\Sigma$ and the evaluation of several scalar products. This requires as most "expensive" step the solution of a linear system with the "Laplace-like matrix" A:

$$y\to M^{-1}\Sigma y \quad\Leftrightarrow\quad y\to By\to A^{-1}By\to(B^TA^{-1}B+C)y\to M^{-1}(B^TA^{-1}B+C)y.$$

Since A^{-1} is not available exactly the evaluation of $A^{-1}By$ has to be done iteratively. For the matrix A (system matrix of the discretization of the vector-Laplace operator) there exist very efficient and robust PCG- or multigrid methods even on very irregular meshes. Usually

these "inner" iterations within the "outer" CG iteration are controlled by an adaptive stopping criterion oriented by the corresponding iteration residual. This means that in the outer CG iteration the matrix $M^{-1}\tilde{A}^{-1}$ used in the defect computation is faulty (e. g., with elementwise errors of size $\mathcal{O}(10^{-8})$) and changes permanently in the course of the iteration. Therefore, the conditions for the good convergence of the outer CG iteration are not fully satisfied, what may result in erratic convergence behavior and residuals remaining above the accuracy level $\mathcal{O}(10^{-8})$ of the inner iteration. This undesirable defect can be cured by embedding the PCG iteration for the Schur complement $M^{-1}\Sigma$ into an outer defect correction iteration. Thereby, the "unprecise" PCG iteration is used as "preconditioner" S of a simple, robust Richardson iteration:

$$y^t \to d^t := M^{-1}\Sigma y^t - M^{-1}(B^T A^{-1} b) \to r^t = S^{-1} d^t \to y^{t+1} := y^t + r^t.$$

In this way, one obtains a simple, robust, and efficient solution method for the discretized Stokes problem. The extension of this approach to the nonsymmetric and indefinite cases $A+N_1$ (Oseen functional iteration) $A + N_1 + N_2$ (Newton iteration) using generalized PCG methods such as the "GMRES or the "biCGstab method has also not been very successful yet. However, there are competing multigrid methods, which are directly applied to the full block-system matrix $\mathcal{A}$ in the saddle point problem (4.2.120) and are more efficient than the described "stabilized" Schur complement-CG iteration. Here, special care has to be taken in choosing an appropriate smoothing iteration, which can cope with the indefinite character of the problem.

4.2.9 Multigrid method

The main idea underlying a "multigrid algorithm" consists in the fast elimination of "high-frequency" components of the error on the finest mesh ("smoothing") by "cheap" relaxation methods (e. e., point–Jacobi or Gauß-Seidel method) and the reduction of the remaining "smooth" low-frequency error part by defect correction on coarser meshes ("coarse-grid correction"). We briefly describe this process.

The multigrid iteration uses a hierarchy of finite element subspaces,

$$\{H_0 \times L_0\} \subset \ldots \subset \{H_l \times L_l\} \subset \ldots \subset \{H_L \times L_L\},$$

which is obtained within a systematic (adaptively controlled) mesh refinement process. The connection between these spaces is given by so-called "prolongation operators" $P_{l-1}^l : \{H_{l-1} \times L_{l-1}\} \to \{H_l \times L_l\}$ and "restriction operators" $R_l^{l-1} : \{H_l \times L_l\} \to \{H_{l-1} \times L_{l-1}\}$. In the finite element context these operators are simply taken as

$$P_{l-1}^l : \quad \text{natural embedding}, \qquad R_l^{l-1} : \quad L^2 \text{ projection}.$$

The main component of a multigrid algorithm is the smoothing operation $S_l : \{H_l \times L_l\} \to \{H_l \times L_l\}$ on the various mesh levels $0 \le l \le L$ ($l = 0$ corresponding to the coarsest and $l = L$ to the finest mesh.). The multigrid iteration

$$\mathcal{M}\xi = \mathcal{M}(l, z^0, \xi) \tag{4.2.133}$$

on level l with starting value z^0 and m_1 pre- and m_2 post-smoothing steps is recursively defined by:

Multigrid Algorithm $\mathcal{M}(l.z^0, \xi)$ for $l \geq 0$:
For $l = 0$ the multigrid algorithm consists in the "exact" solution of the coarsest problem, i. e., $\mathcal{M}(0, z^0, \xi) := \mathcal{A}^{-1}\xi$. For $l \geq 1$ the following iteration is performed:

1. Pre-smoothing m_1-times: $z^1 := S_l^{m_1} z^0$.
2. Residual on level l: $r^l := \xi - \mathcal{A}_l z^0$.
3. Restriction to level $l-1$: $\tilde{r}^{l-1} := R_l^{l-1} r^l$.
4. Coarse-grid corretion starting with $q^0 := 0$: $q := \mathcal{M}(l-1, q^0, \tilde{r}^{l-1})$.
5. Prolongation to level l: $z^2 := z^1 + P_{l-1}^l q$.
6. Post-smooting m_2-times: $\mathcal{M}(l, z^0, \xi) := S_l^{m_2} z^2$.

Is the multigrid algorithm applied γ-times on each mesh level, one speaks in case $\gamma = 1$ of "V-cycle" and in case $\gamma = 2$ of "W-cycle". The variants with $\gamma \geq 3$ are to expensive and not used in practice. In non-standard situations (e. g., strongly nonsymmetric systems) the V-cycle is usually not robust enough so that the more robust but also more expensive W-cycle is prefered. However, if the multigrid iteration is only used as an "inner" iteration for preconditioning a robust "outer" iteration (e. g., the GMRES method) one usually employs the cheaper V-cycle. The so-called "F-cycle" is an attractive compromise between V- and W-cycle.

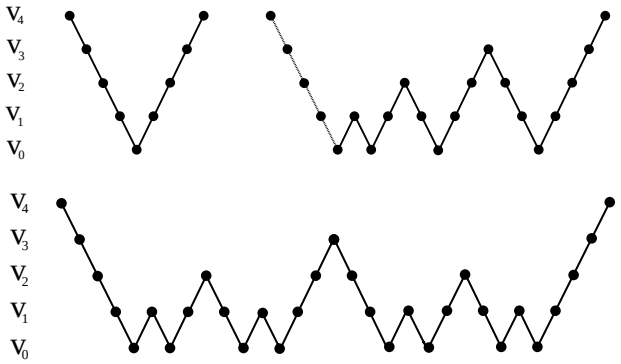


Figure 4.6: Schemes of multigrid V-cycle (upper left), F-cycle (oper right) and W-cycle (bottom)

It is known that a multigrid algorithm as described above, if applied in the standard finite element discretization of scalar elliptic model problems such as the Poisson problem, is of "(almost) optimal complexity". This means that on a mesh $\mathbb{T}_h$ with N_h nodal unknowns the discrete solution u_h is obtained with only $\mathcal{O}(N_h L(N_h))$ operations (uniformly in h). This complexity is improved to the optimal $\mathcal{O}(N_L)$ if the starting value for the multigrid iteration is taken as the solution obtained on the preceding coarser mesh, $u_L^{(0)} := u_{L-1}$. Unfortunately, analogous theoretical results are not available yet for saddle point problems such as the Navier-Stokes equations. However, numerical experience shows that even in such indefinte situations the multigrid concept can work surprisingly well.

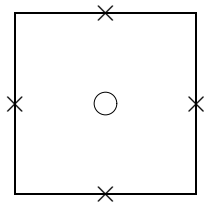
The design of a muligrid algorithm for solving a saddle point problem requires special care. In particular, the choice of the smoothing operation is difficult since the common fixed point iterations such as the point-Jacobi or point-Gauß-Seidel method do not work in this case. This problem can be handles in different ways.

Block-Gauß-Seidel smoothing ("Vanka smoothing"): A very popular smoothing for the (generally nonsymmetric) block matrix $\mathcal{A}$ is obtained by cellwise blocking of velocity and pressure variables within a global Gauß-Seidel iteration. The purpose of this trick is to realize at least locally the indefinite coupling of velocity and pressure variables. This approach was originally proposed for a finite difference discretization of the Stokes problem and has proven very successful also in the context of the finite element method.

We briefly describe the realization of this idea for the nonconforming "rotated" $Q_1^{\rm nc}/P_0^{\rm dc}$ Stokes element. The velocity and pressure variables in a cell T or a patch of cells are numbered consecutively and the corresponding element system matrices indicated by the subscript "loc". These local degrees of freedom are then simultaneously updated within a blockwise Gauß-Seidel iteration:

$$S_{\rm loc} v_{\rm loc}^k + B_{\rm loc} p_{\rm loc}^k = \text{"known"}, \quad B_{\rm loc}^T v_{\rm loc}^k = \text{"known"},$$

where $S_{\rm loc} := A_{\rm loc}$, $S_{\rm loc} := (A+N_1)_{\rm loc}$, or $S_{\rm loc} := (A+N_1+N_2)_{\rm loc}$. This iteration runs over all cells in the current mesh $\mathbb{T}_h$. The local Stokes-like problems have the dimension $d_{\rm loc} = 9$ (in 2D) and $d_{\rm loc} = 19$ (in 3D). The corresponding matrix in 2D is shown below.



$\times$: nodes for v
$\bigcirc$: nodes for p

$$\mathcal{A}_{loc} = \begin{bmatrix} S_{loc,1} & O & B_{loc,1} \\ O & S_{loc,2} & B_{loc,2} \\ -B_{loc,1}^T & -B_{loc,2}^T & 0 \end{bmatrix}.$$

For minimizing costs the main diagonal blocks $S_{{\rm loc},i}$ may by reduced to diagonal matrices by "lumping", $S_{{\rm loc},i} \approx D_{{\rm loc},i}$. Further, for enhancing robustness the iteration may be damped, $v_h^{k+1} = v_h^k + \omega(\tilde{v}_h^{k+1} - v_h^k)$ with some $\omega \in (0,1)$.

We illustrate the performance of the resulting multigrid algorithm for the "rotated" $Q_1^{\rm nc}/P_0^{\rm dc}$ Stokes element applied in approximating the so-called "lid-driven cavity" problem by the Oseen linearization for a range of viscosities $1 \geq \nu \geq 1/5000$.

# cells	1600	6400	25600	# iterations
$\nu = 1$	0.081	0.096	0.121	4
$\nu = 1/100$	0.098	0.099	0.130	6
$\nu = 1/1000$	0.227	0.245	0.168	9
$\nu = 1/5000$	0.285	0.368	0.370	18

Table 4.1: Multigrid convergence rates (2 pre- and 1 post-smoothing with the "Vanka smoother") and number of outer functional iteration steps on uniformly refined meshes.

We remark that a similar "blockwise" iteration can also be used in an incomplete block-LU decomposition. From the common analysis of multigrid methods we know that "pointwise" iterations lose their smoothing property if the mesh has large anisotropies. This is the standard case in resolving boundary layers. The solution is the use of special smoothing, which in the limit of extremly streched cells reduce to "exact" solvers. These are, for example, "line-Gauß-Seidel" and "ILU" iteration. Since the above "Vanka smoother" acts on the velocity variables like a "point-Gauß-Seidel" iteration, there occur problems in case of strongly streched mesh cells. This difficulty can be overcome by patchwise blocking the variables in the direction of the "anisotropic" mesh refinement leading to a method called "stringwise Gauß-Seidel smoothing".

4.3 A nested solution scheme

The practical solution process of the Navier-Stokes problem may be organized in form of a "nested" solution scheme, in which discretization (by a stable Stokes element), linearization (by a Newton-type iteration) and algebraic solution (by a GMRES-multigrid method) is adaptively coupled. The whole process should be controlled by a posteriori error estimates for the different algorithmic components. We formulate this scheme for a nonlinear problem of the following abstract variational form:

$$a(u;\varphi) = f(\varphi) \quad \forall\, \varphi \in V, \tag{4.3.134}$$

where the nonlinear form $a(\cdot;\cdot)$ represents the "energy form" of the Navier-Stokes problem, with $V = H \times L$ and $u = \{v, p\} \in H \times L$, or that of a general quasi-linear elliptic problem, with $V = H_0^1(\Omega) \cap W^{1,\infty}(\Omega)$.

Let be given a desired tolerance TOL for some error measure $E(\cdot)$ (e. g., a norm $\|\cdot\|$ or some more local functional $E(u) = u(a)$) and a maximum mesh complexity $N_{\max}$ (due to the capacity limits of the computer used). Starting from a coarse initial mesh $\mathbb{T}_0$, a hierarchy of successively refined meshes $\mathbb{T}_l$, $l \geq 1$, with $N_l := \#\{T \in \mathbb{T}_l\}$, and corresponding finite element spaces V_l, with $V_l \subset V_{l+1}$, are generated by the following algorithm.

Nested (adaptive) solution algorithm:

(0) *Initialization for* $l = 0$: Compute an initial approximation $u_0 \in V_0$ (coarsest mesh).

(1) *Defect correction iteration for* $l \geq 1$*:* Start with $u_l^{(0)} := u_{l-1} \in V_l$.

(2) *Iteration step:* For $j \geq 0$ form the defect functional

$$d_l^{(j)}(\varphi) := f(\varphi) - a(u_l^{(j)};\varphi), \quad \varphi \in V_l. \tag{4.3.135}$$

Pick a suitable approximation $\tilde{a}'(u_l^{(j)};\cdot,\cdot)$ to the derivative form $a'(u_l^{(j)};\cdot,\cdot)$ (with good stability and solvability properties) and compute a correction $v_l^{(j)} \in V_l$ from the linear equation

$$\tilde{a}'(u_l^{(j)}; v_l^{(j)}, \varphi) = d_l^{(j)}(\varphi) \quad \forall \varphi \in V_l. \tag{4.3.136}$$

For that, a CG/GMRES-type method with multigrid preconditioning may be employed using the hierarchy of already constructed meshes $\{\mathbb{T}_l, ..., \mathbb{T}_0\}$. This "inner" iteration

yields an approximation $\tilde{v}_l^{(j)} \approx v_l^{(j)}$ together with an a posteriori error estimate

$$E(\tilde{v}_l^{(j)} - v_l^{(j)}) \leq \eta_{\text{iter}}^{\text{in}},$$

from which a stopping criterion is obtained by balancing iteration and discretization errors on mesh $\mathbb{T}_l$. Then, update $u_l^{(j+1)} = u_l^{(j)} + \lambda_l \tilde{v}_l^{(j)}$, with some relaxation parameter $\lambda_l \in (0, 1]$, increment j and go back to (2). This process is repeated until an approximation $\tilde{u}_l := u_l^{(j)} \in V_l$ is reached with a sufficiant accuracy, also based on an a posteriori estimate for this "outer" iteration,

$$E(\tilde{u}_l - u_l) \leq \eta_{\text{iter}}^{\text{out}}.$$

(3) *Error estimation and mesh adaptation:* Accept $\tilde{u}_l$ as the solution on mesh $\mathbb{T}_l$ and evaluate an a posteriori error estimate of the form

$$E(\tilde{u}_l - u) \leq \eta_{\text{discr}}(\tilde{u}_l) = \sum_{T \in \mathbb{T}_l} \eta_T(\tilde{u}_l).$$

The so-called "cell-error indicators" $\eta_T(\tilde{u}_l)$ are used to construct a new (refined) mesh $\mathbb{T}_{l+1}$ and corresponding finite element space V_{l+1} by seeking equilibration through the following strategy:

$$\text{if } \eta_T \gg \frac{\text{TOL}}{N_l} \quad \text{refine } T, \qquad \text{if } \eta_T \approx \frac{\text{TOL}}{N_l} \quad \text{keep } T, \qquad \text{if } \eta_T \ll \frac{\text{TOL}}{N_l} \quad \text{coarsen } T.$$

where "refining" and "coarsening" may be realized using the structure of the hierarchical meshes $\mathbb{T}_{l-1} \subset \mathbb{T}_l \subset \mathbb{T}_{l+1}$ (bysection allowing "hanging" nodes). Here, the underlying philosophy is that the additional "iteration errors" due to the use of an only approximate finite element solution $\tilde{u}_l \approx u_l \in V_l$ can be controlled by a combined a posteriori error estimate of the form

$$E(\tilde{u}_l - u) \leq \eta_{\text{discr}} + \eta_{\text{iter}}^{\text{out}} + \eta_{\text{iter}}^{\text{in}}.$$

This induces stopping criteria for the outer and inner algebraic iterations of the form

$$\eta_{\text{iter}}^{\text{out}} + \eta_{\text{iter}}^{\text{in}} \leq \kappa \eta_{\text{discr}}, \tag{4.3.137}$$

where usually, for safety reasons, $\kappa := 1/10$. If then, on some mesh $\mathbb{T}_l$ good equilibration is achieved, there holds

$$E(\tilde{u}_l - u) \approx \sum_{T \in \mathbb{T}_l} \eta_T \approx \sum_{T \in \mathbb{T}_l} \frac{\text{TOL}}{N_l} = \text{TOL},$$

and the solution process can be stopped. The process is also stopped if the current mesh $\mathbb{T}_l$ has already maximal complexity, i. e., $N_l \approx N_{\max}$. Otherwise, if $\eta_{\text{discr}} > \text{TOL}$ and $N_l < N_{\max}$, increment l and go back to (1).

The nested solution process described above requires to choose several parameters in the stopping criteria for the inner and outer algebraic iterations and in the strategy for equilibrating the cell-error indicators on the current mesh. The appropriate choice of these parameters depends very much on the problem to be solved and needs some expert knowledge by the user. There

exists a systematic approach to deriving a posteriori error estimates of the form (4.3.137) even for nonlinear problems as considered here. The theory underlying the so-called "Dual Weighted Residual (DWR) method will be the subject of the next chapter.

Remark 4.11: Usually the evaluation of the *a posteriori* error estimate (4.3.137) involves only the solution of *linearized* problems. Hence, the whole error estimation may amount to only a relatively small fraction of the total cost for the solution process. This has to be compared to the usually much higher cost when working on non-adapted meshes and without efficient stopping criteria.

4.4 Exercises

Exercise 4.1: Let $(v_m)_{m\in\mathbb{N}}$ be a bounded sequence of functions in $V = J_1(\Omega)$, which converges weakly in V and strongly in $J_0(\Omega)$ to some limit $v \in V$:

$$(\nabla v_m, \nabla\varphi) \to (\nabla v, \nabla\varphi),\ \varphi \in V, \qquad \|v_m - v\| \to 0 \quad (m \to \infty).$$

For any fixed $\varphi \in V$ let $(\varphi_m)_{m\in\mathbb{N}}$ be a sequence, which converges strongly in V to φ, i. e., $\|\nabla(\varphi_m - \varphi)\| \to 0$ $(m \to \infty)$. Show that this implies the following convergences:

$$(\nabla v_m, \nabla\varphi_m) \to (\nabla v, \nabla\varphi), \quad (v_m \cdot \nabla v_m, \varphi_m) \to (v \cdot \nabla v, \varphi) \quad (m \to \infty).$$

Exercise 4.2: The key property of the nonlinear form $b(u, v, w) := (u \cdot \nabla v, w)$ in the Navier-Stokes equations is the relation

$$b(u, v, v) = 0, \quad u, v \in V := \{v \in H_0^1(\Omega)^d \,|\, \nabla \cdot v = 0\}.$$

If in the approximation (non-conforming) finite element spaces $V_h \not\subset V$ are used this property is lost on the discrete level. Therefore, one may use instead a symmetrized version of this nonlinear form, such as

$$\tilde{b}(u, v, w) := \tfrac{1}{2}b(u, v, w) - \tfrac{1}{2}b(u, w, v).$$

Show that for this modification there holds

$$\tilde{b}(u, v, w) = b(u, v, w), \quad u, v, w \in V,$$

and, by construction,

$$\tilde{b}(u_h, v_h, v_h) = 0, \quad u_h, v_h \in V_h.$$

Exercise 4.3: A common linearization of the Navier-Stokes equations is the so-called ''Oseen linearization", with a given divergence-free flow field $\bar{v} \in H^1(\Omega)^d$,

$$-\nu\Delta v + \bar{v} \cdot \nabla v + \nabla p = f, \quad \nabla \cdot v = 0, \quad \text{in } \Omega,$$

and the usual boundary conditions on $\partial\Omega = \Gamma_{\text{rigid}} \cup \Gamma_{\text{in}} \cup \Gamma_{\text{out}}$:

$$v = 0 \text{ on } \Gamma_{\text{rigid}}, \quad v = \bar{v} \text{ on } \Gamma_{\text{in}}, \quad -\nu\partial_n v + pn = 0 \text{ on } \Gamma_{\text{out}}.$$

Show the unique weak solvability of this linear problem for the case that $\text{meas}(\Gamma_{\text{rigid}}) > 0$ and $\bar{v} \cdot n \geq 0$ on Γ_{out}. **Hint:** One may use the theorem of Lax-Milgram.

Exercise 4.4: It has been shown in class that the variational stationary Navier-Stokes problem in $\mathbb{R}^d$ $(d = 2, 3)$,

$$\nu(\nabla v, \nabla\varphi) + (v \cdot \nabla v, \varphi) = (f, \varphi) \quad \forall\varphi \in V := \{v \in H_0^1(\Omega)^d \,|\, \nabla \cdot v = 0\},$$

possesses a unique solution $v \in V$ under the smallness condition $c_*^2\nu^{-2}\|f\|_{-1} < 1$ on the data. Show that under the same condition for any starting value $v^0 \in V$ the functional iteration

$$\nu(\nabla v^t, \nabla\varphi) + (v^{t-1} \cdot \nabla v^t, \varphi) = (f, \varphi) \quad \forall\varphi \in V,$$

produces a sequence $(v^t)_{t\in\mathbb{N}} \subset V$, which converges in V to this solution. **Hint:** One may interprete the functional iteration as a fixed point iteration $v^t = g(v^{t-1})$ and employ the Banach fixed point theorem.
Remark: This offers an alternative proof for the existence of weak solutions of the Navier-Stokes equations in the case of small data.

Exercise 4.5: Consider the approximation of the variational (linear) Stokes problem in two dimensions,

$$v \in V = J_1(\Omega): \qquad (\nabla v, \nabla\varphi) = (f, \varphi) \quad \forall\varphi \in V,$$

by the finite element method on quasi-uniform families of triangulations.

a) Construct V-conforming finite element subspaces $V_h \subset V$ from the quintic H^2-conforming Argyris plate element. (Hint: Observe that $\mathrm{div}(\mathrm{rot}) = 0$.)

b) Specify appropriate local nodal bases of these spaces V_h.

c) For the case $f \in L^2(\Omega)^2$ and Ω a convex polygonal domain derive optimal order error estimates in the H^1 and L^2 norms.

d) Give an idea on how the corresponding discrete pressures p_h can be computed once the velocity field v_h is known.

Exercise 4.6: Show that the Q_1^c/P_0^{dc} Stokes element is not "inf-sup" stable in general. To this end consider a uniform cartesian (quadrilateral) mesh $\mathbb{T}_h$ of the unit square $\Omega = (0, 1)^2$ and show that for the so-called "checkerboard" pressure function $q_h^{\mathrm{check}} \in L_h$, which has the alternating constant values ± 1 there holds

$$(q_h^{\mathrm{check}}, \nabla \cdot \varphi_h) = 0, \quad \varphi_h \in H_h.$$

This shows that in this case the "inf-sup" stability estimate cannot hold.

Exercise 4.7: Let the Navier-Stokes problem be approximated by a conforming and uniformly "inf-sup" stable Stokes element with finite element subspaces $H_h \times L_h \subset H \times L$. Suppose that the data are sufficiently small, $c_*^2\nu^{-2}\|f\|_{-1} < 1$, to guarantee the existence of a unique solution $\{v, p\}$. For solving the corresponding discrete nonlinear problems consider the following iteration, in which the nonlinear term is treated fully "explicitly",

$$\begin{aligned}\nu(\nabla v_h^t, \nabla\varphi) - (p_h^t, \nabla \cdot \varphi) &= (f, \varphi) - \tfrac{1}{2}(v_h^{t-1} \cdot \nabla v_h^{t-1}, \varphi) + \tfrac{1}{2}(v_h^{t-1} \cdot \nabla\varphi, v_h^{t-1}) \quad \forall\varphi \in H_h,\\ (\chi, \nabla \cdot v_h^t) &= 0 \quad \forall\chi \in L_h,\end{aligned}$$

for a starting value $v_h^0 \in V_h$, which reduces the solution of a nonlinear algebraic system to a sequence of linear algebraic (Stokes) systems.

a) Give a conditions on the size of the initial error $\|v_h - v_h^0\|$, which implies that this iteration converges and specify the rate of convergence.

b) Define a modified iteration, which under the same condition on the data converges for all starting value $v_h^0 \in V_h$ (with proof).

Exercise 4.8: Let $\{\mathbb{T}_h\}_{h>0}$ be a shape uniform family of triangulations in $\mathbb{R}^2$ and $v \in H$. In class it has been claimed that for the (cellwise) quadratic polynomial $Q_h^{(2)} v|_T \in P_2(T)$ with the properties

$$Q_h^{(2)} v(a) = 0, \;\; a \in \partial^2 \mathbb{T}_h, \qquad \int_\Gamma Q_h^{(2)} v \, ds = \int_\Gamma (v - I_h^{(1)} v) \, ds, \;\; \Gamma \in \partial \mathbb{T}_h,$$

there holds

$$\|\nabla Q_h^{(2)} v\| \le c \|\nabla (v - I_h^{(1)} v)\|.$$

Give a proof of this estimate. (Hint: Consider each cell $T \in \mathbb{T}_h$ separately, write $Q_h^{(2)} v$ in its nodal basis representation and use the trace inequality locally on T.)

Exercise 4.9: Consider the special situation of a family of uniformly cartesian (quadrilateral) meshes in 2D and the discretization of the Stokes problem by the nonconforming "rotated" bilinear Stokes element $\tilde{Q}_1^{\rm nc}/P_0^{\rm dc}$.

a) Show that on any cell T the prescription of the nodal values $\chi_\Gamma(v) := |\Gamma|^{-1}(v_h, 1)_\Gamma$, $\Gamma \subset \partial T$, uniquely determines a polynomial in the space $Q_1^{\rm rot}(T) = {\rm span}\{1, x_1, x_2, x_1^2 - x_2^2\}$ ("unisolvence").

b) Use the technics described in class for proving the uniform "inf-sup" stability of this Stokes element.

Exercise 4.10: It has been shown in class that on triangular meshes in 2D the conforming $\tilde{P}_1^{\rm c}/P_1^{\rm c}$ (MINI) Stokes element and the nonconforming $P_1^{\rm nc}/P_0^{\rm dc}$ (Crouzeix/Raviart) Stokes element are uniformly "inf-sup" stable. Define the analogues of these Stokes elements on tetrahedral meshes in 3D and show their uniform "inf-sup" stability.

Exercise 4.11: Show that the "MINI" Stokes element introduced in class is equivalent to a version of the "stabilized" $P_1^{\rm c}/P_1^{\rm c}$ Stokes element with the pressure equation:

$$(\chi_h, \nabla \cdot v_h^{(1)}) + \sum_{T \in \mathbb{T}_h} \delta_T (\nabla \chi_h, \nabla p_h)_T = \sum_{T \in \mathbb{T}_h} \sum_{i=1}^{2} \frac{\delta_T |T|}{(1, \varphi_{T,i}^b)_T} (\partial_i \chi_h, (f_i, \varphi_{T,i}^b) \varphi_{T,i}^b)_T \quad \forall \chi_h \in L_h,$$

where $v_h^{(1)}$ is the linear part of the MINI-velocity, $\varphi_{T,i}^b$ the "bulb" basis functions on an element T and $\delta_T \approx h_T^2$. (Hint: Split the variational equations of the MINI element into equations for the linear part and its "bulb" component.)

Exercise 4.12: Consider the lowest-order nonconforming $P_1^{\rm nc}/P_0^{\rm dc}$ Stokes element analyzed in

class. Prove for this element the following discrete version of the Poincaré inequality on shape-uniform triangulations $\mathbb{T}_h$:

$$\|v_h\| \le c\|\nabla_h v_h\|, \quad v_h \in H_h,$$

where again ∇_h denotes the cellwise defined gradient operator.
(Hint: One may use a duality argument and the fact that the continuity property of functions in H_h implies $\int_\Gamma [v_h]\, ds = 0$ for $\Gamma \in \partial\mathbb{T}_h, v_h \in H_h$.)

Exercise 4.13: The conforming, "equal order" Stokes element of type P_2^c/P_2^c (continuous, piecewise quadratic velocities and pressures) in its pure form is not "inf-sup" stable. Formulate the corresponding fully consistent stabilized versions of this Stokes element for approximating a) the linear Stokes problem and b) the nonlinear Navier-Stokes problem.

Exercise 4.14: Consider the approximation of the nonlinear Navier-Stokes problem as considered in class seeking a $v \in V = \{\varphi \in H_0^1(\Omega)^2 \,|\, \nabla \cdot \varphi = 0\}$ such that

$$\nu(\nabla v, \nabla\varphi) + (v \cdot \nabla v, \varphi) = (f, \varphi) \quad \forall \varphi \in V,$$

by the finite element method using a conforming "inf-sup" stable Stokes element, e. g., the conforming P_2^c/P_0^{dc} element. Suppose that in this approximation the nonlinear form $n(v, v, \varphi) = (v \cdot \nabla v, \varphi)$ is used in its original nonsymmetrized form, seeking $v_h \in V_h$ such that

$$\nu(\nabla v_h, \nabla\varphi_h) + (v_h \cdot \nabla v_h, \varphi_h) = (f, \varphi_h) \quad \forall \varphi_h \in V_h,$$

In this case, we do not have $n(v, \varphi, \varphi) = 0$, which causes difficulties even in proving existence of discrete solutions.

a) Use the properties of the P_2^c/P_0^{dc} element in 2D to show the estimate

$$|n(v, \varphi, \varphi)| \le c_1 h L(h) \|\nabla v_h\| \|\nabla\varphi_h\|^2, \quad v_h, \varphi_h \in V_h.$$

where $L(h) := \max\{1, \log(1/h)\}$.

b) How does the analogue of this estimate in 3D look like?
Hint: One may use integration by parts, the definition of the discrete space V_h and the Sobolev inequalities (see Exercise 7.2)

$$\|w_h\|_\infty \le cL(h)\|\nabla w_h\| \text{ in 2D, and } \|w_h\|_\infty \le ch^{-1/2}\|\nabla w_h\| \text{ in 3D.}$$

Exercise 4.15: Consider the situation described in Exercise 12.3.

a) Apply the Brouwer fixed point theorem (as in the corresponding argument on the continuous level) to show that for sufficiently small $h \le h_1$ the discrete Navier-Stokes problem is solvable.

b) Prove that for sufficiently small data, $c_*^2 \nu^{-2} \|f\|_{-1} < 1$, the solution obtained in (a) is unique.

Exercise 4.16: Consider again the lowest-order nonconforming P_1^{nc}/P_0^{dc} Stokes element analized in class. Prove for this element the "discrete" versions of the following Sobolev inequalities on quasi-uniform triangulations $\mathbb{T}_h$:

$$(i) \quad \|v_h\|_6 \le c_*^{nc} \|\nabla_h v_h\|, \quad v_h \in H_h,$$
$$(ii) \quad \|v_h\|_3 \le c_*^{nc} \|v_h\|^{1/2} \|\nabla_h v_h\|^{1/2}, \quad v_h \in H_h,$$

where ∇_h denotes again the cellwise defined gradient operator.

(**Hints:** (i) For proving the first inequality, one may use the following estimate for cellwise constant functions $\bar{w}$ on the quasi-uniform mesh $\mathbb{T}_h$ in 2D or 3D (Try to give an argument for this estimate.):

$$\|\bar{w}\|_6 \leq ch^{-1}\|\bar{w}\|.$$

Further, to the (nonconforming) function $v_h \in H_h$, one can associate a "smooth" function $v \in H$ by the relation

$$(\nabla v, \nabla\varphi) = (\nabla_h v_h, \nabla\varphi) \quad \forall\varphi \in H.$$

Then, considering v_h as an approximation to v, by a duality argument, one shows the following error estimate (Try to give a complete argument for this estimate.)

$$\|v - v_h\| \leq ch\|\nabla_h v_h\|.$$

(ii) The second inequality can be obtained by using the Hölder inequality and the first estimate.)

Exercise 4.17: Consider the variational Navier-Stokes problem seeking $v \in V = \{\varphi \in H = H_0^1(\Omega)^d \,|\, \nabla\cdot\varphi = 0\}$ such that

$$\nu(\nabla v, \nabla\varphi) + (v\cdot\nabla v, \varphi) = (f, \varphi) \quad \forall\varphi \in V,$$

in the case of general data ν, f. Suppose that there exists a solution $v \in V$, for which the derivative form

$$a'(v;\psi,\varphi) := \nu(\nabla\psi, \nabla\varphi) + (v\cdot\nabla\psi, \varphi) + (\psi\cdot\nabla v, \varphi), \quad \psi, \varphi \in V,$$

is coercive, i. e., with some constant $\alpha > 0$ there holds

$$\sup_{\varphi\in V} \frac{a'(v;\psi,\varphi)}{\|\nabla\varphi\|} \geq \alpha\|\nabla\psi\|, \quad \psi \in V.$$

Show that then this solution is locally unique in some ball $B_R(v) \subset V$ with radius $R = R(\alpha) > 0$.

Remark: This concerns the situation, such as in the Taylor problem, in which for the same set of data multiple stationary solutions exist which are only locally unique.

Exercise 4.18: On a convex polygonal domain $\Omega \subset \mathbb{R}^2$ the Laplace operator $-\Delta$ (similarly as the Stokes operator) can be defined as a self-adjoint positive definite operator in $L^2(\Omega)$ with dense domain of definition $D(-\Delta) = H_0^1(\Omega) \cap H^2(\Omega)$ and range $R(-\Delta) = L^2(\Omega)$. It is invertible with compact inverse $-\Delta^{-1} : L^2(\Omega) \to L^2(\Omega)$. Therefore, its spectrum $\sigma(-\Delta)$ consists of isolated real positive eigenvalues $0 < \lambda_1 \leq \lambda_2 \leq \lambda_3 \leq \ldots$ with no finite accumulation point. For any value $\lambda \notin \sigma(-\Delta)$ the operator $-\Delta - \lambda I : D(-\Delta) \to L^2(\Omega)$ is onto with bounded inverse $(-\Delta - \lambda I)^{-1}$. The same holds true for the operator $-\Delta - \lambda I : H_0^1(\Omega) \to H^{-1}(\Omega)$ (defined in a variational sense). For such λ the bilinear form

$$a_\lambda(u, \varphi) := (\nabla u, \nabla\varphi) - \lambda(u, \varphi), \quad u, \varphi \in V := H_0^1(\Omega),$$

can, in general, not be V-elliptic. Show that it is always "coercive", i. e.,

$$\sup_{\varphi\in V}\frac{a_\lambda(u,\varphi)}{\|\nabla\varphi\|}\geq\alpha\|\nabla u\|,\quad u\in V,$$

with a constant $\alpha>0$ related to the norm of the corresponding operator $(-\Delta-\lambda I)^{-1}: H^{-1}(\Omega)\to H_0^1(\Omega)$.
Hint: You may use an auxiliary problem with appropriate right-hand side and the boundedness of the operator $(-\Delta-\lambda I)^{-1}: H^{-1}(\Omega)\to H_0^1(\Omega)$.

Exercise 4.19: Let the singularly perturbed Sturm-Liouville problem

$$-\varepsilon u''+u'=0,\quad \text{in } \Omega=(0,1),\quad u(0)=1,\ \ u(1)=0,$$

in 1D be discretized by conforming "linear" finite elements on an equidistant mesh $0=x_0<x_1<\cdots<x_N<x_{N+1}=1$ with "artificial diffusion" set to $\varepsilon_h:=\varepsilon+\delta h$ for stabilizing the transport term. Form the corresponding finite difference equation and derive an explicit formula for its solution. Deduce that the choice $\delta\geq\frac{1}{2}$ leads to a diagonally dominant M-matrix.

Exercise 4.20: Consider the discretization of the p-Laplace-like problem $(1<p<\infty)$

$$\min\ J(u):=\frac{1}{p}\int_\Omega(1+|\nabla u|^2)^{p/2}\,dx-\int_\Omega fu\,dx\quad\text{on } V=H_0^{1,p}(\Omega),$$

in 2D by conforming "linear" finite elements, $V_h\subset V$, on a family of triangulations $\{\mathbb{T}_h\}_{h>0}$. Formulate explicitly the corresponding discrete variational equations and the Newton iteration for their solution.

Exercise 4.21: Let the stationary Navier-Stokes problem with homogeneous Dirchlet boundary conditions be approximated by a conforming "inf-sup" stable Stokes element leading to the finite dimensional problems: Find $\{v_h,p_h\}\in H_h\times L_h$, such that

$$\begin{aligned}\nu(\nabla v_h,\nabla\varphi_h)+\tilde n(v_h,v_h,\varphi_h)-(p_h,\nabla\cdot\varphi_h)&=(f,\varphi_h)\quad\forall\varphi_h\in H_h,\\ (\chi_h,\nabla\cdot v_h)&=0\quad\forall\chi_h\in L_h.\end{aligned}$$

Show in analogy to the continuous level with help of the Brouwer fixed point theorem that these problems always possess solutions, which are unique if the data are sufficiently small, $q=c_*^2\nu^{-2}\|f\|_{-1}<1$.

Exercise 4.22: Let the finite element subspaces $H_h\times L_h\subset H\times L$ be defined on basis of a conforming Stokes element. Further, suppose that there exists an interpolation operator $\pi:h:H\to H_h$ satisfying the relations

$$\begin{aligned}(\chi_h,\nabla\cdot\pi_h v)=(\chi_h,\nabla\cdot v),\quad \chi_h\in L_h,\ v\in H,\\ \|\nabla\pi_h v\|\leq c\|\nabla v\|,\quad v\in H.\end{aligned}$$

Show that then the family of spaces $\{H_h\times L_h\}_{h>0}$ is uniformly "inf-sup" stable. Use this result for proving the uniform "inf-sup" stability of the conforming $\tilde P_1^{\rm c}/P_1^{\rm c}$ Stokes element ("MINI element") in 2D.

Exercise 4.23: Give short answers to the following questions:

1. Which property of a bilinear form $a(\cdot,\cdot)$ on a Hilbert space H with norm $\|\cdot\|_H$ is meant by "H-ellipticity"? Give an example of an H-elliptic bilinear form, which is not symmetric, i. e., not a scalar product.

2. Which properties must a family of triangulations $\{\mathbb{T}_h\}_{h>0}$ of a polygonal domain $\Omega \subset \mathbb{R}^2$ possess in order to be called "quasi-uniform"?

3. Formulate the "minimal surface problem" on a domain $\Omega \subset \mathbb{R}^2$. In what sense is here the boundary condition imposed?

4. Consider the approximation of the Poisson problem on a convex polygonal domain $\Omega \subset \mathbb{R}^2$ by "linear" finite elements on quasi-uniform meshes. What are the best achievable orders of approximation in the L^∞- and the $W^{1,\infty}$-norm if the sulution satisfies $u \in W^{2,\infty}(\Omega)$?

5. What is the content of the Poincaré inequality on a bounded domain $\Omega \subset \mathbb{R}^d$? Formulate conditions on the functions considered, under which this inequality holds true.

6. How does the "inf-sup" stability condition for Stokes elements look like and what is it good for?

7. Give three examples of uniformly "inf-sup" stable pairs of finite element spaces $H_h \times L_h$ for approximating the Stokes problem.

8. How does the functional iteration applied in solving the Navier-Stokes equation on the function space level look like? Under what condition is its global convergence guaranteed?

9. What is the purpose of modifying the nonlinear term $n(u,v,w) := (u \cdot \nabla v, w)$ in the variational Navier-Stokes problem to $\tilde{n}(u,v,w) := \frac{1}{2}n(u,v,w) - \frac{1}{2}n(u,w,v)$? Show that $\tilde{n}(u,v,w) = n(u,v,w)$ for functions $u \in V$ and $v, w \in H$.

10. What is the purpose of the streamline diffusion stabilization in the finite element approximation of the Navier-Stokes problem? What is the dependence of the stabilization parameters δ_T on the local mesh size h_T?

Bibliography

[1] R. Rannacher: *Numerische Mathematik 0 (Einf. in die Numerische Mathematik)*, Lecture Notes, Heidelberg University,
`http://numerik.uni-hd.de/~lehre/notes/`

[2] R. Rannacher: *Numerische Mathematik 1 (Numerik gewöhnlicher Differentialgleichungen)*, Lecture Notes, Heidelberg University,
`http://numerik.uni-hd.de/~lehre/notes/`

[3] R. Rannacher: *Numerische Mathematik 2 (Numerik partieller Differentialglei-chungen)*, Lecture Notes, Heidelberg University,
`http://numerik.uni-hd.de/~lehre/notes/`

[4] R. Rannacher: *Numerische Mathematik 3 (Numerische Methoden der Kontinuumsmechanik)*, Lecture Notes, Heidelberg University,
`http://numerik.uni-hd.de/~lehre/notes/`

(I) General References on Functional Analysis, PDEs and their Numerical Solution

[5] R. A. Adams: *Sobolev Spaces*, Academic Press, New York, 1975.

[6] P. G. Ciarlet and J.L. Lions: *Handbook of Numerical Analysis Volume II, Finite Element Methods I, and Volume IV, Finite Element Methods II*, North-Holland: Amsterdam, 1991.

[7] G. P. Galdi: *An Introduction to the Mathematical Theory of the Navier-Stokes Equations. Vol. 1: Linearized Steady problems, Vol. 2: Nonlinear Steady Problems*, Springer: Berlin-Heidelberg-New York, 1998.

[8] P. M. Halmos: *Finite Dimensional Vector Spaces*, Springer, 1974.

[9] G. Hellwig: *Partial Differential Equations. An Introduction*,
B.G. Teubner, Stuttgart, 1977.

[10] J. Wloka: *Partial Differential Equations*, Cambridge University Press, Cambridge, 1987.

[11] W. R. Strauss: *Partial Differential Equations: An Introduction*,
John Wiley 1992.

[12] P. Lax: *Functional Analysis*, Wiley-Interscience, 2002.

[13] M. Renardy, R. Rogers: *An Introduction to Partial Differential Equations*,
Springer 1993.

[14] J. Joos: *Partial Differential Equations*, Springer 2013.

[15] T. Kato: *Perturbation Theory for Linear Operators*, Springer, 2nd ed., 1980.

[16] F. Riesz and B. Sz.-Nagy: *Functional Analysis*, Dover Publications, 1990.

[17] W. Rudin: *Functional Analysis*, McGraw-Hill Science, 1991.

[18] M. Schechter: *Principles of Functional Analysis*, AMS, 2nd ed., 2001.

[19] R. Temam: *Navier-Stokes Equations. Theory and numerical analysis*. North Holland: Amsterdam, 1987.

[20] A. Tveito and R. Winther: *Introduction to Partial Differential Equations: A Computational Approach*, Springer, 1998.

[21] K. Yosida: *Functional Analysis*, Springer, 6th ed., 1980.

(II) General References on the Finite Element Method

[22] T. Apel: *Anisotropic Finite Elements: Local Estimates and Applications*, B.G.Teubner: Stuttgart-Leipzig, 1999.

[23] O. Axelsson and V. A. Barker: *Finite Element Solution of Boundary Value Problems, Theory and Computation*, Academic Press, 1984.

[24] W. Bangerth and R. Rannacher: *Adaptive Finite Element Methods for Differential Equations*, Lectures in Mathematics, ETH Zürich, Birkhäuser, Basel 2003.

[25] D. Braess: *Finite Elements: Theory, Fast Solvers, and Applications in Solid Mechanics*, Springer 2007 (3rd edition).

[26] P. G. Ciarlet: *The Finite Element Method for Elliptic Problems*, North-Holland 1978.

[27] V. Girault and P.-A. Raviart: *Finite Element Methods for the Navier-Stokes Equations*. Springer: Berlin-Heidelberg-New York, 1986.

[28] A. Quarteroni and A. Valli: *Numerical Approximation of Partial Differential Equations*, Springer, 1994.

[29] G. Strang and G. J. Fix: *An Analysis of the Finite Element Method*, Prentice-Hall, 1973.

[30] O. Axelsson, V. A. Barker: *Finite Element Solution of Boundary Value Problems, Theory and Computation*, Academic Press 1984.

[31] W. Hackbusch: *Elliptic Differential Equations. Theory and Numerical Treatment*, Springer, 1992.

[32] C. Johnson: *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Cambridge University Press 1987.

[33] F. Brezzi and M. Fortin: *Mixed and Hybrid Finite Element Methods*, Springer 1991.

[34] S. C. Brenner, L. R. Scott: *The Mathematical Theory of Finite Element Methods*, Springer 1994.

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

(III) Special References

[36] M. Dobrowolski and R. Rannacher: *Finite element methods for nonlinear elliptic systems of second order*, Math. Nachr. 95, 155-172 (1980).

[37] J. Frehse and R. Rannacher: *Eine L^1-Fehlerabschätzung für diskrete Grundlösungen in der Methode der finiten Elemente*, Bonn. Math. Schr. 89, 92-114 (1976).

[38] J. Frehse and R. Rannacher: *Asymptotic L^∞-error estimates for linear finite element approximations of quasi-linear boundary value problems*, SIAM J. Numer. Anal. 15, 418-431 (1978).

[39] J. G. Heywood and R. Rannacher: *Finite element approximation of the nonstationary Navier-Stokes Problem. I. Regularity of solutions and second order error estimates for spatial discretization*, SIAM J. Numer. Anal. 19, 275–311 (1982).

[40] J. G. Heywood, R. Rannacher, and S. Turek: *Artificial boundary and flux and pressure conditions for the incompressible Navier-Stokes equations*, Int. J. Comput. Fluid Mecj. 22, 325–352 (1996).

[41] R. Rannacher: *Some asymptotic error estimates for finite element approximation of minimal surfaces*, R.A.I.R.O. Anal. Numer. 11, 181-196 (1976).

[42] R. Rannacher: *On nonconforming and mixed finite element methods for plate bending problems. The linear case*, R.A.I.R.O. Anal. Numer. 13, 369-387 (1979).

[43] R. Rannacher: *On finite element approximation of general boundary value problems in nonlinear elasticity*, Calcolo 17, 175-193 (1980).

[44] R. Rannacher: *On the convergence of the Newton-Raphson method for strongly nonlibear problems*, in Nonlinear Computational Mechanics, State of the Art (P. Wriggers and W. Wagner, eds), pp. 11–30, Springer, Berlin-Heidelberg-New York, 1991.

[45] R. Rannacher: *Finite element methods for the incompressible Navier-Stokes equations.* in Fundamental Directions in Mathematical Fluid Mechanics, Galdi GP, Heywood J and Rannacher R (eds). Birkhäuser: Basel-Boston-Berlin, 2000.

[46] R. Rannacher: *Incompressible viscous flow*, in Encyclopedia of Computational Mechanics (E. Stein, R. de Borst, T.J.R.Hughes, eds), Volume 3 'Fluids', John Wiley, Chichester, 2004.

[47] R. Rannacher: *A Short Course on numerical simulation of viscous flow: discretization, optimization and stability analysis*, in Lecture Notes 12th school "Mathematical Theory in Fluid Mechanics", Karcov, Czech Republic, Spring 2011, AIMS, Discrete and Continuous Dynamical Systems - Series S, Vol. 5(6), pp. 1147–1194, 2012.

[48] R. Rannacher: *Pointwise convergence of finite element approximations to quasi-nonlinear elliptic boundary value problems on non-quasi-uniform meshes*, preprint, Heidelberg University, 2016.

[49] R. Rannacher and R. Scott: *Some optimal error estimates for piecewise linear finite element approximations*, Math. Comp. 31, 437-445 (1982).

[50] R. Rannacher and S. Turek: *Simple nonconforming quadrilateral Stokes element*, Numer. Meth. Part. Diff. Equ. 8, 97–111 (1992).