LECTURES

on

A POSTERIORI ERROR ANALYSIS

S. Repin

.....

University of Saarbrücken, 2008

Contents

Chapter 1

Introduction

1.1 A priori and a posteriori conceptions.

In the 20th century, the theory of differential equations was mainly developed in the context of

A Priori Conception

In it, for a problem

$$\mathcal{A}\mathbf{u} = \mathbf{f} \tag{1.1}$$

we must establish:

- I. Existence and uniqueness of u (mathematical correctness);
- II. Regularity (extra properties of u).

In the a priori conception, mathematical analysis of PDE's is concentrated on the EXACT solution u and its QUALITATIVE PROPERTIES.

However, quantitative analysis of solutions to PDE's generates new mathematical problems that are often quite different from those in I and II.

Explicitly, exact solutions of real-life models are known in exceptional cases.

Therefore, approximation methods offer the only way of QUANTITATIVE analysis of PDE's.

"Classical" way:
We study projections of (1.1) to sequences of finite-dimensional spaces

$$\boldsymbol{\mathcal{A}_h u_h = f_h} \quad \textbf{is a projection of} \quad \boldsymbol{\mathcal{A}u = f} \textbf{ on } \boldsymbol{V_h \subset V}, \tag{1.2}$$

compute approximations u_h numerically and try to justify their convergence to u.

In other words, in the a priori conception mathematical analysis of errors is mainly reduced to the question: HOW TO APPROXIMATE EXACT SOLUTION "IN PRINCIPLE".

Classical approximation theory (60'-80') says that

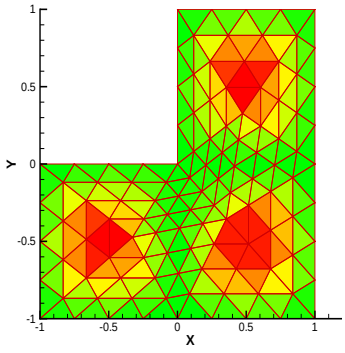
$$\|u - u_h\|_V \leq Ch^k, \quad C > 0, k > 0 \tag{1.3}$$

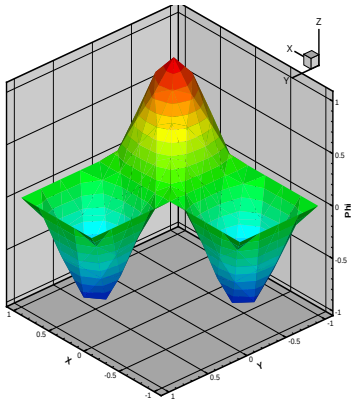
provided that

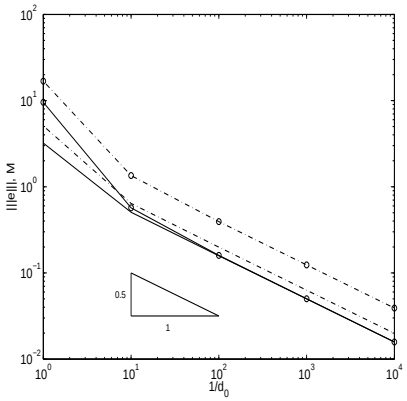
- **(a)** u **has an extra regularity;**
- **(b) all** V_h **are "regular" in some sense;**
- **(c)** u_h **is the exact finite dimensional solution.**

In practice, (a), (b), and (c) are often violated.
Even if they are satisfied, the constant C **is either unknown or highly overestimated. In spite of that, it was often (implicitly) assumed that if (I) and (II) have been positively solved and a priori convergence estimate has been established then the model *is valid for numerical analysis.* In other words, a priori conception views Numerical Experiment as the LAST (and "technical") step more related to engineers and codemakers than to mathematicians.**

Another widely speared "belief", is that results obtained by combining standard blocks (codes) give almost exact solutions provided that the dimensionality of the corresponding discrete problems is sufficiently large. Numerous standard codes and program complexes produce such results and represent them in a nice graphical form. We are suggested to believe in these pictures/numbers. Should we always believe?







The general principle of scientific objectivity suggests that the mathematical experiment must obey the same strict authenticity rules as those commonly accepted in natural sciences and we need to answer the question:

WHAT IS THE ACCURACY OF MATHEMATICAL EXPERIMENTS?
THEN WE WOULD KNOW WHAT THE DATA COMPUTED INDEED MEAN.

To understand the importance of this question the reader is offered to solve the problem below.

A "baby" coupled problem. Find z satisfying the differential equation

$$z'' - 9z' - 10z = 0, \quad z = z(x), \quad x \in [0, 8],$$
$$z(0) = 1,\ z'(0) = a_{N-1} - a_N,$$

where a is a solution of the system $Ba = f$ of the dimensionality N

$$b_{ij} = \frac{2S_i^2 S_j^2}{\pi} \int_0^\pi \big(\sin(i\xi)\sin(j\xi) + \sin(i + j^2)\xi\big)\, \mathrm{d}\xi,$$
$$i, j = 1, 2, ...N,\ f_i = (i+1)^4 i,\ S_i = \sum_{k=0}^{+\infty} \left(\frac{i}{i+1}\right)^k.$$

Task 1.1.1 *For $N = 10, 50, 100, 200$ find $z(8)$ analytically and compare with numerical results obtained by computing the sums numerically, finding definite integrals with help of quadratures formulas, solving the system of linear simultaneous equations by a numerical method, and integrating the differential equation by a certain (e.g., Euler) method.*

Errors arising in quantitative analysis of PDE's

U **Physical object/process**

$\Downarrow$

$\boldsymbol{\varepsilon_1}$ $\longrightarrow$ Error of a model

$\Downarrow$

u **Differential model** $Au = f$

$\Downarrow$

$\boldsymbol{\varepsilon_2}$ $\longrightarrow$ Approximation error

$\Downarrow$

u_h **Discrete model** $A^h u_h = f_h$

$\Downarrow$

$\boldsymbol{\varepsilon_3}$ $\longrightarrow$ Computational error

$\Downarrow$

u_h^ϵ **Numerical solution** $A^h u_h^\epsilon = f_h + \epsilon.$

MODELING ERROR

Let U be a physical value that characterizes some process and u be a respective value obtained from the mathematical model. Then the quantity

$$\epsilon_1 = |U - u|$$

is an **error of the mathematical model.**

Mathematical model always presents an "abridged" version of a physical object.
Therefore, $\epsilon_1 > 0$**.**

TYPICAL SOURCES OF MODELING ERRORS

- (a) "Second order" phenomena are neglected in a mathematical model.
- (b) Problem data are defined with an uncertainty.
- (c) Dimension reduction is used to simplify a model.

APPROXIMATION ERRORS

Let u_h be a solution on a mesh of the size h. Then, u_h encompasses the **approximation error**

$$\epsilon_2 = |u - u_h|.$$

Classical error control theory is mainly focused on approximation errors.

In the next section, we give a concise overview of the a priori asymptotic methods in error estimation.

NUMERICAL ERRORS

Finite–dimensional problems are also solved approximately, so that instead of u_h we obtain u_h^ϵ. The quantity

$$\epsilon_3 = |u_h - u_h^\epsilon|$$

shows an error of the numerical algorithm performed with a concrete computer. This error includes

- **roundoff errors,**
- **errors arising in iteration processes and in numerical integration,**
- **errors caused by possible defects in computer codes.**

Roundoff errors. **Numbers in a computer are presented in a floating point format:**

$$x = \pm\Big(\frac{i_1}{q} + \frac{i_2}{q^2} + ... + \frac{i_k}{q^k}\Big)q^\ell, \quad i_s < q.$$

These numbers form the set $R_{q\ell k} \subset \mathbb{R}$. q is the base of the representation, $\ell \in [\ell_1, \ell_2]$ is the power.

The set $R_{q\ell k}$ is not closed with respect to the operations $+, -, *$!

Operations with plane vectors ARE RESTRICTED to those associated with the dots marked! Certainly, in modern computers the amount of dots is much bigger, but the principal structure is the same.

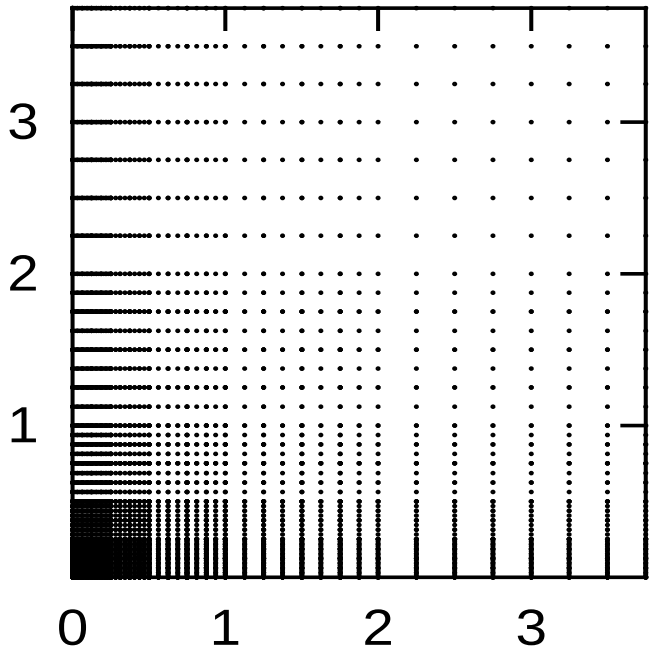


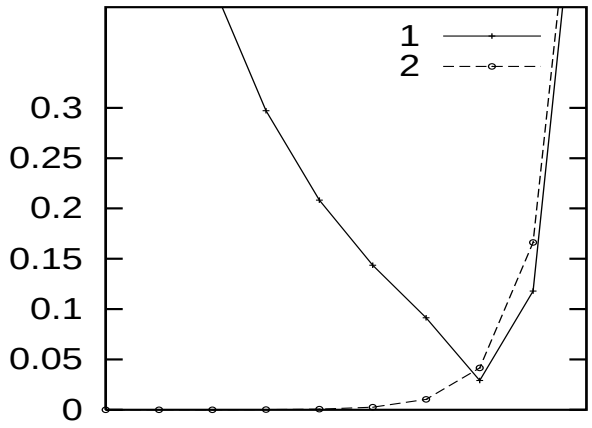
Figure 1.1: The set $R_{q\ell k} \times R_{q\ell k}$

Example.

$$
\begin{aligned}
&k = 3,\\
&a = \left(\frac{1}{2} + 0 + 0\right) * 2^5, \qquad b = \left(\frac{1}{2} + 0 + 0\right) * 2^1\\
&b \Rightarrow \left(0 + \frac{1}{2} + 0\right) * 2^2 \Rightarrow \left(0 + 0 + \frac{1}{2}\right) * 2^3 \Rightarrow (0 + 0 + 0) * 2^4
\end{aligned}
$$

$$\mathbf{a + b = a}!?$$

Definition 1.1.1 *The smallest floating point number which being added to 1 gives a quantity other than 1 is called* **the machine accuracy**.



Numerical integration

$$\int_{\mathbf{b}}^{\mathbf{a}} \mathbf{f(x)dx} \cong \sum_{\mathbf{i=1}}^{\mathbf{n}} \mathbf{c_i f(x_i)h} = \sum_{\mathbf{i=1}}^{\mathbf{n/2}} \mathbf{c_i} \overset{\sim 1}{\mathbf{f(x_i)h}} + \mathbf{c_{n/2+1}} \overset{\sim \delta}{\mathbf{f(x_{n/2+1})h}} + ...$$

Two principal classes of problems in Mathematical Modeling

I. Computations on the basis of a reliable (certified) model.

Here modeling error ϵ_1 is assumed to be small and u_h^ϵ gives a desired information on U.

Then

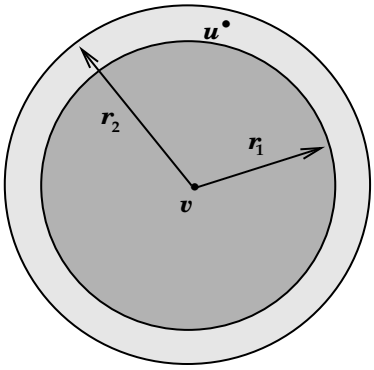
$$\|\mathbf{U} - \mathbf{u}_\mathbf{h}^\epsilon\| \leq \quad \epsilon_1 + \boxed{\epsilon_2 + \epsilon_3}. \tag{1.4}$$

II. Verification of a mathematical model.

Here physical data U and numerical data u_h^ϵ are compared to judge on the quality of a mathematical model

$$\|\epsilon_1\| \leq \|\mathbf{U} - \mathbf{u}_\mathbf{h}^\epsilon\| + \boxed{\epsilon_2 + \epsilon_3}. \tag{1.5}$$

Thus, two major problems of mathematical modeling, namely, **reliable computer simulation**, and **verification of mathematical models by comparing physical and mathematical experiments**, require efficient methods able to provide COMPUTABLE AND REALISTIC estimates of $\boxed{\epsilon_2 + \epsilon_3}$.



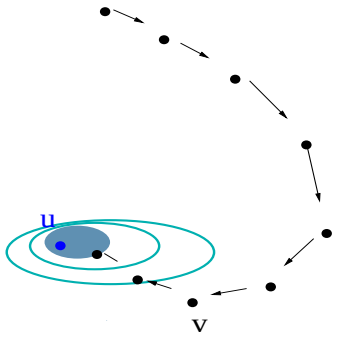
SUMMARY: Reliable QUANTITATIVE ANALYSIS of PDE's cannot be performed without solving the following

MAIN ERROR CONTROL PROBLEM

Given the data (coefficients,a domain, boundary conditions) of a boundary-value problem and a function v from the corresponding (energy) space V, compute the radii r_1 and r_2 of two balls $\mathcal{B}(v, r_1)$ and $\mathcal{B}(v, r_2)$ centered at v such that

$$u \notin \mathcal{B}(v, r_1) \quad \text{and} \quad u \in \mathcal{B}(v, r_2). \tag{1.6}$$

We say that a method used to solve the above problem is *sharp* if one can find r_1 and r_2 such that $r_2 - r_1 \leq \epsilon$ for any given ϵ.



If we wish to QUALITATIVELY analyze models based on PDE's then in addition to problems (I) and (II) we need to solve

- **Problem III. Find computable estimates for NEIGHBORHOODS of a generalized solution u.**

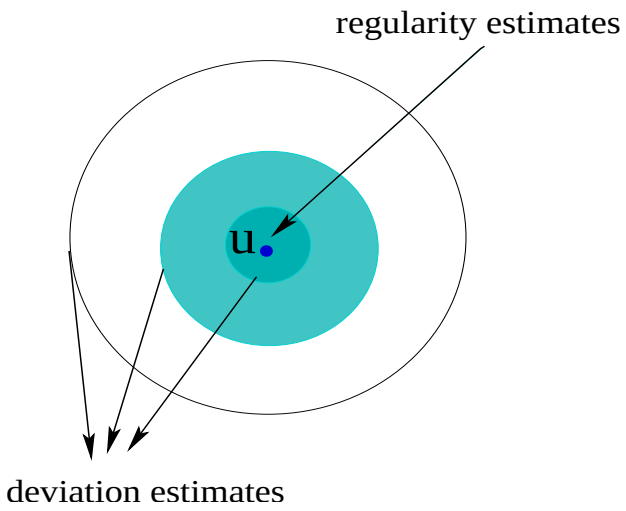
If neighborhoods are generated by the topology of energy (Banach) space V, then we need to have estimates

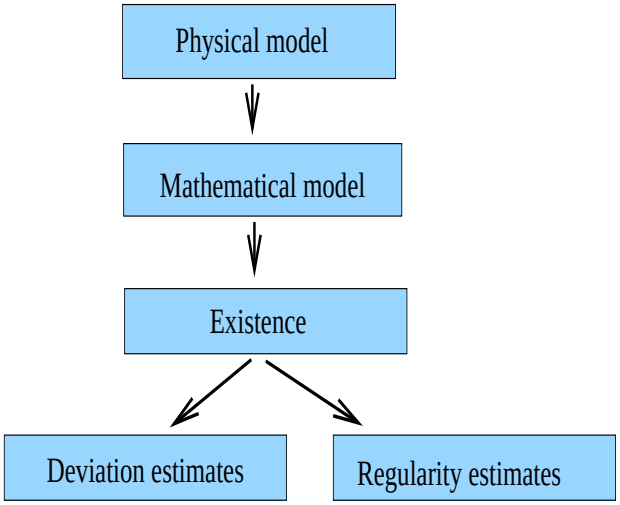
$$\underline{\mathfrak{M}}\ (v, \mathcal{D}) \le \|u - v\|_V \le \overline{\mathfrak{M}}\ (v, \mathcal{D}), \qquad \forall v \in V, \tag{1.7}$$

where $\mathcal{D}$ denotes the set of known data and the functionals $\overline{\mathfrak{M}}$ (error majorant) and $\underline{\mathfrak{M}}$ (error minorant) which must be

- **directly computable;**
- **valid for any admissible approximations;**
- **do not attract special (e.g. extra regularity) properties of u or u_h.**

We call (1.7) guaranteed A POSTERIORI ESTIMATES or DEVIATION ESTIMATES.





In a sense Problem III is opposite to the regularity analysis. In spite of its clear practical meaning, it is much less investigated than existence (I) and regularity (II).

1.2 Mathematical background and notation

1.2.1 Vectors and tensors

By $\mathbb{R}^d$ and $\mathbb{M}^{d\times d}$ we denote the spaces of real d-dimensional vectors and $d\times d$ matrices, respectively. The scalar product of vectors is denoted by $\cdot$, and for the product of tensors we use the symbol :, i.e.,

$$u \cdot v = u_i v_i, \qquad \tau : \sigma = \tau_{ij}\sigma_{ij},$$

where summation (from 1 to d) over repeated indices is implied. The norms of vectors and tensors are defined as follows:

$$|a| := \sqrt{a \cdot a}, \qquad |\sigma| := \sqrt{\sigma : \sigma}.$$

Henceforth, the symbol := means "equals by definition". The multiplication of a matrix $A \in \mathbb{M}^{d\times d}$ and a vector $b \in \mathbb{R}^d$ is the vector, which we denote Ad. Matrixes are usually denoted by capital letters (matrixes associated with stresses and strains are denoted by Greek letters σ and ε). Any tensor τ is decomposed into the *deviatoric* part τ^{D} and the *trace* $\operatorname{tr}\tau := \tau_{ii}$, so that $\tau := \tau^{\mathrm{D}} + \frac{1}{d}\mathbb{I}\operatorname{tr}\tau$, where $\mathbb{I}$ is the unit tensor. It is easy to check that

$$\tau : \mathbb{I} = \operatorname{tr}\tau, \quad \tau^{\mathrm{D}} : \mathbb{I} = 0, \tag{1.8}$$

$$|\tau|^2 = |\tau^{\mathrm{D}}|^2 + \frac{1}{d}\operatorname{tr}\tau^2, \tag{1.9}$$

so that τ is decomposed into two parts (which sometimes are called deviatorical and spherical).

We will use the algebraic Young's inequality

$$2ab \le \beta a^2 + \frac{1}{\beta}b^2, \tag{1.10}$$

which is valid for any $\beta > 0$.

For any pair of vectors a and b and any $\beta > 0$ we have a similar estimate

$$2a \cdot b \le \beta|a|^2 + \frac{1}{\beta}|b|^2, \tag{1.11}$$

which implies the inequalities

$$|a+b|^2 \le (1+\beta)|a|^2 + \frac{1+\beta}{\beta}|b|^2, \tag{1.12}$$

$$|a+b|^2 \ge \frac{1}{1+\beta}|a|^2 - \frac{1}{\beta}|b|^2. \tag{1.13}$$

Similarly, for a pair of tensors σ and τ we have

$$2\tau : \sigma \le \beta|\tau|^2 + \frac{1}{\beta}|\sigma|^2, \tag{1.14}$$

$$|\tau+\sigma|^2 \le (1+\beta)|\tau|^2 + \frac{1+\beta}{\beta}|\sigma|^2. \tag{1.15}$$

If H is a Hilbert space with scalar product $(.,.)$ and norm $\|\,.\,\|$ associated with the product, then it is easy to extend (1.11)–(1.13) to the elements of H.

The inequality (1.8) is a particular form of the more general Young's inequality

$$ab \le \frac{1}{p}(\beta a)^p + \frac{1}{p'}\left(\frac{b}{\beta}\right)^{p'}, \qquad \frac{1}{p}+\frac{1}{p'} = 1. \tag{1.16}$$

Another integral relation is

$$\int\limits_\Omega \operatorname{curl} a \cdot v\, dx = \int\limits_\Omega a \cdot \operatorname{curl} v\, dx - \int\limits_\Gamma (a \times n)\cdot v\, ds. \tag{1.17}$$

In the above relations, we assume that the functions are sufficiently regular so that the corresponding volume and surface integrals exist."

1.2.2 Spaces of functions

We denote a bounded connected domain in $\mathbb{R}^d$ by Ω and its boundary (which is assumed to be Lipschitz continuous) by Γ. Usually, ω stands for an open subset of Ω. The closure of sets is denoted by a bar and the Lebesgue measure of a set ω by the symbol $|\omega|$.

By $L^p(\omega)$ we denote the space of functions summable with power p with norm

$$\|w\|_{p,\omega} := \left(\int_\omega |w|^p dx \right)^{1/p}.$$

Also, we use the simplified notation

$$\|w\|_p := \|w\|_{p,\Omega}, \qquad \|w\| := \|w\|_{2,\Omega}.$$

The vector-valued functions with components that are square summable in Ω form the Hilbert space $L^2(\Omega, \mathbb{R}^d)$. Analogously, $L^2(\Omega, \mathbb{M}^{d\times d})$ is the Hilbert space of tensor-valued functions (sometimes we use the special notation Σ for this space). If tensor-valued functions are assumed to be symmetric, then we write $\mathbb{M}_s^{d\times d}$ (and Σ_s instead of $L^2(\Omega, \mathbb{M}_s^{d\times d})$). For $v \in L^2(\Omega, \mathbb{R}^d)$ and $\tau \in L^2(\Omega, \mathbb{M}^{d\times d})$, the norms are defined by the relations

$$\|v\|^2 := \int_\Omega |v|^2 dx \quad \text{and} \quad \|\sigma\|^2 := \int_\Omega |\sigma|^2 dx.$$

Since no confusion may arise, we denote the norm of $L^2(\Omega)$ and the norm of the space $L^2(\Omega, \mathbb{R}^d)$ by $\|\cdot\|$. The space of measurable essentially bounded functions is denoted by $L^\infty(\Omega)$. It is equipped with the norm

$$\|u\|_{\infty,\Omega} = \operatorname{ess\,sup}_{x\in\Omega} |u(x)|.$$

By $\overset{\circ}{C}{}^\infty(\Omega)$ we denote the space of all infinitely differentiable functions with compact supports in Ω. The spaces of k-times differentiable scalar- and vector-valued functions are denoted by $C^k(\Omega)$ and $C^k(\Omega, \mathbb{R}^d)$, respectively; $\overset{\circ}{C}{}^k(\Omega)$ is the subspace of $C^k(\Omega)$ that contains functions vanishing at the boundary; $P^k(\Omega)$ denotes the set of polynomial functions defined in $\Omega \subset \mathbb{R}^d$, i.e., $v \in P^k(\Omega)$ if

$$v = \sum_{|\alpha|\le m} a_\alpha x^\alpha, \quad m \le k,$$

where $\alpha := (\alpha_1, \dots, \alpha_d)$ is the so-called multi-index,

$$|\alpha| = \alpha_1 + \alpha_2 + \ldots + \alpha_d, \qquad a_\alpha = a_{\alpha_1,\ldots\alpha_d},$$

and $x^\alpha = x^{\alpha_1} x^{\alpha_2} \dots x^{\alpha_d}$.

For partial derivatives we keep the standard notation and write

$$\frac{\partial f}{\partial x_i} \quad \text{or} \qquad f_{,i}.$$

Usually, we understood them in a generalized sense: a function $g = f_i$ is called the generalized derivative of $f \in L^1(\Omega)$ with respect to the x_i if it satisfies the relation

$$\int_\Omega f w_{,i} dx = -\int_\Omega g w dx, \qquad \forall w \in C_0^1(\Omega). \tag{1.18}$$

Generalized derivatives of higher orders are defined by similar integral relations.

By $\{\!\{g\}\!\}_S$ we denote the mean value of a function g on S, i.e.,

$$\{\!\{g\}\!\}_S := \frac{1}{|S|} \int\limits_S g \, dx$$

and $\widetilde{g}_S := g - \{\!\{g\}\!\}_S$. The functions with zero mean form the space

$$\widetilde{L}^2(\Omega) := \left\{ q \in Q \,\middle|\, \{\!\{q\}\!\}_\Omega = 0 \right\}.$$

The space $H(\Omega, \mathrm{div}\,)$ is a subspace of $L^2(\Omega, \mathbb{R}^d)$ that contains vector-valued functions with square-summable divergence, and $H(\Omega, \mathrm{Div}\,)$ is a subspace of Σ that contains tensor-valued functions with square-summable divergence, i.e.,

$$\begin{aligned} H(\Omega, \mathrm{div}\,) &:= \{v \in L^2(\Omega, \mathbb{R}^d) \quad | \ \mathrm{div}\, v := \{v_{i,i}\} \in L^2(\Omega)\}, \\ H(\Omega, \mathrm{Div}\,) &:= \{\tau \in L^2(\Omega, \mathbb{M}^{d\times d}) \ | \ \mathrm{Div}\, \tau := \{\tau_{ij,j}\} \in L^2(\Omega, \mathbb{R}^d)\}. \end{aligned}$$

Both spaces $H(\Omega, \mathrm{div}\,)$ and $H(\Omega, \mathrm{Div}\,)$ are Hilbert spaces endowed with scalar products

$$(u, v)_{\mathrm{div}} := \int_\Omega (u \cdot v + \mathrm{div}\, u \, \mathrm{div}\, v) dx$$

and

$$(\sigma, \tau)_{\mathrm{Div}} := \int_\Omega (\sigma : \tau + \mathrm{Div}\, \sigma \cdot \mathrm{Div}\, \tau) dx,$$

respectively. The norms $\|\cdot\|_{\rm div}$ and $\|\cdot\|_{\rm Div}$ are associated with the above-defined scalar products.

Similarly, $H(\Omega, {\rm curl}\,)$ is the Hilbert space of vector-valued functions having square-summable rotor, i.e.,

$$H(\Omega, {\rm curl}\,) := \{v \in L^2(\Omega, \mathbb{R}^d) \quad | \ {\rm curl}\, v \in L^2(\Omega)\},$$

where ${\rm curl}\, v := \big(v_{3,2} - v_{2,3}; v_{1,3} - v_{3,1}; v_{2,1} - v_{1,2}\big)$. This space can be defined as the closure of smooth functions with respect to the norm

$$\|w\|_{\rm curl} := \big(\|w\|^2_\Omega + \|{\rm curl}\, w\|^2_\Omega\big)^{1/2}.$$

The Sobolev spaces[1] $W^{m,p}(\Omega)$ (where m and p are positive integer numbers) contain functions summable with power p the generalized derivatives of which up to order l belong to L^p. For a function $f \in W^{m,p}(\Omega)$, the norm is defined as usual:

$$\|f\|_{m,p,\Omega} = \left(\int\limits_\Omega \sum_{|\alpha|\le m} |D^\alpha f|^p \, dx\right)^{1/p}.$$

Here $\alpha = \{\alpha_1, \dots \alpha_2\}$ is the multi index and

$$D^\alpha v = \frac{\partial^{|\alpha|} v}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}$$

is the derivative of order $|\alpha|$.

The Sobolev spaces with $p = 2$ are denoted by the letter H, i.e.,

$$H^m(\Omega) := \left\{v \in L^2(\Omega) \ | \ D^\alpha v \in L^2(\Omega), \quad \forall m: \ |\alpha| \le m\right\}.$$

These spaces belong to the class of Hilbert spaces. A subset of $H^m(\Omega)$ formed by the functions vanishing on Γ is denoted by $\overset{\circ}{H}{}^m(\Omega)$.

[1]Introduced in S. L. Sobolev. *Some Applications of Functional Analysis in Mathematical Physics*, Izdt. Leningrad. Gos. Univ., Leningrad, 1955,
English version: Translation of Mathematical Monographs, Volume 90, American Mathematical Society, Providence, RI, 1991.

Embedding Theorems

Relationships between the Sobolev spaces and $L^p(\Omega)$ and $C^k(\Omega)$ are given by *Embedding Theorems*.

Theorem 1.2.1 *If* $p, q \geq 1$, $\ell > 0$ *and* $\ell + \frac{n}{q} \geq \frac{n}{p}$, *then* $W^{\ell,p}(\Omega)$ *is continuously embedded in* $L^q(\Omega)$. *Moreover, if* $\ell + \frac{n}{q} > \frac{n}{p}$, *then the embedding operator is compact.*

Theorem 1.2.2 *If* $\ell - k > \frac{n}{p}$, *then* $W^{\ell,p}(\Omega)$ *is compactly embedded in* $C^k(\overline{\Omega})$.

1.2.3 Boundary traces

The functions in Sobolev spaces have counterparts on Γ (and on other manifolds of lower dimensions) that are associated with spaces of *traces*. Thus, there exist some bounded operators mapping the functions defined in Ω to functions defined on the boundary. For example, the operator $\gamma : H^1(\Omega) \to L^2(\Gamma)$ is called the *trace* operator if it satisfies the following conditions:

$$\gamma v = v \mid_{\Gamma}, \qquad \forall v \in C^1(\Omega), \tag{1.19}$$

$$\|\gamma v\|_{2,\Gamma} \leq c_{T\Gamma} \|v\|_{1,2,\Omega}, \tag{1.20}$$

where $c_{T\Gamma}$ is a positive constant independent of v. From these relations, we observe that γv is a natural generalization of the trace defined for a continuous function (in the pointwise sense). The image of γ is a subset of $L^2(\Gamma)$, which is the space $H^{1/2}(\Gamma)$. The functions from other Sobolev spaces are also known to have traces in Sobolev spaces with fractional indices. Thus, $\gamma \in \mathcal{L}\left(H^1(\Omega), H^{1/2}(\Gamma)\right)$ and the space $H_0^1(\Omega)$ is the kernel of γ.

Also, for any $\phi \in H^{1/2}(\Gamma)$, one can define a continuation operator

$$\mu \in \mathcal{L}(H^{1/2}(\Gamma), H^1(\Omega))$$

such that

$$\mu\phi = w, \qquad w \in H^1(\Omega), \qquad \gamma w = \phi \qquad \text{on } \Gamma$$

and

$$\|\phi\|_{H^{1/2},\Gamma} \le c_\gamma \|w\|_{1,2,\Omega}\,, \qquad \|w\|_{1,2,\Omega} \le c_\mu \|\phi\|_{H^{1/2},\Gamma}\,. \tag{1.21}$$

Using the operator γ, we define subspaces of functions vanishing on Γ or on some part Γ_1 of Γ. Usually, such subspaces are marked by the zero subindex, e.g.,

$$V_0 := \{v \in V \mid \gamma v = 0 \text{ a.e. on } \Gamma_1\}\,,$$

Henceforth, we understand the boundary values of functions in the sense of traces, so that the phrase "$u = \phi$ on Γ" means that the trace γu of a function u defined in Ω coincides with a given function ϕ defined on Γ (for the sake of simplicity, we usually omit γ). If for two functions u and v defined in Ω we say that $u = v$ on Γ, then we mean that $\gamma(u - v) = 0$ on Γ.

1.2.4 Generalized derivatives and Sobolev spaces with negative indexes

For $f \in L^2(\Omega)$, the functional

$$\langle f_{,i}, \varphi\rangle := -\int_\Omega f \frac{\partial \varphi}{\partial x_i}\, dx \tag{1.22}$$

is linear and continuous not only for functions in $\overset{\circ}{C}{}^\infty(\Omega)$ but also for all functions of the space $\overset{\circ}{H}{}^1(\Omega)$ (this fact follows from the density of smooth functions in $\overset{\circ}{H}{}^1(\Omega)$ and known theorems on the continuation of linear functionals). Such functionals can be viewed as generalized derivatives of square summable functions. They form the space $H^{-1}(\Omega)$ dual to $\overset{\circ}{H}{}^1(\Omega)$. It is easy to see that the quantity

$$\big|\!\big| f_{,i} \big|\!\big| := \sup_{\substack{\varphi \in \overset{\circ}{H}{}^1(\Omega) \\ \varphi \neq 0}} \frac{|\langle f_{,i}, \varphi\rangle|}{\|\nabla \varphi\|_\Omega} \tag{1.23}$$

is nonnegative and finite. It can be used to introduce the norm for $H^{-1}(\Omega)$.

1.2.5 Functional inequalities

In the subsequent chapters, we use several inequalities well known in functional analysis. For convenience of the reader, we collect and discuss them below.

First, we recall the inequality

$$|a \cdot b| \leq \left(\sum_{i=1}^{d} |a_i|^{\alpha} \right)^{\frac{1}{\alpha}} \left(\sum_{i=1}^{d} |b_i|^{\alpha'} \right)^{\frac{1}{\alpha'}}, \tag{1.24}$$

where $\frac{1}{\alpha'} + \frac{1}{\alpha} = 1$ and $a, b \in \mathbb{R}^d$. It is known as the discrete Hölder inequality. The Hölder inequality in functional form is as follows:

$$\int_{\Omega} uv dx \leq \|u\|_{\alpha,\Omega} \|v\|_{\alpha',\Omega}. \tag{1.25}$$

Let u and v be two functions in $L^{\alpha}(\Omega)$. Then

$$\begin{aligned} \int_{\Omega} (u+v)^{\alpha} dx = \int_{\Omega} u(u+v)^{\alpha-1} dx + \int_{\Omega} v(u+v)^{\alpha-1} dx \leq \\ \leq \|u\|_{\alpha,\Omega} \left(\int_{\Omega} (u+v)^{(\alpha-1)\alpha'} \right)^{1/\alpha'} + \|v\|_{\alpha,\Omega} \left(\int_{\Omega} (u+v)^{(\alpha-1)\alpha'} \right)^{1/\alpha'} = \\ = (\|u\|_{\alpha,\Omega} + \|v\|_{\alpha,\Omega}) \left(\int_{\Omega} (u+v)^{(\alpha-1)\alpha'} \right)^{(\alpha-1)/\alpha} \end{aligned}$$

and we arrive at the Minkovski inequality

$$\left(\int_{\Omega} (u+v)^{\alpha} dx \right)^{1/\alpha} \leq \|u\|_{\alpha,\Omega} + \|v\|_{\alpha,\Omega} \tag{1.26}$$

For the functions in $\overset{\circ}{H}{}^1(\Omega)$, we have the Friedrichs inequality

$$\|w\|_{\Omega} \leq C_{F\Omega} \|\nabla w\|_{\Omega}, \quad \forall w \in \overset{\circ}{H}{}^1(\Omega), \tag{1.27}$$

where $C_{F\Omega}$ is a positive constant independent of w. It is not difficult to observe that the constant in (1.27) satisfies the relation

$$\frac{1}{C_{F\Omega}} = \lambda_{\Omega} := \inf_{\substack{w \in \overset{\circ}{H}{}^1(\Omega) \\ w \neq 0}} \frac{\|\nabla w\|}{\|w\|}. \tag{1.28}$$

Let $\Omega \subset \widehat{\Omega}$. For any $w \in \overset{\circ}{H}{}^1(\Omega)$, we can define $\widehat{w} = w$ in Ω and $\widehat{w}(x) = 0$ for any $x \in \widehat{\Omega} \setminus \Omega$. Obviously, $\widehat{w} \in \overset{\circ}{H}{}^1(\widehat{\Omega})$. Therefore,

$$\lambda_\Omega \geq \inf_{\substack{\widehat{w} \in \overset{\circ}{H}{}^1(\widehat{\Omega}) \\ \widehat{w} \neq 0}} \frac{\|\nabla \widehat{w}\|}{\|\widehat{w}\|} = \lambda_{C_{\widehat{\Omega}}} = \frac{1}{C_{F\widehat{\Omega}}}$$

and we conclude that $C_{F\Omega} \leq C_{F\widehat{\Omega}}$.

Assume that

$$\Omega \subset \Pi := \{x \in \mathbb{R}^d \mid a_i < x < b_i, \quad b_i - a_i = l_i\}.$$

Then,

$$\lambda_\Omega \geq \lambda_\Pi = \pi \sqrt{\sum_i \frac{1}{l_i^2}},$$

and we obtain an explicit upper bound for $C_{F\Omega}$.

For $w \in H^1(\Omega)$, the Friedrichs inequality has a more general form

$$\|w\|_\Omega^2 \leq c_{F\Omega}^2 \left(\|\nabla w\|_\Omega^2 + \int_\Gamma |w|^2 \, ds \right). \tag{1.29}$$

where $C_{F\Omega}^2$ can be estimated from above by the quantity

$$\frac{1}{\pi^2 \sum\limits_i \frac{1}{l_i^2}} \max\{1, \bar{c}\}, \quad \text{where} \quad \bar{c} = \max_\Gamma \left\{ \frac{1}{\phi} |\nabla \phi \cdot n| \right\}$$

and $\phi(x)$ is the first eigenfunction of the Laplace operator in Π The reader can find estimates of the constants in Friedrichs inequality in the books by S. Mikhlin.

For $w \in H^1(\Omega)$, the Poincaré inequality reads

$$\|w\|_\Omega^2 \leq C_{P\Omega}^2 \left(\|\nabla w\|_\Omega^2 + \Big(\int_\Omega w \, ds \Big)^2 \right). \tag{1.30}$$

From (1.30) it follows that

$$\|w\|_\Omega \le C_{P\Omega}\|\nabla w\|_\Omega \qquad \forall w \in \widetilde{L}^2(\Omega). \tag{1.31}$$

If

$$\Omega = \Pi_l := \{x \in \mathbb{R}^d \; x_i \in (0, l_i), \; l_i > 0\},$$

then the Poincaré inequality takes the form

$$\|w\|^2_\Omega \le \frac{1}{|\Pi_l|}\left(\int\limits_{\Pi_l} w\, dx\right)^2 + \frac{d}{2}\int\limits_{\Pi_l}\sum_{i=1}^{d} l_i^2 w_{,i}^2 dx. \tag{1.32}$$

In continuum mechanics, of importance is the following assertion known as the Korn's inequality. Let Ω be an open, bounded domain with Lipschitz continuous boundary. Then

$$\int\limits_\Omega \left(|w|^2 + |\varepsilon(w)|^2\right) dx \ge C_{K\Omega}\|w\|^2_{1,2,\Omega} \quad \forall w \in H^1(\Omega, \mathbb{R}^d), \tag{1.33}$$

where $C_{K\Omega}$ is a positive constant independent of w and $\varepsilon(w)$ denotes the symmetric part of the tensor ∇w, i.e.,

$$\varepsilon_{ij}(w) = \frac{1}{2}\left(\frac{\partial w_i}{\partial x_j} + \frac{\partial w_j}{\partial x_i}\right).$$

It is not difficult to verify that the left-hand side of (1.33) is bounded from above by the H^1-norm of w. Thus, it represents a norm equivalent to $\|.\|_{1,2,\Omega}$. The kernel of $\varepsilon(w)$ is called the space of rigid deflections and is denoted by $\mathbf{R}(\Omega)$. If $w \in \mathbf{R}(\Omega)$, then it can be represented in the form $w = w_0 + \omega_0 x$, where w_0 is a vector independent of x and ω_0 is a skew-symmetric tensor with coefficients independent of x. It is easy to understand that the dimension of $\mathbf{R}(\Omega)$ is finite and equals $d + \frac{d(d-1)}{2}$.

For the functions in $\overset{\circ}{H}{}^1(\Omega)$, the Korn's inequality is easy to prove. Indeed,

$$|\varepsilon(w)|^2 = \frac{1}{4}(w_{i,j} + w_{j,i})(w_{i,j} + w_{j,i}) =$$
$$= \frac{1}{4}(w_{i,j}w_{i,j} + w_{j,i}w_{j,i} + 2w_{i,j}w_{j,i}) = \frac{1}{2}(|\nabla w|^2 + w_{i,j}w_{j,i}),$$

where the summation over repeated indices is implied. Therefore, for any $w \in \overset{\circ}{C}{}^2(\Omega)$ we have

$$\int_\Omega |\varepsilon(w)|^2 dx = \frac{1}{2}\int_\Omega \Big(|\nabla w|^2 + w_{i,j}w_{j,i}\Big)dx = \frac{1}{2}\int_\Omega \Big(|\nabla w|^2 - w_i w_{j,ij}\Big)dx =$$
$$= \frac{1}{2}\int_\Omega \Big(|\nabla w|^2 + w_{i,i}w_{j,j}\Big)dx = \frac{1}{2}\int_\Omega \Big(|\nabla w|^2 + |w_{i,i}|^2\Big)dx \geq \frac{1}{2}\|\nabla w\|^2.$$

Hence,

$$\|\nabla w\| \leq \sqrt{2}\|\varepsilon(w)\| \qquad \forall w \in \overset{\circ}{C}{}^2(\Omega). \tag{1.34}$$

Since $\overset{\circ}{C}{}^2(\Omega)$ is dense in $\overset{\circ}{H}{}^1(\Omega)$, this inequality is also valid for functions in $\overset{\circ}{H}{}^1(\Omega)$. The proofs of the Korn's inequality (1.33) are much more complicated.

1.2.6 Convex analysis

Convex sets and functions

Consider a Banach space V. A set $K \subset V$ is called *convex* if $\lambda_1 v_1 + \lambda_2 v_2 \in K$ for all $v_1, v_2 \in K$ and all $\lambda_1, \lambda_2 \in \mathbb{R}_+$ such that $\lambda_1 + \lambda_2 = 1$.

Convex hull $\mathrm{conv}K$ is the set of all convex combinations of all the elements of K, i.e.,

$$\mathrm{conv}K = \left\{ v \in V \mid v = \sum_{i=1}^m \lambda_i v_i,\ v_i \in K,\ \sum_{i=1}^m \lambda_i = 1,\ \lambda_i \geq 0 \right\}.$$

It is obvious that $K = \mathrm{conv}K$ if and only if K is a convex set. Let K be a convex set. A functional $J : K \to \mathbb{R}$ is said to be *convex* if

$$J(\lambda_1 v_1 + \lambda_2 v_2) \leq \lambda_1 J(v_1) + \lambda_2 J(v_2) \tag{1.35}$$

for all $v_1, v_2 \in K$ and all $\lambda_1, \lambda_2 \in \mathbb{R}_+$ such that $\lambda_1 + \lambda_2 = 1$. A functional J is called *strictly convex* if

$$J(\lambda_1 v_1 + \lambda_2 v_2) < \lambda_1 J(v_1) + \lambda_2 J(v_2) \tag{1.36}$$

for all $v_1, v_2 \in K$ (such that $v_1 \neq v_2$) and $\lambda \in (0, 1)$.

A functional J is called *concave* (resp., *strictly concave*) if the functional $(-J)$ is convex (resp., strictly convex).

The functional

$$\chi_K(v) = \begin{cases} 0 & \text{if } v \in K, \\ +\infty & \text{if } v \notin K \end{cases}$$

is called the *characteristic functional* of the set K. It is clear that it is convex if and only if the set K is convex.

Definition 1.2.1 *A functional $J : V \to \overline{\mathbb{R}}$ is called* proper *if $J(v) > -\infty$ for any $v \in V$ and $J \not\equiv +\infty$.*

Any functional $J : V \to \mathbb{R}$ is characterized by two sets:

$$domJ := \{v \in V \mid J(v) < +\infty\},$$
$$epiJ := \{(v, \alpha) \mid v \in V,\ \alpha \in \mathbb{R},\ J(v) \le \alpha\}.$$

The first set contains elements (functions) that give finite values to J. The second one (called *supergraph* or *epigraph*) consists of "points" $(v, \alpha) \in V \times \overline{\mathbb{R}}$ that lie "above" the graph.

The set $epiJ$ is convex if and only if J is a convex functional.

Proposition 1.2.1 *If J is a proper convex functional, then*

$$V_\alpha := \{v \in V \mid J(v) \le \alpha,\ \alpha \in \mathbb{R}\}$$

is convex.

Operations with convex sets and functionals

Intersection of convex sets K_1 and K_2 is a convex set $K_1 \cap K_2$. If J_1 and J_2 are two convex functionals defined on a convex set K then the functionals $\alpha_1 J_1 + \alpha_2 J_2$ (for $\alpha_1, \alpha_2 \in \mathbb{R}_+$) and $\max\{J_1, J_2\}$ are also convex. It is worth noting that the latter fact remains valid for any amount of convex functionals, i.e., the upper bound taken over any set of convex functionals is a convex functional. Therefore, convex functionals are often represented as upper bounds of affine functionals.

By definition, the space V^* consists of all linear continuous functionals on V. It is called *topologically dual* to V. The value of $v^* \in V^*$ on $v \in V$ is denoted by $\langle v^*, v\rangle$. This product generates a *duality pairing* of the spaces V and V^*. If V is a Banach space, then V^* can also be normed by setting

$$\|v^*\|_* := \sup_{v\in V} \frac{\langle v^*, v\rangle}{\|v\|}. \tag{1.37}$$

Henceforth, we assume that the supremum (or infimum) of a quotient is taken with respect to all elements of V, except for the zero element $\mathcal{O}_V$.

Any affine functional defined on elements of V has the form $\langle v^*, v\rangle - \alpha$, where $v^* \in V^*$ and $\alpha \in \mathbb{R}$.

A functional space is called *reflexive* if it coincides with the bidual space V^{**} (i.e., if there exists a one-to-one mapping of V to V^{**} and back that preserves the metric). All the Hilbert spaces are reflexive. The same is true for the spaces L^p with $1 < p < +\infty$.

The theorem of F. Riesz asserts that for Hilbert spaces, any functional $v^* \in V^*$ can be written in the form of a scalar product introduced in such a space, i.e.,

$$(u, v) \;=\; \langle v^*, v\rangle, \qquad \forall v \in V, \tag{1.38}$$

where u is uniquely determined.

The functional $J^* : V^* \to \mathbb{R}$ defined by the relation

$$J^*(v^*) = \sup_{v\in V}\{\langle v^*, v\rangle - J(v)\} \tag{1.39}$$

is said to be *dual* (or *conjugate*) to J.

Remark 1.2.1 *If J is a smooth function that increases at infinity faster than any linear function, then J^* is the Legendre transform of J. The dual functionals were studied by Young, Fenchel, Moreau, and Rockafellar.*

The functional J^* is also called polar *to J.*

The functional

$$J^{**}(v) = \sup_{v^*\in V^*} \{\langle v^*, v\rangle - J^*(v^*)\} \tag{1.40}$$

is called the *second conjugate* to J (or *bipolar*). If J is a convex functional attaining finite values, then J coincides with J^{**}.

To illustrate the definitions of conjugate functionals, consider functionals defined on the Euclidean space E^d. In this case, V and V^* consist of the same elements: d-dimensional vectors (denoted by ξ and ξ^*, respectively) and the quantity $\langle \xi^*, \xi \rangle$ is given by the scalar product $\xi^* \cdot \xi$.

Let $A = \{a_{ij}\}$ be a positive definite matrix. We have the following pair of mutually conjugate functionals:

$$J(\xi) = \frac{1}{2} A\xi \cdot \xi \quad \text{and} \quad J^*(\xi^*) = \frac{1}{2} A^{-1}\xi^* \cdot \xi^*. \tag{1.41}$$

Another example is given by the functionals

$$J(\xi) = \frac{1}{\alpha}|\xi|^\alpha \quad \text{and} \quad J^*(\xi^*) = \frac{1}{\alpha'}|\xi^*|^{\alpha'}, \tag{1.42}$$

where $\frac{1}{\alpha} + \frac{1}{\alpha'} = 1$. If φ is an odd convex function, then $(\varphi(\|u\|_V))^* = \varphi^*(\|u^*\|_{V^*})$.

Subdifferential

Let a functional $J : V \to \mathbb{R}$ takes a finite value at $v_0 \in V$. The functional J is called subdifferentiable at v_0 if there exists an affine minorant l such that $J(v_0) = l(v_0)$. A minorant with this property is called the exact minorant at v_0.

Obviously, any affine minorant exact at v_0 has the form

$$l(v) = \langle v^*, v - v_0 \rangle + J(v_0), \quad l(v) \le J(v), \quad \forall v \in V. = \langle v^*, v \rangle - (\langle v^*, v_0 \rangle - J(v_0)). \tag{1.43}$$

From this relation, we see that if

$$J^*(v^*) < +\infty, \tag{1.44}$$

then the quantity $\langle v^*, v_0 \rangle - J(v_0)$ is also finite, so that such a minorant exists. The element v^* is called a *subgradient* of J at v_0. The set of all subgradients of J at v_0 forms a *subdifferential*, which is usually denoted by $\partial J(v_0)$. It may be empty, may contain one element or infinitely many elements.

An important property of convex functionals follows directly from the fact that they have an exact affine minorant at any point (at which the functional attains a finite value). Assume that J is a convex functional and $v^* \in \partial J(v_0)$. Then there exists an affine minorant such that

$$\langle v^*, v\rangle - \alpha \le J(v), \ \ \forall v \in V,$$

and $\langle v^*, v_0\rangle - \alpha = J(v_0)$. Hence, we obtain

$$J(v) - J(v_0) \ge \langle v^*, v - v_0\rangle. \tag{1.45}$$

The inequality (1.45) represents the basic incremental relation for convex functionals. For proper convex functionals, there exists a simple criterion that enables one verify whether or not an element v^* belongs to the set $\partial J(v)$.

Proposition 1.2.2 *The following two statements are equivalent:*

$$J(v) + J^*(v^*) - \langle v^*, v\rangle = 0, \tag{1.46}$$
$$v^* \in \partial J(v), \tag{1.47}$$
$$v \in \partial J^*(v^*). \tag{1.48}$$

Proof. Assume that $v^* \in \partial J(v)$. In accordance with (1.45), we have

$$J(w) \ge J(v) + \langle v^*, w - v\rangle, \quad \forall w \in V.$$

Hence,

$$\langle v^*, v\rangle - J(v) \ge \langle v^*, w\rangle - J(w), \quad \forall w \in V$$

and, consequently,

$$\langle v^*, v\rangle - J(v) \ge \sup_{w\in V}\{\langle v^*, w\rangle - J(w)\} = J^*(v^*). \tag{1.49}$$

However, by the definition of J^*, we know that for any v and v^*

$$J^*(v^*) \ge \langle v^*, v\rangle - J(v). \tag{1.50}$$

We observe that (1.49) and (1.50) imply (1.46).

Assume that $v \in \partial J^*(v^*)$. Then $J^*(w^*) \geq J^*(v^*) + \langle w^* - v^*, v\rangle$, so that

$$\langle v^*, v\rangle - J^*(v^*) \geq \sup_{w^* \in V^*} \{\langle w^*, v\rangle - J^*(w^*)\} = J^{**}(v).$$

On the other hand,

$$\langle v^*, v\rangle - J^*(v^*) \geq J^{**}(v) = J(v),$$

and we again arrive at (1.46).

Assume that (1.46) holds. By the definition of J^*, we obtain

$$0 = J(v) + J^*(v^*) - \langle v^*, v\rangle \geq J(v) - J(w) - \langle v^*, v - w\rangle,$$

where w is an arbitrary element of V. Thus,

$$J(w) - J(v) \geq \langle v^*, w - v\rangle, \quad \forall w \in V,$$

which means that $J(v) + \langle v^*, v - w\rangle$ is an exact affine minorant of J (at v) and, consequently, (1.47) holds.

The proof of (1.48) is quite similar.

Definition 1.2.2 *Let J and J^* be a pair of conjugate functionals. Then*

$$D_J(v, v^*) := J(v) + J^*(v^*) - \langle v^*, v\rangle$$

is called the compound *functional.*

From Proposition 1.2.2 it follows that D_J is nonnegative and vanishes only if the arguments satisfy (1.47) and (1.48), which are also called the *duality relations* and very often represent the constitutive relations of a physical model. Compound functionals play an important role in the a posteriori error estimation of nonlinear problems. They serve as penalty functionals that penalize errors caused by dissatisfaction of the duality relations. For this reason, we denote the compound functionals by the letter D.

Note that the relation $D_J(v, v^*) \geq 0$ generates inequalities that can be viewed as generalizations of the Young's inequality (cf. (1.10)–(1.16)):

$$\langle v^*, v\rangle \leq J(v) + J^*(v^*). \tag{1.51}$$

In particular, if V and V^* coincide with $\mathbb{R}^d$ and $J(v) = \frac{|v|^\alpha}{\alpha}$, then $J^*(v^*) = \frac{|v^*|^{\alpha'}}{\alpha'}$ and (1.37) implies the estimate

$$v^* \cdot v \le \frac{|v|^\alpha}{\alpha} + \frac{|v^*|^{\alpha'}}{\alpha'}, \qquad \forall\, v, v^* \in \mathbb{R}^d. \tag{1.52}$$

For convex functions we have the Jensen's inequality

Proposition 1.2.3 *Assume that J is a proper convex functional defined on V, v_i, $i = 1, 2, \dots, d$, are given elements of V and $\lambda_i \in \mathbb{R}_+$ meet the condition*

$$\sum_{i=1}^{d} \lambda_i = 1.$$

Then

$$J\left(\sum_{i=1}^{d} \lambda_i v_i\right) \le \sum_{i=1}^{d} \lambda_i J(v_i). \tag{1.53}$$

This inequality also has an integral form. Let $J : \mathbb{R} \to \mathbb{R}$ be a convex function, $\Omega \subset \mathbb{R}^d$, $v : \Omega \to \mathbb{R}$ be a continuous function, and $\lambda : \Omega \to \mathbb{R}$ be an integrable function that satisfies the conditions

$$\int_\Omega \lambda(x)\, dx = 1, \quad \lambda(x) \ge 0 \quad \text{in } \Omega.$$

Then, the inequality reads as follows:

$$J\left(\int_\Omega \lambda(x) v(x)\, dx\right) \le \int_\Omega \lambda(x) J(v(x))\, dx. \tag{1.54}$$

Gâteaux differentiation

Finally, we recall some basic notions related to the differentiation of convex functionals.

Definition 1.2.3 *We say that J has a weak derivative $J'(v_0) \in V^*$ (at the point v_0) in the sense of Gâteaux if*

$$\lim_{\lambda \to +0} \frac{J(v_0 + \lambda w) - J(v_0)}{\lambda} = \langle J'(v_0), w \rangle, \quad \forall w \in V. \tag{1.55}$$

Assume that J is differentiable in the above sense and $v^* \in \partial J(v_0)$. Then for any $v \in V$ we know that

$$J(v) - J(v_0) \geq \langle v^*, v - v_0 \rangle.$$

Set $v = v_0 + \lambda w$, where $\lambda > 0$. Now, we have

$$J(v_0 + \lambda w) - J(v_0) \geq \lambda \langle v^*, w \rangle.$$

Therefore,

$$\langle J'(v_0), w \rangle = \lim_{\lambda \to +0} \frac{J(v_0 + \lambda w) - J(v_0)}{\lambda} \geq \langle v^*, w \rangle,$$

and

$$\langle J'(v_0) - v^*, w \rangle \geq 0$$

for any $w \in V$. This inequality means that, in such a case, the Gâteaux derivative coincides with v^*.

1.2.7 Uniformly convex functionals

Consider a pair of topologically dual spaces Y and Y^* is its topologically dual counterpart. Let $\Upsilon : Y \to \overline{\mathbb{R}}$ be a nonnegative functional such that $\Upsilon(y) = 0$ if and only if $y = \mathcal{O}_Y$ (zero element of Y).

Definition 1.2.4 *A convex functional $J : Y \to \overline{\mathbb{R}}$ is called uniformly convex in $B(\mathcal{O}_Y, \rho)$ if there exists a functional $\Upsilon_\rho \not\equiv 0$ such that for all $y_1, y_2 \in B(\mathcal{O}_Y, \rho)$ the following inequality holds:*

$$J\left(\frac{y_1 + y_2}{2}\right) + \Upsilon_\rho(y_1 - y_2) \leq \frac{1}{2}\left(J(y_1) + J(y_2)\right). \tag{1.56}$$

From (1.56) it is clear that any uniformly convex functional is convex in $B(\mathcal{O}_Y, \rho)$. Now we establish two important inequalities that hold for uniformly convex functionals assuming that Υ_ρ is even, i.e., $\Upsilon(y) = \Upsilon(-y) \quad \forall y \in B(\mathcal{O}_Y, \rho)$.

Proposition 1.2.4 *If $J : Y \to \overline{\mathbb{R}}$ is uniformly convex and Gâteaux differentiable in $B(\mathcal{O}_Y, \rho)$, then for any $y, z \in B(\mathcal{O}_Y, \rho)$ the following relations hold:*

$$J(z) \geq J(y) + \langle J'(y), z - y \rangle + 2\Upsilon_\rho(z - y) \tag{1.57}$$

and

$$\langle J'(z) - J'(y), z - y\rangle \geq 4\Upsilon_\rho(z - y). \tag{1.58}$$

Proof. By convexity and differentiability we have

$$\Upsilon_\rho(z - y) + J\left(\frac{z + y}{2}\right) \leq \frac{1}{2}J(z) + \frac{1}{2}J(y)$$

and

$$J(y) + \left\langle J'(y), \frac{z - y}{2}\right\rangle \leq J\left(\frac{z + y}{2}\right).$$

Hence,

$$\Upsilon_\rho(z - y) \leq J(z) - J(y) - \Big\langle J'(y), z - y\Big\rangle. \tag{1.59}$$

Analogously, we deduce the estimate

$$2\Upsilon_\rho(y - z) \leq J(y) - J(z) + \Big\langle J'(z), z - y\Big\rangle. \tag{1.60}$$

It is easy to see that (1.58) follow from (1.59) and (1.60).

1.3 Generalized formulations of BVP's and existence of solution

"Solution of a (boundary-value) problem
is a notion of indefinite meaning."
H. Poincaré

Take the problem

$$\Delta u + f = 0 \qquad \text{in } \Omega \tag{1.61}$$

$$u = u_0 \qquad \text{on } \partial\Omega \tag{1.62}$$

as the basic example.

How should we understand u?

In the 19th century the problem was understood in the classical sense:
find $u \in C^2(\Omega) \cap C(\bar{\Omega})$ such that (1.62) is satisfied in the pointwise sense and

$$u_{,11} + u_{,22} + f = 0$$

at ALL points of Ω, where $u_{,ss}$ is understood as the classical derivative.

Immediately the question rises: can we always find such u? Unlike for ODE's, this question occurred to be so difficult that the answer was found only about one hundred years of studies that completely reconstructed the whole mathematical building. On this way, a lot of mistakes was made and at the same time the fundament of modern PDE was created by outstanding mathematicians as Weierstrass, Banach, Hilbert, Poincaré, Sobolev, Courant, Ladyzhenskaya and many others.

The concept of generalized solutions to PDE's came from *Petrov-Bubnov-Galerkin method*[2]

The idea was to find $u_N = \sum_{i=1}^{N} \alpha_i w_i$

$$\int_\Omega (\Delta u_N + f) w_i \, dx = 0 \qquad \forall w_i, \quad i = 1, 2, ...N$$

[2]B. G. Galerkin. Beams and plates. Series in some questions of elastic equilibrium of beams and plates (approximate translation of the title from Russian). *Vestnik Ingenerov, St.-Peterburg*, 19(1915), 897-908.

Equivalently, it means that u_N is such that the residual of the differential equation is orthogonal to the finite dimensional space V_N formed by linearly independent w_i.

The key idea of generalized solutionis a logical extension of the Petrov-Galerkin idea, namely:

Generalized solution is a function that makes the residual orthogonal to V:

$$\int_\Omega (\Delta u + f) w \, dx = 0 \qquad \forall w$$

Integration by parts leads to the so–called generalized formulation of the problem: find $u \in \overset{\circ}{H}{}^1(\Omega) + u_0$ such that

$$\int_\Omega \nabla u \cdot \nabla w \, dx = \int_\Omega f w \, dx \qquad \forall w \in \overset{\circ}{H}{}^1(\Omega)$$

This idea admits wide extensions [3] to many differential problems representable in the form: for a certain linear continuous functional f (from the space V^* topologically dual to V) find u such that

$$B(u, w) =< f, w > \qquad w \in V.$$

Here a symmetric form $B : V \times V \to R$, where V is a Hilbert space, is called $V - elliptic$ if $\exists c_1 > 0, c_2 > 0$ such that

$$B(u, u) \geq c_1 \|u\|^2, \quad \forall u \in V$$

$$\mid B(u, v) \mid \leq c_2 \|u\| \|v\|, \quad \forall u, v \in V$$

How to prove that such a statement is correct?

First proofs were based on

[3] see e.g., O. A. Ladyzhenskaya, *The boundary value problems of mathematical physics.* Springer-Verlag, New York, 1985

Theorem 1.3.1 (Lax-Milgram Lemma) *For a bilinear form B there exists a linear bounded operator $A \in \mathcal{L}(\mathcal{V}, \mathcal{V})$ such that*

$$B(u, v) = (Au, v), \ \ \forall u, v \in V$$

It has an inverse $A^{-1} \in \mathcal{L}(\mathcal{V}, \mathcal{V})$, such that $\|A\| \le c_2$, $\|A^{-1}\| \le \frac{1}{c_1}$.

Proof. 1. Take $u \in V$ and consider a linear functional

$$v \to B(u, v) = l(v),$$

where u is a given element in V. We have

$$\mid l(v) \mid = \mid B(u, v) \mid \le c_2 \|u\| \|v\|.$$

Therefore, $l(v) \in V^*$ and there exists $\omega \in V$, such that

$$l(v) = B(u, v) = (\omega, v). \ \ \forall v \in V$$

Set $\omega = Au$. Evidently, $A : V \to V$ is a linear operator and

$$B(u, v) = (Au, u), \ \ \forall u, v \in V$$
$$(Au, u) \le c_2 \|u\| \|v\|.$$

Set $v = Au$, then

$$(Au, Au) \le c_2 \|u\| \|v\| = c_2 \|u\| \|Au\|$$

and, consequently,

$$\|Au\| \le c_2 \|u\|,$$

so that $\|A\| \le c_2$.

2. Now we show that $L = A(u)$ is a subspace in V (i.e., close linear manifold). The fact that L is a lineal follows from the linearity of A.

Note that

$$A(u, u) = B(u, u) \ge c_1 \|u\|^2.$$

On the other hand,

$$(Au, u) \le \|Au\| \|u\|$$

From here, we conclude that

$$\|Au\| \geq c_1\|u\|, \tag{1.63}$$

Next, L contains all limits of converging sequences. i.e., if $\omega^n \to \omega$ in V and $\omega^n \in L$, then $\omega \in L$.

Indeed, $\exists u^n \in V$, such that $Au^n = \omega^n$. By (1.63), we have

$$\|\omega^m - \omega^n\| \geq c_1\|u^m - u^n\|$$

and u^n is a fundamental sequence. Since V is a full space, this sequence converges to $u \in V$. Since A is a continuous operator, $Au^n \to Au$.

On the other hand, $Au^n = \omega^n \to \omega$ as $n \to \infty$. Thus, $\omega = Au$ and, therefore, $\omega \in L$.

3. L is a subspace of V. Assume that $\exists u_0 \in V$ such that $u_0 \notin L$.

By the Banach-Han theorem $\exists \ell(v) \in V^*$ such that

$$\ell(u_0) = 1, \;\; \ell(v) = 0. \;\; \forall v \in L$$

The functional $\ell(u)$ admits the presentation $\ell(v) = (v, \omega_*)$, where $\omega_* \in V$. Hence,

$$(u_0, \omega_*) = 1, \;\; (v, \omega_*) = 0 \;\; \forall v \in L$$

and

$$(u_0, \omega_*) = 1, \;\; (Au, \omega_*) = 0 \;\; \forall u \in V$$

Set $u = \omega_*$, then $(A\omega_*, \omega_*) = 0$. On the other hand,

$$(A\omega_*, \omega_*) \geq c_1\|\omega_*\|^2$$

and $\omega_* \equiv 0$. But then $(u_0, \omega_*) = 1$ is not true.

Hence, $A(V)$ coincides with V and, therefore A is a one-to-one mapping of V to V, which means that A^{-1} exists.

In (1.63) $\|Au\| \geq c_1\|u\|$ we set $v = Au$. Then

$$\|v\| \geq c_1\|A^{-1}v\| \;\; \text{or} \;\; \|A^{-1}v\| \leq \frac{1}{c_1}\|v\| \;\; \forall v \in V$$

and we find that

$$\|A^{-1}\| \leq \frac{1}{c_1}.$$

Existence via LM Lemma

Consider the abstract problem: find $u \in V$, such that

$$B(u, v) = l(v), \quad \forall v \in V \tag{1.64}$$

where $l \in V^*$.

Theorem 1.3.2 *Let B be a V-elliptic bilinear form and $l \in V^*$. Then, (1.64) has a unique solution and*

$$\|u\| \leq \frac{1}{c_1}\|l\|. \tag{1.65}$$

Proof. There exists $w \in V$ such that

$$l(v) = (v, w), \quad \forall v \in V$$

and $\|u\| = \|l\|$. By LM Lemma, $B(u, v) = (Au, v)$ and (1.64) is equivalent to

$$(Au, v) = (w, v), \quad \forall v \in V$$

which is equivalent to $Au = w$ or $u = A^{-1}w$. Hence, u is unique. By LM Lemma, we also conclude that

$$\|u\| = \|A^{-1}w\| \leq \|A^{-1}\| \, \|w\| \leq \frac{1}{c_1}\|w\|.$$

1.3.1 Variational approach to elliptic PDE's

Variational approach arose in the 19th century shortly after the first PDE's have been presented.

It brings the origin from the Fermat theorem:

If f is a differentiable function which attains minimum at $\bar{x}$, then $f'(\bar{x}) = 0$.

Later, when L. Euler created the calculus of variations he extended this principle to 1D variational problems and established that the minimizer of the integral $\int_0^T g(t, y, \dot{y})dt$ is described by an ODE (later named Euler equation).

It is easy to show that this principle can be extended to multidimensional variational problems.

Consider the problem:

Find $u(x)$ such that $u = u_0$ on $\partial\Omega$ and

$$J(u) = \inf_v J(v), \tag{1.66}$$

where infimum is taken over all admissible v (i.e., such that $J(v)$ is finite) satisfying $v = u_0$ on $\partial\Omega$.

What is the relation that must satisfy u if

$$J(v) = \int_\Omega (\frac{1}{2}|\nabla v|^2 - fv)dx \quad ?$$

Let w be an admissible (smooth) function vanishing at the boundary. Then, for any $\lambda > 0$ we have

$$J(u) \le J(u + \lambda w) = \int_\Omega (\frac{1}{2}|\nabla(u + w)|^2 - f(u + w))dx \tag{1.67}$$

Then

$$\int_\Omega (\lambda\nabla u \cdot \nabla w + \frac{\lambda^2}{2}|\nabla w|^2 - \lambda f v)dx \ge 0$$

$$\int_\Omega (\nabla u \cdot \nabla w - \lambda f v)dx \geq \int_\Omega -\frac{\lambda}{2}|\nabla w|^2 dx$$

Since λ is arbitrary (!), we find that (1.67) can hold only if

$$\int_\Omega (\nabla u \cdot \nabla w - f v)dx \geq 0 \quad \forall w$$

Take $-w$ instead of w, then we arrive at the conclusion that

$$\int_\Omega (\nabla u \cdot \nabla w - f v)dx = 0 \quad \forall w. \tag{1.68}$$

In fact, we have derived the generalized formulation of a boundary-value problem using the variational argumentation.

Regrettably, in the 19th century instead of paying attention to (1.68) they continued manipulations in order to obtain "solutions" expressed in terms of the classical derivatives.

Certainly, the classical statement also follows from (1.68) if we use the relations

$$a \cdot \nabla w + w \operatorname{div} a = \operatorname{div}(aw)$$
$$\int_\Omega \operatorname{div}(aw)dx = \int_{\partial\Omega} (a \cdot n) w ds$$

and transform (1.68) as follows

$$\int_\Omega \nabla u \cdot \nabla w \, dx = -\int_\Omega (\operatorname{div} \nabla u) w dx + \int_{\partial\Omega} (\nabla u \cdot n) w ds.$$

Since $w = 0$ on $\partial\Omega$, we arrive at

$$\int_\Omega (\operatorname{div} \nabla u + f) w dx = 0 \quad \forall w \tag{1.69}$$

Now, we use Du-Bois-Reymond Lemmathat says that if g is continuous and

$$\int_\Omega g w dx = 0 \quad \forall \text{ smooth } w \text{ vanishing on } \partial\Omega,$$

then $g = 0$.

Hence, we conclude that if the minimizer is sufficiently regular, then it is a solution of the problem

$$\Delta u + f = 0, \qquad \text{in } \Omega \tag{1.70}$$

$$u = u_0 \qquad \text{on } \partial\Omega. \tag{1.71}$$

In 19th century this approach was believed to give a way of proving that the classical solution (i.e., a function u satisfying (1.70)–(1.70)) indeed exists. This way was strongly advocated by Reimann who offered the following simple "existence proof" for the case $f = 0$ and smooth $\partial\Omega$:

It is clear that for any smooth w satisfying $w = u_0$ on $\partial\Omega$ we have

$$J(w) = \frac{1}{2}\int_\Omega |\nabla w|^2\, dx \geq 0.$$

Hence, values of the (energy) functional are bounded from below and we can construct a sequence of smooth functions $\{w_s\}$ such that

$$J(w_s) \to \quad \text{exact lower bound.}$$

From here, it was concluded that there exists a smooth function that corresponds to the minimal value of the functional.

However, shortly Karl Weierstrass discovered a logical gap in this argumentation: *a sequence of smooth functions may have a nonsmooth limit, for which the equation $\Delta u = 0$ has no sense.*

In spite of that this simple "proof" have failed, it occurred to be very thought-provoking, especially for Weierstrass, who started his fundamental studies of variational problems.

Regrettably, at that time proper understanding of existence problems was hardly possible because one of the main parts of modern mathematics, namely, FUNCTIONAL ANALYSIS, did not exist. Weierstrass was one of those have created its fundament.

Theorem 1.3.3 (Weierstrass 1) *Let K be a closed bounded set $\mathbb{R}^d$ and J be a continuous function defined on K. Then, the problem*

$$\inf_{v\in K} J(v) \tag{1.72}$$

has a solution $u \in K$ such that $J(u)$ gives the infimum.

Proof. Let $\{v_k\}$ be a minimizing sequence, i.e., $J(v_k) \to \inf\limits_K J$. We can extract a converging subsequence out of it (Boltzano-Weierstrass Lemma), which we denote $\{v_{k_s}\}$. Since K is closed we know that the limit of this sequence (we denote it by u) belongs to K. Since J is continuous, we find that

$$\inf_K J = \lim_{s\to\infty} J(v_{k_s}) = J(u).$$

Thus, u is the minimizer.

Regrettably this Theorem cannot be applied to our case. Our problem is as follows

$$\inf_{w\in V_0} \int_\Omega (\frac{1}{2}|\nabla w|^2 - fw)dx$$

where $V_0 = \overset{\circ}{H}{}^1(\Omega)$, i.e.,

$$K = \overset{\circ}{H}{}^1(\Omega).$$

This set is not bounded !

Theorem 1.3.4 (Weierstrass 2) *Let V be a full metric space, $K \subset V$ be a compact set and J be a lower semicontinuous functional*[4] *defined and finite on K. Then, the problem*

$$\inf_{v\in K} J(v) \tag{1.73}$$

has a solution $u \in K$ such that $J(u)$ gives the infimum.

Proof. Let $\{v_k\}$ be a minimizing sequence, i.e., $J(v_k) \to \inf\limits_K J$. Since K is compact, we can extract a converging subsequence out of it, which we denote $\{v_{k_s}\}$. Since K is closed we know that the limit of this sequence (we denote it by u) belongs to K. Since J is lower semicontinuous, we find that

$$\inf_K J = \lim_{s\to\infty} J(v_{k_s}) \geq J(u).$$

[4]We recall that the functional is lower semicontinuous at v_0 if $\lim_{v_s\to v_0} J(v_s) \geq J(v_0)$

Thus, u is the minimizer.

In [Weierstrass 2] we have rather strong conditions for the set K (compactness). Except some special cases, it is impossible to guarantee this property.

Therefore, the idea is to *reduce requirements for K and strengthen for J.*

Namely: we replace *compactness* by *weak compactness.*

This is a very practical change because any closed bounded subset of a Hilbert space is weakly compact!

Theorem 1.3.5 (Weierstrass 3) *Let K be weakly compact and J be a weakly lower semicontinuous functional*[5] *defined on K. Then, the problem*

$$\inf_{v\in K} J(v) \tag{1.74}$$

has a solution $u \in K$ such that $J(u)$ gives the infimum.

Proof. Let $\{v_k\}$ be a minimizing sequence, i.e., $J(v_k) \to \inf\limits_K J$. We can extract a weakly converging subsequence

$$\{v_{k_s}\} \rightharpoonup u \in K.$$

Since J is weakly lower semicontinuous, we find that

$$\inf_K J = \lim_{s\to\infty} J(v_{k_s}) \geq J(u).$$

Thus, u is the minimizer.

Theorem [Weierstrass 3] is more suitable for us because the set

$$K := \{w \in V_0 \mid J(w) \leq J(v_1)\}$$

is bounded. Indeed,

$$\begin{aligned}
&\frac{1}{2}\|\nabla w\|^2 - \int_\Omega f w dx \leq J(v_1),\\
&\|\nabla w\|^2 \leq J(v_1) + \|f\|\|w\| \leq J(v_1) + \|f\| C_F \|\nabla w\|\\
&\qquad \leq (J(v_1) + \frac{1}{2}C_F^2\|f\|^2) + \frac{1}{2}\|\nabla w\|^2.
\end{aligned}$$

[5] i.e., the functional that possesses lower semicontinuity with respect to weakly converging sequences.

and $\|\nabla w\| \leq$ const.

How to verify weak lower semicontinuity ?

Hopefully, there is a simple rule:

Convex lower semicontinuous functional is weakly lower semicontinuous.

Typically, the functionals arising in variational formulations of boundary-value problems are *continuous* and convexity is easy to check.

We recall that J is convex if

$$J(\lambda_1 v_1 + \lambda_2 v_2) \leq \lambda_1 J(v_1) + \lambda_2 J(v_2), \quad \lambda_1 + \lambda_2 = 1,\ \lambda_i \geq 0.$$

Since

$$\begin{aligned}(\lambda_1 v_1 + \lambda_2 v_2)^2 = \lambda_1^2 v_1^2 + 2\lambda_1\lambda_2 v_1 v_2 + \lambda_2^2 v_2^2 \\ \leq\ \lambda_1^2 v_1^2 + \lambda_1\lambda_2 v_1^2 + \lambda_1\lambda_2 v_2^2 + \lambda_2^2 v_2^2 \\ \leq\ \lambda_1 v_1^2 + \lambda_2 v_2^2\end{aligned}$$

we observe that quadratic functionals are convex.

Thus, for our particular problem [Weierstrass 3] is enough to establish existence of a minimizer and, consequently, existence of a solution to PDE.

However, the method is extendable to a much wider class of problems.

Definition 1.3.1 *The functional J is called coercive on $K \subset V$ if*

$$J(v_k) \to +\infty \quad \text{as} \quad \|v_k\|_V \to +\infty \tag{1.75}$$

Coercivity plays an important role in establishing existence results.

Lemma 1.3.1 *Let J is coercive, then the set*

$$V_\alpha := \{v \in V \mid J(v)) \leq \alpha\}$$

is bounded.

Proof. Assume the opposite, i.e., that V_α is unbounded and it is not contained in any ball

$$B(0,d) = \{v \in V \mid \|v\|_V \le d\}.$$

This means that for any integer k, one can find $v_k \in V_\alpha$ such that $\|v_k\|_V > k$. Then, b the coercivity we conclude that

$$J(v_k) \to +\infty \quad \text{as} \quad k \to +\infty.$$

But this is impossible because the elements of V_α are such that the functional does not exceed α.

Theorem 1.3.6 (Weierstrass 4) [6] *Let $J : K \to R$ be convex, continuous and coercive, i.e., and K be a convex closed subset of a Hilbert space V. Then the problem*

$$\inf_{w \in K} J(w)$$

has a minimizer u. If J is strictly convex, then the minimizer is unique.

Proof. Let $\{v_k\}$ be a minimizing sequence, i.e., $J(v_k) \to \inf_K J$. The set

$$K_1 := \{v \in K \mid J(v) \le J(v_1)\}$$

is bounded (see Lemma). Evidently, it is also closed. In a Hilbert space all closed bounded sets are weakly compact. Therefore, we can extract a weakly converging subsequence

$$\{v_{k_s}\} \rightharpoonup u \in K_1.$$

Since J is convex and continuous, it is weakly lower semicontinuous, we find that

$$\inf_K J = \lim_{s \to \infty} J(v_{k_s}) \ge J(u).$$

Hence u is the minimizer.

[6] see, e.g., I. Ekeland and R. Temam. *Convex analysis and variational problems.* North-Holland, Amsterdam, 1976.

Assume that J is strictly convex, i.e.,

$$J(\lambda_1 v_1 + \lambda_2 v_2) < \lambda_1 J(v_1) + \lambda_2 J(v_2), \quad \lambda_1 + \lambda_2 = 1,\ \lambda_i > 0.$$

If u_1 and u_2 are two different minimizers, then we immediately arrive at a contradiction because

$$J(\lambda_1 u_1 + \lambda_2 u_2) < \lambda_1 J(u_1) + \lambda_2 J(u_2) = \inf_K J.$$

Example 1.
Take $J(w) = \frac{1}{2}B(w,w) - < f, w >$ and let $K = V$.

Then

$$\frac{1}{2}B(w,w) \geq c_1 \|w\|_V^2, \qquad | < f, w > | \ \leq \ \|f\|_{V^*} \|w\|_V.$$

We see, that

$$J(w) \geq \ c_1 \|w\|_V^2 - \|f\|_{V^*} \|w\|_V \to +\infty \quad \text{as } \|w\|_V \to +\infty$$

Since J is strictly convex and continuous we conclude that a minimizer exists and unique.

Example 2.
Take $J(w) = \int_\Omega (\frac{1}{p}|\nabla w|^p - fw)dx$
and let
$K = W^{1,p}(\Omega)$,
where $p > 1$.

This functional is convex and continuous on $W^{1,p}$. Its coercivity is also obvious (indeed $a|x|^p - bx$ tends to infinity if $x \to +\infty$).

This variational problem is related to a nonlinear PDE called $p - Laplacian$.

Example 3 Consider the problem
$J(w) = \int_\Omega (\frac{\nu}{2}|\nabla w|^2 + k_*|\nabla w| - fw)dx$
and let
$K = \overset{\circ}{H}{}^1(\Omega)$.

This nonlinear model is related to the so-called *Bingham fluid* (μ, k_* are positive constants defined by viscosity and plasticity properties of the fluid).

Task 1.3.1 *Using Theorem [Weierstrass 4] prove that this variational problem has a unique solution.*

Example 4. Consider the problem
$J(w) = \int_\Omega (\frac{1}{2}|\nabla w|^2 - cw)dx$
and let
$K = \{w \in \overset{\circ}{H}{}^1(\Omega) \mid |\nabla w| \leq k_*\}$.

This nonlinear model is related to the so-called *elasto-plastic torsion* of a long bar.

Task 1.3.2 *Using Theorem [Weierstrass 4] prove that this variational problem has a unique solution.*

Comment. Limits of applicability: Violations of the conditions in [Weierstrass 4] that arise in practical problems are due to:

- 1. Nonconvexity of the functional J.
- 2. Nonconvexity of K.
- 3. Nonreflexivity of V.

1. Nonconvex problems.Phase transitions in solids:

$$\int_\Omega (g(\nabla w) - fw)dx, \quad g = \min\{g_1, g_2\}$$

Here g_1 and g_2 are two energy functionals related to two different phases. In these problems a minimizing sequence may (strongly) converge to nothing and "solutions" are presented by structures with rapidly oscillating layers.

Lower semicontinuous regularization is required, which amounts constructing CONVEX or QUASICONVEX envelope of g.

2. Optimal control problems with control in the main operator part.

$$\inf_K J(\eta, w)$$
$$K := \{(\eta, w) \mid A(\eta)w + f = 0\}$$

The set K may be nonconvex and, therefore, may be not weakly compact. Nonexistence arises in the form of the so called "sliding regimes".

Mathematically, the so-called G-closure of the operator set is required.

3. Problems with linear growth.

Typical problem is the *nonparametric minimal surface problem*

$$J(v) = \int_\Omega \sqrt{1 + |\nabla w|^2} dx.$$

The functional J is defined on the Sobolev space $V := W^{1,1}$, so that we set

$$K := \{w \in V \mid w = u_0 \text{ on } \partial\Omega\}.$$

This functional is *convex* and continuous on V. Since $J(w) \geq \|\nabla w\|$ it is *coercive* on K. Also, K is convex and closed (with respect to convergence in V).

However, the variational problem may have no solution because $W^{1,1}$ is a *nonreflexive space*. For such spaces, we cannot say that CONVEXITY+BOUNDEDNESS implies WEAK COMPACTNESS.

Practically important classes of engineering problems related to such problems are: Capillary surface and perfect plasticity problems.

> Here minimizing sequence may converge to a discontinuous function. Therefore, special approximation methods are required.

1.3.2 PDE analysis via minimax theory

Henceforth, we will consider the problem

$$\begin{aligned} &\operatorname{div} A\nabla u + f = 0 \quad \text{in } \Omega\,, \\ &u = u_0 \qquad \text{on } \partial_1\Omega\,, \\ &A\nabla u \cdot n = F \text{on } \partial_2\Omega\,, \end{aligned}$$

$$c_1^2|\xi|^2 \leq A(x)\xi \cdot \xi \leq c_2^2|\xi|^2 \qquad \forall \xi \in \mathbb{R}^d\,, \text{ for a.e. } x \in \Omega\,,$$

where

$$u_0 \in H^1(\Omega)\,,\ f \in L_2(\Omega)\,,\ F \in L_2(\partial_2\Omega)\,.$$

Notation.

$$
\begin{aligned}
&V := H^1(\Omega) \qquad \text{basic space},\\
&V_0 := \{v \in V \ \mid \ v = 0 \ \text{ on } \partial_1\Omega\},\\
&V_0 + u_0 := \{v \in V \ \mid \ v = w + u_0, \ w \in V_0\} \quad \text{energy space}\\
&\widehat{V} := L_2(\Omega) \qquad \text{extended (nonconforming) energy space},\\
&Q := L_2(\Omega; \mathbb{R}^d) \qquad \text{extended space for fluxes}\\
&\widehat{Q} := H(\Omega; \operatorname{div}) \qquad \text{space for fluxes}\\
&\widehat{Q}_{\partial_2\Omega} := \{y \in \widehat{Q} \ \mid \ y \cdot n\big|_{\partial_2\Omega} \in L_2(\partial_2\Omega)\} \quad \text{reduced space for fluxes.}
\end{aligned}
$$

We recall that $\|q\|_{\operatorname{div}}$ is the norm in $H(\Omega; \operatorname{div})$:

$$
\|q\|_{\operatorname{div}} := (\|q\|^2 + \|\operatorname{div} q\|^2)^{1/2} \quad \forall q \in Q
$$

and

$$
\begin{aligned}
&|\!|\!| \ q \ |\!|\!| := \left(\int\limits_\Omega Aq \cdot q \, dx \right)^{1/2}, \quad q \in Q\\
&|\!|\!| \ q \ |\!|\!|_* := \left(\int\limits_\Omega A^{-1}q \cdot q \, dx \right)^{1/2}
\end{aligned}
$$

Note that,

$$
\bar{c}_1^2|\xi|^2 \le A^{-1}(x)\xi \cdot \xi \le \bar{c}_2^2|\xi|^2 \qquad \forall \xi \in \mathbb{R}^d, \text{ for a.e. } x \in \Omega
$$

with $\bar{c}_1 = 1/c_2$, $\bar{c}_2 = 1/c_1$.

In the so-called *mixed* formulations of PDE's the solution can is viewed as a pair of functions that give a saddle point to the Lagrangian

$$
L(v, q) := \int_\Omega \left(\nabla v \cdot q - \frac{1}{2} A^{-1}q \cdot q \right) dx - \boldsymbol{\ell}(v) \, ,
$$

where $\boldsymbol{\ell}(v) = \int_\Omega f v \, dx + \int_{\partial_2\Omega} F v \, ds$.

The problem of finding $(u,p) \in V_0 + u_0 \times Q$ such that

$$L(u,q) \;\le\; L(u,p) \;\le\; L(v,p) \quad \forall q \in Q\,,\ \forall v \in V_0 + u_0 \tag{1.76}$$

leads to is the so-called Primal Mixed Formulation

Which relations follow from (1.76)?

Take the left-hand inequality

$$L(u, p + \lambda\eta) \le L(u,p), \quad \forall \eta \in Q,\ \lambda > 0$$

Then

$$\int_\Omega (\lambda \nabla u \cdot \eta - \frac{\lambda^2}{2} A^{-1}\eta \cdot \eta - \lambda A^{-1} q \cdot \eta) dx \le 0;$$

$$\int_\Omega (\nabla u \cdot \eta - A^{-1} q \cdot \eta) dx \le \int_\Omega \frac{\lambda}{2} A^{-1}\eta \cdot \eta;$$

$$\int_\Omega (\nabla u \cdot \eta - A^{-1} q \cdot \eta) dx = 0 \quad \forall \eta \in Q$$

Take the right-hand inequality

$$L(u,p) \le L(u + w, p), \quad \forall w \in V_0.$$

Then

$$\int_\Omega (\nabla w \cdot p - fw) dx - \int_{\partial_2\Omega} F w ds \ge 0 \quad \forall w \in V_0.$$

If $p \in \hat{\hat{Q}}_{\partial_2\Omega}$ then we conclude that

$$\operatorname{div} p + f = 0 \quad \text{and} \quad p \cdot n = F \text{ on } \partial_2\Omega.$$

Thus, we see that the saddle-point $(u,p) \in (V_0 + u_0) \times Q$ satisfies the relations

$$\int_\Omega \left(A^{-1}p - \nabla u\right) \cdot q \, dx = 0 \qquad \forall q \in Q\,, \tag{1.77}$$

$$\int_\Omega p \cdot \nabla w \, dx - \boldsymbol{\ell}(w) = 0 \qquad \forall w \in V_0\,. \tag{1.78}$$

In the Primal Mixed Formulation the *constitutive relation*

$$p = A\nabla u$$

is satisfied in $L_2(\Omega)$- sense and the *conservation law*

$$\operatorname{div} p + f = 0$$

and the boundary condition $p\cdot n = F$ on $\partial_2\Omega$ are satisfied in a weak sense.

An introduction to the theory of saddle-points.

Consider the abstract saddle-point problem. Let $K \subset V$ and $M \subset Q$ Find $(u,p) \in K \times M$ such that

$$L(u,q) \;\le\; L(u,p) \;\le\; L(v,p) \quad \forall q \in M\,,\ \forall v \in K \tag{1.79}$$

Which conditions could guarantee that the minimax problem is stated correctly and the saddle point exists?

First, we assume that

- V and Q are *reflexive* Banach spaces (e.g., Hilbert spaces), K and M be *convex and closed subsets* of V and M, respectively.
- The functional $v \mapsto L(v,q)$ be *convex and lower semicontinuous* for any $q \in M$.
- The functional $q \mapsto L(v,q)$ be *concave and upper semicontinuous* for any $v \in K$.

However, these conditions are not sufficient to guarantee that a saddle-point exists!

Note that L generates two functionals:

$$J(v) := \sup_{q\in M} L(v,q) \tag{1.80}$$

and

$$I^*(q) := \inf_{v\in K} L(v,q). \tag{1.81}$$

The functionals J and I^* generate two variational problems.

Problem $\mathcal{P}$. Find $u \in K$ such that

$$J(u) = \inf \mathcal{P} := \inf_{v \in K} J(v). \tag{1.82}$$

Problem $\mathcal{P}^*$. Find $p \in M$ such that

$$I^*(p) = \sup \mathcal{P}^* := \sup_{q \in M} I^*(q). \tag{1.83}$$

Henceforth, Problems $\mathcal{P}$ and $\mathcal{P}^*$ are called *primal* and *dual*, respectively. They are closely related to the *saddle-point* problem.

How are these two problems related?

First, we establish one relation that holds regardless of the structure of the Lagrangian.

Sup Inf and Inf Sup

Lemma 1.3.2 *Let $L(x, y)$ be a functional defined on the elements of two nonempty sets X and Y. Then*

$$\sup_{y \in Y} \inf_{x \in X} L(x, y) \le \inf_{x \in X} \sup_{y \in Y} L(x, y). \tag{1.84}$$

Proof. It is easy to see that

$$L(x, y) \ge \inf_{\xi \in X} L(\xi, y), \quad \forall x \in X,\ y \in Y.$$

Taking the supremum over $y \in Y$, we obtain

$$\sup_{y \in Y} L(x, y) \ge \sup_{y \in Y} \inf_{\xi \in X} L(\xi, y), \quad \forall x \in X.$$

The left-hand side depends on x, while the right-hand side is a number. Thus, we may take infimum over $x \in X$ and obtain the inequality

$$\inf_{x \in X} \sup_{y \in Y} L(x, y) \ge \sup_{y \in Y} \inf_{\xi \in X} L(\xi, y).$$

Therefore, we always have

$$\sup \mathcal{P}^* \leq \inf \mathcal{P} \tag{1.85}$$

Let us prove that if a saddle-point exists, then its components are solutions of Problems $\mathcal{P}$ and $\mathcal{P}^*$ and (1.85) holds as equality.

First, we present a simple saddle-point criterion.

Proposition 1.3.1 *If there exist a constant α such that*

$$L(u,q) \leq \alpha, \ \forall q \in M, \tag{1.86}$$

and

$$L(v,p) \geq \alpha, \ \forall v \in K, \tag{1.87}$$

then (u,p) is a saddle point. Moreover, we have the relation

$$\alpha = \inf_{v\in K} \sup_{q\in M} L(v,q) = \sup_{q\in M} \inf_{v\in K} L(v,q). \tag{1.88}$$

From (1.86) and (1.87), we obtain

$$L(u,p) \leq \alpha \leq L(u,p).$$

Therefore, $L(u,p) = \alpha$ and

$$L(u,q) \leq L(u,p) \leq L(v,p), \ \forall v \in K, \ \forall y \in M,$$

which means that (u,p) is a saddle point. Since

$$\sup_{q\in M} L(u,q) = L(u,p) = \alpha$$

and

$$\inf_{v\in K} L(v,p) = L(u,p) = \alpha,$$

we have

$$\inf \mathcal{P} = \inf_{v\in K} \sup_{q\in M} L(v,q) \le \sup_{q\in M} L(u,q) = \alpha,$$
$$\sup \mathcal{P}^* = \sup_{q\in M} \inf_{v\in K} L(v,q) \ge \inf_{v\in K} L(v,p) = \alpha.$$

In view of Lemma, we arrive at (1.88) .

Proposition 1.3.2 *The set of all saddle points of the Lagrangian L has the form $K_0 \times M_0$, where K_0 and M_0 are convex subsets of K and M, respectively.*

Proof. Let (u_1, p_1) and (u_2, p_2) be two different saddle points. Then

$$L(u_1,q) \le L(u_1,p_1) = \alpha = L(u_2,p_2) \le L(v,p_1),$$
$$L(u_2,q) \le L(u_1,p_1) = \alpha = L(u_2,p_2) \le L(v,p_2),$$

where v and q are arbitrary elements of the sets K and M, respectively. From the first relation, we obtain

$$L(u_2,p_1) \ge \alpha,$$

and from the second one we have

$$L(u_2,p_1) \le \alpha.$$

Now, Proposition 1.3.1 implies that (u_2, p_1) is a saddle point. The same conclusion is obviously true for (u_1, p_2). Let u_1 and u_2 be two different elements of K_0. Then,

$$L(u_1,q) \le L(u_1,p_1) = \alpha,\ \forall q \in M,$$
$$L(u_2,q) \le L(u_2,p_1) = \alpha,\ \forall q \in M.$$

We assumed that $v \mapsto L(v,q)$ is convex. Therefore, for any positive $\lambda_1 + \lambda_2 = 1$ we have

$$L\left(\lambda_1 u_1 + \lambda_2 u_2, q\right) \le \lambda_1 L(u_1,q) + \lambda_2 L(u_2,q) \le \alpha.$$

Since $L(v,p_1) \ge \alpha$, $\forall v \in K$, we deduce the opposite inequality

$$L\left(\lambda_1 u_1 + \lambda_2 u_2, p_1\right) \ge \alpha.$$

It remains to conclude that $L\left(\lambda_1 u_1 + \lambda_2 u_2, p_1\right) = \alpha$. Hence, $\lambda_1 u_1 + \lambda_2 u_2 \in K_0$.

Now we state the main theorem that establishes a link between the solutions of Problems $\mathcal{P}$ and $\mathcal{P}^*$ and the saddle points of Problems $\mathcal{L}$.

Theorem 1.3.7 *The following two statements are equivalent:*

1. *there exists a pair of elements $u \in K$ and $p \in M$ such that*

$$J(u) = \inf \mathcal{P}, \tag{1.89}$$
$$I^*(p) = \sup \mathcal{P}^*, \tag{1.90}$$
$$\inf \mathcal{P} = \sup \mathcal{P}^* \tag{1.91}$$

2. *(u, p) is a saddle point of the Lagrangian L on $K \times M$.*

Moreover, any of the above two assertions implies the principal relation

$$I^*(p) = L(u, p) = J(u). \tag{1.92}$$

Proof. Let the first assumption be true. Then

$$L(u, q) \le \sup_{q \in M} L(u, q) = J(u) = \alpha, \ \forall q \in M,$$
$$L(v, p) \ge \inf_{v \in K} L(v, p) = I^*(p) = \alpha, \ \forall v \in K.$$

According to Proposition 1.3.1, (u, p) is a saddle point. Let (u^*, p) be a saddle point, i.e.,

$$L(u, q) \le L(u, p) \le L(v, p), \ \forall v \in K, q \in M.$$

From this double inequality we obtain

$$J(u) = \sup_{q \in M} L(u, q) \le L(u, p) \le$$
$$\le L(v, p) \le \sup_{q \in M} L(v, q) = J(v), \ \forall v \in K,$$

and

$$I^*(p) = \inf_{v \in K} L(v, p) \ge L(u, p) \ge$$
$$\ge L(u, q) \ge \inf_{v \in K} L(v, q) = I^*(q), \ \forall q \in M.$$

Hence, $u \in K$ and $p \in M$ are solutions. Furthermore,

$$L(u,p) \le \sup_{q \in M} L(u,q) = J(u) \le L(u,p),$$
$$L(u,p) \ge \inf_{v \in K} L(v,p) = I^*(p) \ge L(u,p),$$

and the relation (1.92) follows.

Before closing this concise review of saddle-point theory, we shall present two assertions that may be useful in checking the correctness of particular saddle-point problems.

Theorem 1.3.8 *If the assumptions imposed on L hold and the sets K and M are bounded, then L possesses at least one saddle point and*

$$\inf \mathcal{P} = \sup \mathcal{P}^*.$$

Theorem 1.3.9 *If the assumptions imposed on L hold and the sets K hold and there exist elements $p_0 \in M$ and $u_0 \in K$ such that*

$$L(v_k, p_0) \to +\infty \qquad \text{for any sequence } \{v_k\} \in K \text{ such that } \|v_k\|_V \to +\infty,$$
$$L(u_0, q_k) \to -\infty \qquad \text{for any sequence } \{q_k\} \in M \text{ such that } \|q_k\|_{Y^*} \to +\infty\,.$$

Then, L has at least one saddle point.

Combining the conditions of the above two theorems, one can prove, for example, that a saddle point exists if K is bounded and the coercivity condition for q holds (or M is bounded and coercivity condition for v holds).

It is also worth noting that the basic relation

$$\inf \mathcal{P} = \sup \mathcal{P}^*$$

is true even if only one of the coercivity conditions hold. Proofs of all these results and a more detailed exposition of saddle-point theory can be found in Ekeland and Temam.

Now we apply the general theory to diffusion problem Take the Lagrangian

$$L(v,q) = \int_\Omega (\nabla v \cdot q - A^{-1}q \cdot q - fv)dx - \int\limits_{\partial_2\Omega} Fvds$$

It generates two functionals.

$$J(v) := \sup_{q\in Q} L(v,q) = \frac{1}{2} \parallel\!\!\parallel \nabla v \parallel\!\!\parallel^2 - \boldsymbol{\ell}(v)$$

leads to minimization problem

$$\inf_{w\in V_0+u_0} J(w),$$

which has a unique solution u (apply Theorem [Weierstrass 4]).

Consider another problem:

$$\sup_{v\in V_0+u_0} \int_\Omega (\nabla v \cdot q - A^{-1}q \cdot q - fv)dx - \int\limits_{\partial_2\Omega} Fvds$$

We can represent $v = u_0 + w$, where $w \in V_0$. Obviously this supremum is finite only for

$$q \in Q_\ell := \{q \in Q \mid \int_\Omega (\nabla w \cdot q - fw)dx - \int\limits_{\partial_2\Omega} Fwds = 0 \qquad \forall w \in V_0\}$$

and for such q we have

$$I^*(q) := -\frac{1}{2} \parallel\!\!\parallel q \parallel\!\!\parallel_*^2 - \boldsymbol{\ell}(u_0) + \int\limits_\Omega \nabla u_0 \cdot q \, dx \, .$$

Hence, we arrive at the *dual* problem:

$$\sup_{q\in Q_\ell} \left(-\frac{1}{2} \parallel\!\!\parallel q \parallel\!\!\parallel_*^2 - \boldsymbol{\ell}(u_0) + \int\limits_\Omega \nabla u_0 \cdot q \, dx \right)$$

Note that the functional $-I^*$ is convex and continuous on Q. Moreover, it is coercive. The set Q_ℓ is an affine manifold, so that it is convex. It is easy to see that it is closed with respect to convergence in Q.

By [Weierstrass 4] we establish existence of a maximizer p.

It remains to show that

$$J(u) = I^*(p).$$

We know that $J(u) \geq I^*(p)$. Take $\bar{q} := A\nabla u$. Note that

$$\int_\Omega A\nabla u \cdot \nabla w dx = \ell(w) \qquad \forall w \in V_0.$$

Thus $\bar{q} \in Q_\ell$ and

$$\int_\Omega \nabla u_0 \cdot \bar{q} dx - \ell(u_0) = \int_\Omega \nabla u \cdot \bar{q} - \ell(u) +$$
$$\int_\Omega \nabla(u_0 - u) \cdot \bar{q} - \ell(u_0 - u) = \int_\Omega \nabla u \cdot A\nabla u - \ell(u).$$

Also

$$-\frac{1}{2} \parallel \bar{q} \parallel_*^2 = -\frac{1}{2}\int_\Omega A^{-1}A\nabla u \cdot \nabla u \, dx = -\frac{1}{2}\int_\Omega A\nabla u \cdot \nabla u$$

and we find that

$$I^*(\bar{q}) = \frac{1}{2} \parallel \nabla u \parallel - \ell(u) = J(u).$$

Hence the saddle point formulation is correct!

Such a pair (u, p) exists and satisfies the relations

$$\inf_{v \in V_0 + u_0} J(v) := \inf \mathcal{P} = L(u, p) = \sup \mathcal{P}^* := \sup_{q \in Q_\ell} I^*(q)\,, \tag{1.93}$$

> We have found one more formulation of our boundary-value problem which is mathematically correct. It can be used to find approximate solutions by algorithms developed for saddle-point problems.

However, there exists ANOTHER saddle-point formulation of the same problem.

Dual Mixed Method (DMM)

Another mixed formulation arises if we represent L in a somewhat different form. First, we introduce the functional $g : (V_0 + u_0) \times \widehat{Q} \to \mathbb{R}$ by the relation

$$g(v,q) := \int_\Omega (\nabla v \cdot q + v(\operatorname{div} q))\, dx\,.$$

We have

$$L(v,q) = \int_\Omega \left(\nabla v \cdot q - \frac{1}{2} A^{-1} q \cdot q \right) dx - \boldsymbol{\ell}(v) =$$

$$= g(v,q) - \int_\Omega v(\operatorname{div} q)\, dx - \frac{1}{2} \| q \|_*^2 - \boldsymbol{\ell}(v)\,.$$

Introduce the set

$$\widehat{Q}_F := \{ q \in \widehat{Q} \mid g(w,q) = \int_{\partial_2\Omega} Fw\, ds \ \ \forall w \in V_0 \}\,.$$

Note that for $q \in \widehat{Q}_F$ we have

$$g(v,q) = g(w + u_0, q) = g(w,q) + g(u_0,q) =$$

$$= \int_{\partial_2\Omega} Fw\, ds + g(u_0,q) \ \forall w \in V_0\,.$$

Therefore, if the variable q is taken not from Q but from the narrower set $\widehat{Q}_F$, then the Lagrangian can be written as

$$\widehat{L}(v,q) = g(v,q) - \int_\Omega (v(\operatorname{div} q) - fv) dx - \frac{1}{2} \| q \|_*^2 - \int_{\partial_2\Omega} Fv\, ds =$$

$$- \frac{1}{2} \| q \|_*^2 - \int_\Omega v(\operatorname{div} q)\, dx - \int_\Omega fv\, dx - \int_{\partial_2\Omega} Fu_0\, ds + g(u_0,q)\,.$$

We observe that the new Lagrangian $\widehat{L}$
is defined on a wider set of primal functions $v \in \widehat{V}$, but uses a narrower set $\widehat{Q}_F$ for the fluxes.

The problem of finding $(\widehat{u}, \widehat{p}) \in \widehat{V} \times \widehat{Q}_F$ such that

$$\widehat{L}(\widehat{u}, \widehat{q}) \;\leq\; \widehat{L}(\widehat{u}, \widehat{p}) \;\leq\; \widehat{L}(\widehat{v}, \widehat{p}) \quad \forall \widehat{q} \in \widehat{Q}_F \,,\; \forall \widehat{v} \in \widehat{V} \tag{1.94}$$

lead to is the so-called Dual Mixed Method[7]

From (1.94) we obtain the necessary conditions for the dual mixed formulation.Since

$$\widehat{L}(\widehat{u}, \widehat{q}) \leq \widehat{L}(\widehat{u}, \widehat{p}) \quad \forall \widehat{q} \in \widehat{Q}_F,$$

we have

$$-\frac{1}{2} \parallel\!\!| \, \widehat{p} + \lambda\eta \, |\!\!\|_*^2 - \int_\Omega \widehat{u}(\operatorname{div}(\widehat{p} + \lambda\eta) - f\widehat{u}) dx - \int_{\partial_2\Omega} F u_0 \, ds + g(u_0, \widehat{p} + \lambda\eta) \leq$$

$$-\frac{1}{2} \parallel\!\!| \, \widehat{p} \, |\!\!\|_*^2 - \int_\Omega \widehat{u}(\operatorname{div} \widehat{p}) \, dx - \int_\Omega f\widehat{u} \, dx - \int_{\partial_2\Omega} F u_0 \, ds + g(u_0, \widehat{p}),$$

where λ is a real number and η is a function in $\widehat{Q}_0 := \widehat{Q}_F$ with $F = 0$. Now, arrive at the relation

$$-\lambda \int_\Omega (A^{-1}\widehat{p} \cdot \eta + \widehat{u}(\operatorname{div} \eta)) dx + \lambda g(u_0, \eta) \leq \frac{\lambda^2}{2} \int_\Omega A^{-1}\eta \cdot \eta dx.$$

Rewrite it as

$$\int_\Omega (A^{-1}\widehat{p} \cdot \eta + \widehat{u}(\operatorname{div} \eta)) dx - g(u_0, \eta) \geq \frac{\lambda}{2} \int_\Omega A^{-1}\eta \cdot \eta dx.$$

Since $\lambda > 0$ can be taken arbitrarily small, the latter relation may hold only if

$$\int_\Omega (A^{-1}\widehat{p} \cdot \eta + \widehat{u} \operatorname{div} \eta) dx - g(u_0, \eta) \geq 0.$$

But η is an arbitrary element of a linear manifold $\widehat{Q}_0$, so that $+\eta$ can be replaced by $-\eta$ what leads to the conclusion that

$$\int_\Omega (A^{-1}\widehat{p} \cdot \eta + \widehat{u} \operatorname{div} \eta) dx - g(u_0, \eta) = 0 \quad \forall \eta \in \widehat{Q}_0.$$

[7] see, e.g., F. Brezzi and M. Fortin

From

$$\widehat{L}(\widehat{u},\widehat{p}) \;\leq\; \widehat{L}(\widehat{u}+\widehat{v},\widehat{p}) \quad \forall \widehat{v} \in \widehat{V} := L^2(\Omega)$$

we observe that the terms of $\widehat{L}$ linear with respect to the "pressure" must vanish. Namely, we obtain

$$\int_\Omega (\widehat{v}\mathrm{div}\,\widehat{p} + f\widehat{v})dx = 0$$

Thus, we arrive at the system

$$\int_\Omega \left(A^{-1}\widehat{p}\cdot\widehat{q} + (\mathrm{div}\,\widehat{q})\widehat{u}\right)\, dx = g(u_0,\widehat{q}) \quad \forall \widehat{q} \in \widehat{Q}_0\,, \tag{1.95}$$

$$\int_\Omega (\mathrm{div}\,\widehat{p} + f)\widehat{v}\, dx = 0 \qquad \forall \widehat{v} \in \widehat{V}\,. \tag{1.96}$$

We observe that now the condition

$$\mathrm{div}\,\widehat{p} + f = 0$$

is satisfied in a "strong" (L_2) sense, the Neumann type boundary condition is viewed as the essential boundary condition, and the relation

$$\widehat{p} = A\nabla\widehat{u}$$

and the Dirichlet type boundary condition are satisfied in a weak sense.

These properties of the DMM lead to that the respective finite dimensional formulations are better adapted to the satisfaction of the equilibrium type relations for the fluxes. This fact is important in many applications where a sharp satisfaction of the equilibrium relations is required.

The Lagrangian $\widehat{L}$ also generates two functionals

$$\widehat{J}(\widehat{v}) := \sup_{\widehat{q}\in\widehat{Q}_F} \widehat{L}(\widehat{v},\widehat{q}) \quad \text{and} \quad \widehat{I}^*(\widehat{q}) := \inf_{\widehat{v}\in\widehat{V}} \widehat{L}(\widehat{v},\widehat{q})\,.$$

The two corresponding variational problems are

$$\inf_{\widehat{v}\in\widehat{V}} \widehat{J}(\widehat{v}) \quad \text{and} \quad \sup_{\widehat{q}\in\widehat{Q}_F} \widehat{I}^*(\widehat{q}).$$

Theyare called Problems $\widehat{\mathcal{P}}$ and $\widehat{\mathcal{P}}^*$, respectively. Note that the functional $\widehat{J}$ (unlike J) has no simple explicit form. However, we can prove the solvability of Problem $\widehat{\mathcal{P}}$ by the following Lemma.

Lemma 1.3.3 *For any $\widehat{v} \in \widehat{V}$ and $F \in L_2(\partial_2\Omega)$ there exists $p^v \in \widehat{Q}_F$ such that*

$$\operatorname{div} p^v + \widehat{v} = 0 \ \text{ in } \Omega\,, \tag{1.97}$$

$$\| p^v \|_* \leq C_\Omega \left(\|\widehat{v}\| + \|F\|_{\partial_2\Omega}\right) . \tag{1.98}$$

Proof. We know that the boundary-value problem

$$\begin{aligned} \operatorname{div} A\nabla u^v + \widehat{v} = 0 \quad & \text{in } \Omega\,, \\ u^v = 0 \quad & \text{on } \partial_1\Omega\,, \\ A\nabla u^v \cdot n = F \quad & \text{on } \partial_2\Omega \end{aligned}$$

possesses the unique solution $u^v \in V_0$.

For this problem the energy estimate

$$\| \nabla u^v \| \leq C_\Omega \left(\|\widehat{v}\| + \|F\|_{\partial_2\Omega}\right)$$

holds. Let $p^v := A\nabla u^v$. We have

$$\operatorname{div} p^v + \widehat{v} = 0.$$

Obviously, $p^v \in \widehat{Q}_F$ and, since

$$\| p^v \|_*^2 = \int_\Omega A^{-1}(A\nabla u^v) \cdot (A\nabla u^v)\, dx = \| \nabla u^v \|^2\,,$$

we find that (1.98) also holds.

□

By the Lemma we can easily prove the coercivity of $\widehat{J}$ on $\widehat{V}$. Indeed,

$$\widehat{J}(\widehat{v}) \geq \widehat{L}(\widehat{v}, \alpha p^v) =$$
$$-\frac{1}{2} \| \alpha p^v \|_*^2 - \alpha \int_\Omega \widehat{v}(\operatorname{div} p^v)\, dx - \int_\Omega f\widehat{v}\, dx - \int_{\partial_2\Omega} Fu_0\, ds + g(u_0, \alpha p^v) =$$
$$= -\frac{1}{2}\alpha^2 \| p^v \|_*^2 + \alpha\|\widehat{v}\|^2 - \|f\|\|\widehat{v}\| + g(u_0, \alpha p^v) - \int_{\partial_2\Omega} Fu_0\, ds\,.$$

Here $|g(u_0, \alpha p^v)| \leq \alpha \|p^v\|_{\operatorname{div}} \|u_0\|_{1,2,\Omega}$ and

$$\|p^v\|_{\operatorname{div}}^2 = \|p^v\|^2 + \|\operatorname{div} p^v\|^2 \leq \frac{1}{\bar{c}_1} \| p^v \|_*^2 + \|\widehat{v}\|^2 \leq$$
$$\leq \frac{1}{\bar{c}_1} C_\Omega^2 \left(\|\widehat{v}\| + \|F\|_{\partial_2\Omega}\right)^2 + \|\widehat{v}\|^2\,.$$

Therefore

$$\widehat{J}(\widehat{v}) \geq -\frac{1}{2}\alpha^2 C_\Omega^2 \|\widehat{v}\|^2 + \alpha\|\widehat{v}\|^2 + \Theta(\|\widehat{v}\|) + \Theta_0\,,$$

where $\Theta(\|\widehat{v}\|)$ contains the terms linear with respect to $\|\widehat{v}\|$ and Θ_0 does not depend on $\widehat{v}$. Take $\alpha = 1/C_\Omega^2$. Then

$$\widehat{J}(\widehat{v}) \geq \frac{1}{2C_\Omega^2}\|\widehat{v}\|^2 + \Theta(\|\widehat{v}\|) + \Theta_0 \longrightarrow +\infty \text{ as } \|\widehat{v}\| \to \infty\,.$$

It is not difficult to prove that the functional $\widehat{J}$ is convex and lower semicontinuous. Therefore, Problem $\widehat{\mathcal{P}}$ has a solution $\widehat{u}$.

Inf-Sup condition for the dual mixed formulation

Lemma implies the *inf-sup* condition

$$\inf_{\substack{\phi\in L^2(\Omega)\\ \psi\in L^2(\partial_2\Omega)}} \sup_{q\in\widehat{Q}_F} \frac{\int_\Omega \phi \operatorname{div} q\, dx + \int_{\partial_2\Omega} \psi q\cdot n\, ds}{\|q\|_{\operatorname{div}} \left(\|\phi\|^2 + \|\psi\|_{\partial_2\Omega}^2\right)^{1/2}} \geq C_0 > 0\,.$$

The Dual Problem with respect to the Lagrangian $\widehat{L}$: Let us now construct the dual functional $\widehat{I}^*$. It is easy to see that

$$\widehat{I}^*(\widehat{q}) = \inf_{\widehat{v}} \widehat{L}(\widehat{v},\widehat{q}) =$$

$$= \inf_{\widehat{v}} \left\{ -\frac{1}{2} \| \widehat{q} \|_*^2 - \int_\Omega v(\operatorname{div}\widehat{q})dx - \int_\Omega fvdx - \int_{\partial_2\Omega} Fu_0 ds + g(u_0,\widehat{q}) \right\} =$$

$$= -\frac{1}{2} \| \widehat{q} \|_*^2 + g(u_0,\widehat{q}) - \int_{\partial_2\Omega} Fu_0\, ds$$

provided that $\operatorname{div}\widehat{q} + f = 0$ (in the L_2-sense). In all other cases $\widehat{I}^*(\widehat{q}) = -\infty$.

Since $\operatorname{div}\widehat{q} = -f$, we find that the dual functional for such a case has the form

$$\widehat{I}^*(q) = -\frac{1}{2} \| \widehat{q} \|_*^2 + \int_\Omega (\nabla u_0 \cdot \widehat{q} - f u_0)\, dx - \int_{\partial_2\Omega} Fu_0\, ds$$

$$= \int_\Omega \nabla u_0 \cdot \widehat{q} dx - \frac{1}{2} \| \widehat{q} \|_*^2 - \boldsymbol{\ell}(u_0)\,,$$

Since $\widehat{q} \in \widehat{Q}_F$, we have (recall that $\operatorname{div}\widehat{q} = -f$)

$$\int_\Omega \nabla w \cdot \widehat{q} dx = -\int_\Omega (\operatorname{div}\widehat{q}) w dx + \int_{\partial_2\Omega} Fw ds \quad \forall w \in V_0\,.$$

we see that $\widehat{q}$ satisfies the relation

$$\int_\Omega \nabla w \cdot \widehat{q} dx = \boldsymbol{\ell}(w) \quad \forall w \in V_0\,.$$

In other cases, $\widehat{I}^*(\widehat{q}) = -\infty$.

Thus, Problems $\mathcal{P}^*$ and $\widehat{\mathcal{P}}^*$ coincide and are reduced to the maximization of I^* on the set $Q_{\boldsymbol{\ell}}$. This means that

$$\sup \mathcal{P}^* = \sup \widehat{\mathcal{P}}^*.$$

Since the saddle point of $\widehat{L}$ exists, we have

$$\widehat{L}(\widehat{u},\widehat{p}) = \inf \widehat{\mathcal{P}} = \sup \widehat{\mathcal{P}}^*,$$

but

$$\sup \widehat{\mathcal{P}}^* = \sup \mathcal{P}^* = \inf \mathcal{P}.$$

Thus, we infer that

$$\inf \widehat{\mathcal{P}} = \inf \mathcal{P}.$$

Thus, we conclude that $u \in V_0 + u_0$ (minimizer of $\mathcal{P}$) also minimizes $\widehat{J}$ on $\widehat{V}$. Analogously, if $p \in Q_\ell$ is the maximizer of Problem $\mathcal{P}^*$, then

$$\int_\Omega \nabla w \cdot p \, dx = \int_\Omega f w \, dx + \int_{\partial_2\Omega} F w \, ds \quad \forall w \in V_0 \, .$$

From here we see that $\operatorname{div} p + f = 0$ a.e. in Ω and, hence,

$$\int_\Omega (\nabla w \cdot p + (\operatorname{div} p) w) \, dx = \int_{\partial_2\Omega} F w \, ds \quad \forall w \in V_0 \, ,$$

that is $p \in \widehat{Q}_F$. Thus, p is also the maximizer of Problem $\widehat{\mathcal{P}}^*$.

The reverse statement that the solutions of $\widehat{\mathcal{P}}$, $\widehat{\mathcal{P}}^*$ are also the solutions of $\mathcal{P}$, $\mathcal{P}^*$ is not difficult to prove as well.

Hence, both mixed formulations have the same solution (u, p) which is in fact the generalized solution of our problem.

1.4 A priori error estimation methods

First error relation

First we present the algebraic identity

$$\begin{aligned} &\frac{1}{2}B(u-v,u-v) = \frac{1}{2}B(v,v) - < f, v > + \\ &+ < f, u > - \frac{1}{2}B(u,u) - B(u, v-u) + < f, v-u > = \\ &= J(v) - J(u) - B(u, v-u) + < f, v-u > \end{aligned} \tag{1.99}$$

From this identity we derive two important results:

- (a) Minimizer u satisfies $B(u,w) =< f,w > \forall w$;
- (b) Error is subject to the difference of functionals.

Integral identity

Let us show (a), i.e., that from (1.99) it follows the identity

$$B(u, v-u) =< f, v-u > \qquad \forall v \in K.$$

Indeed, assume the opposite, i.e. $\exists \bar{v} \in K$ such that

$$B(u, \bar{v}-u) - < f, \bar{v}-u >= \delta > 0 \qquad (\bar{v} \neq u!)$$

Set $\widetilde{v} := u + \alpha(\bar{v}-u)$, $\alpha \in \mathbb{R}$. Then $\widetilde{v} - u = \alpha(\bar{v}-u)$ and

$$\frac{1}{2}B(u-\widetilde{v}, u-\widetilde{v}) + B(u, \widetilde{v}-u) - < f, \widetilde{v}-u >= \\ = \frac{\alpha^2}{2}B(\bar{v}-u, \bar{v}-u) + \alpha\delta = J(\widetilde{v}) - J(u) \geq 0$$

However, for arbitrary α such an inequality cannot be true. Denote $a = B(\bar{v}-u, \bar{v}-u)$. Then in the left–hand side we have a function $1/2\alpha^2 a^2 + \alpha\delta$, which always attains negative values for certain α. For example, set $\alpha = -\delta/a^2$. Then, the left–hand side is equal to $-\frac{1}{2}\delta^2/a^2 < 0$ and we arrive at a contradiction.

Error estimate

Now, we show (b). From

$$\frac{1}{2}B(u-v, u-v) = \\ = J(v) - J(u) - B(u, v-u) + < f, v-u >$$

we obtain the error estimate [8]:

$$\frac{1}{2}B(u-v,u-v)=J(v)-J(u). \tag{1.100}$$

which immediately gives the projection estimate.

Projection estimate

Let u_h be a minimizer of J on $V_h \subset V$. Then

$$\frac{1}{2}B(u-u_h,u-u_h)=J(u_h)-J(u)\le J(v_h)-J(u)=$$
$$=\frac{1}{2}B(u-v_h,u-v_h)\quad \forall\, v_h\in V_h.$$

and we observe that

$$\boxed{B(u-u_h,u-u_h)=\inf_{v_h\in V_h}B(u-v_h,u-v_h)} \tag{1.101}$$

Projection type estimates serve a basis for deriving a priori convergence estimates.

Interpolation in Sobolev spaces

A priori rate convergence estimates are based upon two two key points:
PROJECTION ERROR ESTIMATE and INTERPOLATION OF FUNCTIONS IN SOBOLEV SPACES.

Interpolation theory investigates the difference between a function in a Sobolev space and its piecewise polynomial interpolant. Basic estimate on a simplex T_h is

$$|v-\Pi_h v|_{m,t,T_h}\le C(m,n,t)\left(\frac{h}{\rho}\right)^m h^{2-m}\|v\|_{2,t,T_h},$$

and on the whole domain

$$|v-\Pi_h v|_{m,t,\Omega_h}\le Ch^{2-m}\|v\|_{2,t,\Omega_h}.$$

Here h is a the element size and ρ is the inscribed ball diameter.

[8] S. G. Mikhlin. *Variational methods in mathematical physics*. Pergamon, Oxford, 1964.

Asymptotic convergence estimates

Typical case is $m = 1$ and $t = 2$. Since

$$B(u - u_h, u - u_h) \leq B(u - \Pi_h u, u - \Pi_h u) \leq c_2 \|u - \Pi_h u\|^2$$

for

$$B(w, w) = \int_\Omega \nabla w \cdot \nabla w \, dx$$

we find that

$$\|\nabla(u - u_h)\| \leq Ch|u|_{2,2,\Omega}.$$

provided that

- Exact solution is H^2 – regular;
- u_h is the Galerkin approximation;
- Elements do not "degenerate" in the refinement process.

A priori convergence estimates cannot guarantee that the error monotonically decreasesas $h \to 0$.

Besides, in practice we are interested in the error of a concrete approximation on a particular mesh. Asymptotic estimates could hardly be helpful in such a context because, in general, the constant C serves for the *whole class* of approximate solutions of a particular type. Typically it is either unknown or highly overestimated.

Remark 1.4.1 *A priori convergence estimates have mainly a theoretical value: they show that an approximation method is correct "in principle.*

Remark 1.4.2 *For these reasons, a quite different approach to error control is rapidly developing. Nowadays it has already formed a new direction:* *A Posterio*

Chapter 2

A posteriori error estimation methods developed in 1900-1975

2.1 Runge's rule

At the end of 19th century a heuristic error control method was suggested by C. Runge who investigated numerical integration methods for ordinary differential equations.

> Heuristic rule of C. Runge If the difference between two approximate solutions computed on a coarse mesh $\mathcal{T}_h$ with mesh size h and refined mesh $\mathcal{T}_{h_{ref}}$ with mesh size h_{ref} (e.g., $h_{ref} = h/2$) has become small, then both $u_{h_{ref}}$ and u_h are probably close to the exact solution.

In other words, this rule can be formulated as follows:

$$\boxed{\text{If} \qquad u_h - u_{h_{ref}} \text{ is small then } \; u_{h_{ref}} \text{ is close to } u}$$

where $\cdot$ is a certain functional or mesh-dependent norm.

> Also, the quantity $u_h - u_{h_{ref}}$ can be viewed (in terms of modern terminology) as a certain a posteriori error indicator.

Runge's heuristic rule is simple and was easily accepted by numerical analysts.

Remark 2.1.1 *However, if we do not properly define the quantity $\mathcal{I}\cdot$, for which $\mathcal{I}\cdot u_h - u_{h_{ref}}$ is small, then the such a principle may be not true.*

One can present numerous examples where two subsequent elements of an approximation sequence are close to each other, but far from a certain joint limit. For example, such cases often arise in the minimization (maximization) of functionals with "saturation" type behavior or with a "sharp–well" structure. Also, the rule may lead to a wrong presentation if, e.g., the refinement has not been properly done, so that new trial functions were added only in subdomains were an approximation is almost coincide with the true solution. Then two subsequent approximations may be very close, but at the same time not close to the exact solution.

Remark 2.1.2 *Also, in practice, we need to now precisely what the word "close" means, i.e. we need to have a more concrete presentation on the error. For example, it would be useful to establish the following rule:*

If $\quad u_h - u_{h_{ref}} \leq \epsilon \quad$ *then* $\quad \|u_h - u\| \leq \delta(\epsilon)$,

where the function $\delta(\epsilon)$ *is known and computable.*

In subsequent lectures we will see that for a wide class of boundary–value problems it is indeed possible to derive such type generalizations of the Runge's rule.

2.2 The estimate of Prager and Synge

Prager and Synge derived an estimate on the basis of purely geometrical grounds [1]. In modern terms, there result for the problem

$$\begin{aligned} &\Delta u + f = 0, \qquad \text{in } \Omega, \\ &u = 0, \qquad \text{on } \partial\Omega \end{aligned}$$

reads as follows:

$$\|\nabla(u - v)\|^2 + \|\nabla u - \tau\|^2 = \|\nabla v - \tau\|^2,$$

[1] W. Prager and J. L. Synge. Approximation in elasticity based on the concept of function spaces, *Quart. Appl. Math.* 5(1947)

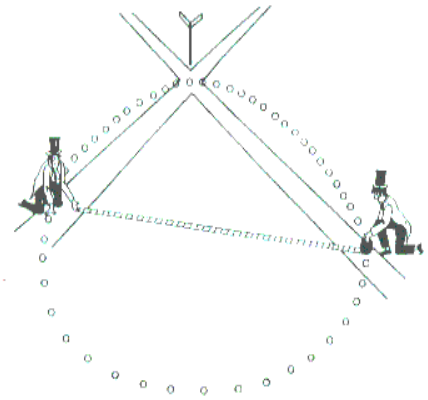


Figure 2.1: "Two blind men and their hypercircle".

where τ is a function satisfying the equation $\operatorname{div}\tau + f = 0$.
We can easily prove it by the *orthogonality relation*

$$\int_\Omega \nabla(u-v)\cdot(\nabla u - \tau)\,dx = 0 \qquad (\operatorname{div}(\nabla u - \tau) = 0\,!).$$

From here, it also follows that

$$\|\nabla(u-v)\| = \inf_{q\in Q_f} \|\nabla v - q\|. \tag{2.1}$$

This relation and its analogs for more complicated problems generate various a posteriori estimates that use equilibration of the dual variable (flux).

2.3 Estimate of Mikhlin

A similar estimate follows from the First error relation and can be justified by variational arguments[2]*. It is as follows:*

$$\frac{1}{2}\|\nabla(u-v)\|^2 \le J(v) - inf J,$$

where

$$J(v) := \frac{1}{2}\|\nabla v\|^2 - (f,v), \qquad inf J := \inf_{v\in \overset{\circ}{H}_1(\Omega)} J(v).$$

Dual problem

Since

$$inf J = \sup_{\tau\in Q_f}\left\{-\frac{1}{2}\|\tau\|^2\right\},$$

where

$$Q_f := \left\{\tau \in L_2(\Omega, R^d) \mid \int_\Omega \tau\cdot\nabla w\, dx = \int_\Omega f w\, dx \qquad \forall w \in \overset{\circ}{H}{}^1\right\},$$

we find that

$$\frac{1}{2}\|\nabla(u-v)\|^2 \le J(v) + \frac{1}{2}\|\tau\|^2, \qquad \forall \tau \in Q_f.$$

Since

$$\begin{aligned} J(v) + \tfrac{1}{2}\|\tau\|^2 &= \frac{1}{2}\|\nabla v\|^2 - \int_\Omega f v\, dx + \frac{1}{2}\|\tau\|^2 = \\ &= \frac{1}{2}\|\nabla v\|^2 - \int_\Omega \tau\cdot\nabla v\, dx + \frac{1}{2}\|\tau\|^2 = \\ &= \frac{1}{2}\|\nabla v - \tau\|^2 \end{aligned}$$

[2] S. G. Mikhlin. *Variational methods in mathematical physics.* Pergamon, Oxford, 1964.

we arrive at the estimate

$$\boxed{\frac{1}{2}\|\nabla(u-v)\|^2 \le \frac{1}{2}\|\nabla v - \tau\|^2, \qquad \forall \tau \in Q_f.} \tag{2.2}$$

Comment. Estimates of Prager and Synge and of Mikhlin are valid for any $v \in \overset{\circ}{H}_1(\Omega)$*, so that, formally, that they can be applied to any conforming approximation of the problem. However, from the practical viewpoint these estimates have an essential drawback:*
they use a function τ *in the set* Q_f *defined by the differential relation,*
which may be difficult to satisfy exactly. Probably by this reason further development of a posteriori error estimates for Finite Element Methods (especially in 80'-90') was mainly based on different grounds.

2.4 A posteriori estimates for iteration methods

2.4.1 Fixed point theorem

Consider a Banach space (X, d) and a continuous operator

$$\mathfrak{T} : X \to X.$$

Definition 2.4.1 *A point $x_\odot$ is called a fixed point of $\mathfrak{T}$ if*

$$x_\odot = \mathfrak{T} x_\odot \,. \tag{2.3}$$

Approximations of a fixed point are usually constructed by the iteration sequence

$$x_i = \mathfrak{T} x_{i-1} \qquad i = 1, 2, \dots \,. \tag{2.4}$$

Two basic problems:

(a) find the conditions that guarantee convergence of x_i to $x_\odot$,

(b) find computable estimates of the error $e_i = d(x_i, x_\odot)$.

Definition 2.4.2 *An operator $\mathfrak{T} : X \to X$ is called q-contractive on a set $S \subset X$ if there exists a positive real number q such that the inequality*

$$d(\mathfrak{T} x, \mathfrak{T} y) \le q\, d(x, y) \tag{2.5}$$

holds for any elements x and y of the set S.

2.4.2 Banach theorem

Theorem 2.4.1 (S. Banach) *Let $\mathfrak{T}$ be a q-contractive mapping of a closed nonempty set $S \subset X$ to itself with $q < 1$. Then, $\mathfrak{T}$ has a unique fixed point in S and the sequence x_i obtained by (2.4) converges to this point.*

Proof. It is easy to see that

$$d(x_{i+1}, x_i) = d(\mathfrak{T}x_i, \mathfrak{T}x_{i-1}) \leq qd(x_i, x_{i-1}) \leq ... \leq q^i d(x_1, x_0).$$

Therefore, for any $m > 1$ we have

$$\begin{aligned} d(x_{i+m}, x_i) \leq \\ &\leq d(x_{i+m}, x_{i+m-1}) + d(x_{i+m-1}, x_{i+m-2}) + ... + d(x_{i+1}, x_i) \leq \\ &\leq q^i(q^{m-1} + q^{m-2} + ... + 1)d(x_1, x_0)\,. \end{aligned} \quad (2.6)$$

Since

$$\sum_{k=0}^{m-1} q^k \leq \frac{1}{1-q}\,,$$

(2.6) implies the estimate

$$d(x_{i+m}, x_i) \;\leq\; \frac{q^i}{1-q} d(x_1, x_0). \quad (2.7)$$

Let $i \to \infty$, then the right-hand side of (2.7) tends to zero, so that $\{x_i\}$ is a Cauchy sequence. It has a limit in $y \in X$.

Then, $d(x_i, y) \to 0$ and

$$d(\mathfrak{T}x_i, \mathfrak{T}y) \;\leq\; qd(x_i, y) \to 0$$

so that $d(\mathfrak{T}x_i, \mathfrak{T}y) \to 0$ and $\mathfrak{T}x_i \to \mathfrak{T}y$. Pass to the limit in (2.4) as $i \to +\infty$. We observe that

$$\mathfrak{T}y = y.$$

Hence, any limit of such a sequence is a fixed point.

It is easy to prove that a fixed point is unique.

Assume that there are two different fixed points $x^1_\odot$ and $x^2_\odot$, i.e.

$$\mathfrak{T}x^k_\odot = x^k_\odot, \qquad k = 1, 2.$$

Therefore,

$$d(x^1_\odot, x^2_\odot) = d(\mathfrak{T}x^1_\odot, \mathfrak{T}x^2_\odot) \leq qd(x^1_\odot, x^2_\odot)\,.$$

But $q < 1$, and thus such an inequality cannot be true.

2.4.3 A priori convergence estimate

Let

$$e_j = d(x_j, x_\odot)$$

denote the error on the j-th step. Then

$$e_j = d(\mathfrak{T}x_{j-1}, \mathfrak{T}x_\odot) \leq qe_{j-1} \leq q^j e_0.$$

and

$$e_j \leq q^j e_0. \tag{2.8}$$

This estimate gives a certain presentation on that how the error decreases. However, this a priori upper bound may be rather coarse.

2.4.4 A posteriori estimates for contractive mappings

A posteriori estimates

The proposition below furnishes upper and lower estimates of e_j, which are easy to compute provided, that the number q (or a good estimate of it) is known.

Theorem 2.4.2 ([3]) *Let $\{x_j\}_{j=0}^{\infty}$ be a sequence obtained by the iteration process*

$$x_i \;=\; \mathfrak{T}x_{i-1} \qquad i = 1, 2, ...$$

with a mapping $\mathfrak{T}$ satisfying the condition $\|\mathfrak{T}\| = q \leq 1$. Then, for any x_j, $j > 1$, the following estimate holds:

$$M_{\ominus}^{j} := \frac{1}{1+q} d(x_{j+1}, x_j) \leq e_j \leq M_{\oplus}^{j} := \frac{q}{1-q} d(x_j, x_{j-1}). \tag{2.9}$$

Proof. The upper estimate in (2.9) follows from (2.7). Indeed, put $i = 1$ in this relation. We have

$$d(x_{1+m}, x_1) \leq \frac{q}{1-q} d(x_1, x_0)\,.$$

Since $x_{1+m} \to x_{\odot}$ as $m \to +\infty$, we pass to the limit with respect to m and obtain

$$d(x_{\odot}, x_1) \leq \frac{q}{1-q} d(x_1, x_0)\,.$$

We may view x_{j-1} as the starting point of the sequence. Then, in the above relation $x_0 = x_{j-1}$ and $x_1 = x_j$ and we arrive at the following *upper bound of the error:*

$$d(x_{\odot}, x_j) \leq \frac{q}{1-q} d(x_j, x_{j-1})\,.$$

The lower bound of the error follows from the relation

$$d(x_j, x_{j-1}) \leq d(x_j, x_{\odot}) + d(x_{j-1}, x_{\odot}) \leq (1+q) d(x_{j-1}, x_{\odot}),$$

which shows that

$$d(x_{j-1}, x_{\odot}) \geq \frac{1}{1+q} d(x_j, x_{j-1})\,.$$

Note that

$$\frac{M^j_{\oplus}}{M^j_{\ominus}} = \frac{q(1+q)}{1-q}\frac{d(x_j, x_{j-1})}{d(x_{j+1}, x_j)} \geq \frac{1+q}{1-q},$$

we see that that the efficiency of the upper and lower bounds given by (2.9) deteriorates as $q \to 1$.

If X is a normed space, then

$$d(x_{j+1}, x_j) = \|R(x_j)\|,$$

where

$$R(x_j) := \mathfrak{T}x_j - x_j$$

is the residual of the basic equation (2.3). Thus, the upper and lower estimates of errors are expressed in terms of the *residuals of the respective iteration equation* computed for two neighbor steps:

$$\frac{1}{1+q}\|R(x_j)\| \le e_j = d(x_j, x_\odot) \le \frac{q}{1-q}\|R(x_{j-1})\|.$$

2.4.5 Corollaries

In the iteration methods, it is often easier to analyze the operator

$$\mathfrak{T} = T^n := \underbrace{TT...T}_{n \text{ times}}$$

where T is a certain mapping.

Proposition 2.4.1 (1) *Let $T : S \to S$ be a continuous mapping such that $\mathfrak{T}$ is a q-contractive mapping with $q \in (0,1)$. Then, the equations*

$$x = Tx \qquad \text{and} \qquad x = \mathfrak{T}x$$

have one and the same fixed point, which is unique and can be found by the above described iteration procedure.

Proof. By the Banach Theorem, we observe that the operator $\mathfrak{T}$ has a unique fixed point $\boldsymbol{\xi}_\odot$.

Let us show that $\boldsymbol{\xi}_\odot$ is a fixed point of T. First, we note that

$$T\boldsymbol{\xi}_\odot = T(\mathfrak{T}\boldsymbol{\xi}_\odot) = T\mathfrak{T}^2\boldsymbol{\xi}_\odot = ... = T\mathfrak{T}^i\boldsymbol{\xi}_\odot = T^{(1+in)}\boldsymbol{\xi}_\odot = T^{in}T\boldsymbol{\xi}_\odot. \quad (2.10)$$

Denote $x_0 = T\boldsymbol{\xi}_\odot$. By (2.10) we conclude that for any i

$$T\boldsymbol{\xi}_\odot = \mathfrak{T}^i x_0. \quad (2.11)$$

Passing to the limit on the right-hand side in (2.11), we arrive at the relation $T\boldsymbol{\xi}_{\odot} = \boldsymbol{\xi}_{\odot}$, which means that $\boldsymbol{\xi}_{\odot}$ is a fixed point of the operator T.

Let $\widetilde{x_{\odot}}$ be a fixed point of T. Then,

$$\widetilde{x_{\odot}} = T^2\widetilde{x_{\odot}} = .. = T^n\widetilde{x_{\odot}} = \mathfrak{T}\widetilde{x_{\odot}}$$

and we observe that $\widetilde{x_{\odot}}$ is a fixed point of $\mathfrak{T}$. Since the saddle point of $\mathfrak{T}$ exists and is unique, we conclude that

$$\boldsymbol{\xi}_{\odot} = \widetilde{x_{\odot}}.$$

Remark 2.4.1 *This assertion may be practically useful if it is not possible to prove that T is q–contractive, but this fact can be established for a certain power of T.*

2.4.6 Iteration methods for bounded linear operators

Consider a bounded linear operator $\mathcal{L} : X \to X$, where X is a Banach space. Given $b \in X$, the iteration process is defined by the relation

$$x_j = \mathcal{L}\, x_{j-1} + b. \tag{2.12}$$

Let $x_\odot$ be a fixed point of (2.12) and

$$\|\mathcal{L}\| = q < 1.$$

By applying the Banach Theorem it is easy to show that

$$\{x_j\} \to x_\odot.$$

Indeed, let $\bar{x}_j = x_j - x_\odot$. Then

$$\bar{x}_j = \mathcal{L}x_{j-1} + b - x_\odot = \mathcal{L}(x_{j-1} - x_\odot) = \mathcal{L}\bar{x}_{j-1}\,. \tag{2.13}$$

Since

$$0_X = \mathcal{L}\, 0_X,$$

we note that the zero element 0_X is a unique fixed point of the operator $\mathcal{L}$. By the Banach theorem $\bar{x}_j \to 0_X$ and, therefore, $\{x_j\} \to x_\odot$.

Therefore, we have an *a priori* estimate

$$\|x_j - x_\odot\|_X = \|\bar{x}_j - 0_X\|_X \le \frac{q^j}{1-q}\,\|\bar{x}_1 - \bar{x}_0\|_X = \frac{q^j}{1-q}\,\|R(x_0)\|_X \tag{2.14}$$

and the *a posteriori* one

$$\|x_j - x_\odot\|_X \le \frac{q}{1-q}\,\|R(x_{j-1})\|_X\,, \tag{2.15}$$

where $R(z) = \mathcal{L}z + b - z$ is the *residual* of the functional equation considered.

By applying the general theory, we also obtain a lower bound of the error

$$\|x_j - x_\odot\|_X \ge \frac{1}{1+q}\,\|x_{j+1} - x_j\|_X = \frac{1}{1+q}\,\|R(x_j)\|_X\,. \tag{2.16}$$

Hence, we arrive at the following estimates for the error in the linear operator equation:

$$\frac{1-q}{q}\,\|x_j - x_\odot\|_X \le \|R(x_{j-1})\|_X \le (1+q)\,\|x_{j-1} - x_\odot\|_X\,.$$

2.4.7 Iteration methods in linear algebra

Important applications of the above results are associated with systems of linear simultaneous equations and other algebraic problems. Set $X = \mathbb{R}^d$ and assume that $\mathcal{L}$ is defined by a nondegenerate matrix $A \in \mathbb{M}^{d\times d}$ decomposed into three matrixes

$$A = A_\ell + A_d + A_r,$$

where A_ℓ, A_r, and A_d are certain lower, upper, and diagonal matrices, respectively.

Iteration methods for systems of linear simultaneous equations associated with A are often represented in the form

$$B\frac{x_i - x_{i-1}}{\tau} + A\,x_{i-1} = f\,. \tag{2.17}$$

In (2.17), the matrix B and the parameter τ may be taken in various ways (depending on the properties of A). We consider three frequently encountered cases:

(a) $B = A_d$,

(b) $B = A_d + A_\ell$,

(c) $B = A_d + \omega A_\ell$, $\tau = \omega$.

For $\tau = 1$, (a) and (b) lead to the methods of Jacobi and Zeidel, respectively. In (c), the parameter ω must be in the interval $(0, 2)$. If $\omega > 1$, we have the so-called "upper relaxation method", and $\omega < 1$ corresponds to the "lower relaxation method".

The method (2.17) is reduced to (2.12) if we set

$$\mathcal{L} = \mathbb{I} - \tau B^{-1} A \qquad \text{and} \qquad b = \tau B^{-1} f\,, \tag{2.18}$$

where $\mathbb{I}$ is the unit matrix. It is known that x_i converges to $x_\odot$ that is a solution of the system

$$A\, x_\odot = f \tag{2.19}$$

if an only if all the eigenvalues of $\mathcal{L}$ are less than one.

Obviously, B and τ should be taken in such a way that they guarantee the fulfillment of this condition.

Assume that $\|\mathcal{L}\| \le q < 1$. In view of (2.14)-(2.16), the quantities

$$M^i_\oplus = \; q(1-q)^{-1} \, \|R(x_{i-1})\|\,, \tag{2.20}$$

$$M^{0i}_\oplus = \; q^i(1-q)^{-1} \, \|R(x_0)\|\,, \tag{2.21}$$

$$M^i_\ominus = \; (1+q)^{-1} \, \|R(x_i)\| \tag{2.22}$$

furnish upper and lower bounds of the error for the vector x_i.

Remark 2.4.2 *It is worth noting that from the practical viewpoint finding an upper bound for $\|\mathcal{L}\|$ and proving that it is less than 1 presents a special and often not easy task.*

If q is very close to 1, then the convergence of an iteration process may be very slow. As we have seen, in this case, the quality of error estimates is also degraded. A well–accepted way for accelerating the convergence consists of using a modified system obtained from the original one by means of a suitable *preconditioner* P^{-1} and solving the system

$$\left(P^{-1}A\right) x = P^{-1} f$$

with a smaller condition number. Of cause, the best preconditioner is the unknown matrix A^{-1}. Therefore, a preconditioner is often constructed from the parts of A that are not difficult to invert (e.g., in the simplest case it is taken as the matrix inverse to the diagonal part of A). This iteration technique is well presented in the literature[4]

[4]see, e.g., O. Axelsson. *Iterative solution methods.* Cambridge University Press, Cambridge, 1994.

Task 2.4.1 *Consider the problem*

$$Ax = f$$

for a symmetric matrix A with coefficients

$$a_{ij} = \kappa/ij \quad \text{if } i \neq j, \quad \kappa = 0.1$$
$$a_{ii} = i.$$

In this case $\lambda_{max}(A) = 200$, $\lambda_{min}(A) = 0.8224$ and $Cond(A) = 24L10.2030$ Solve the system by the iteration method

$$x_{i+1} = (\mathbb{I} - \tau B^{-1}A)\, x_i + \tau B^{-1}F$$

with $B = A_D$ and $x_0 = \{0, 0, ...0\}$, determine q and define two–sided error bounds.

In this example $n = 200$, q= 0.662, and $\tau = 0.760$. The values of the error and the estimates are presented below.

Table 2.1:

i	$M^i_{\ominus}$	$\|e\|$	$M^i_{\oplus}$	$M^{0i}_{\oplus}$
1	.187145E+03	.412471E+03	.245893E+04	.245893E+04
2	.452820E+02	.104019E+03	.610732E+03	.162904E+04
3	.123433E+02	.311517E+02	.147774E+03	.107924E+04
4	.405504E+01	.116679E+02	.402813E+02	.714995E+03
5	.166633E+01	.517711E+01	.132333E+02	.473684E+03
6	.767379E+00	.244532E+01	.543792E+01	.313815E+03
7	.366283E+00	.117450E+01	.250428E+01	.207902E+03
8	.176340E+00	.566166E+00	.119533E+01	.137735E+03
16	.515722E-03	.165576E-02	.349042E-02	.511127E+01
17	.248671E-03	.798371E-03	.168302E-02	.338621E+01
18	.119903E-03	.384956E-03	.811515E-03	.224336E+01
19	.578146E-04	.185617E-03	.391295E-03	.148623E+01
20	.278769E-04	.895001E-04	.188673E-03	.984624E+00

2.4.8 Applications to integral equations

Many problems in science and engineering can be stated in terms of integral equations. One of the most typical cases is to find a function $x_\odot(t) \in C[a,b]$ such that

$$x_\odot(t) = \lambda \int_a^b K(t,s)\, x_\odot(s)\, ds + f(t), \tag{2.23}$$

where $\lambda \geq 0$, K (the kernel) is a continuous function for

$$(x,t) \in Q := \{a \leq s \leq b,\ a \leq t \leq b\}$$

and

$$|K(t,s)| \leq M, \qquad \forall (t,s) \in Q.$$

Also, we assume that $f \in C[a,b]$.

Let us define the operator $\mathfrak{T}$ as follows:

$$y(t) := \mathfrak{T}x(t) := \lambda \int_a^b K(t,x)x(s)\, ds + f(t) \tag{2.24}$$

and show that $\mathfrak{T}$ maps continuous functions to continuous ones. Let t_0 and $t_0 + \Delta t$ belong to $[a,b]$. Then,

$$|y(t_0+\Delta t) - y(t_0)| \leq$$
$$\leq |\lambda| \int_a^b |K(t_0+\Delta t, s) - K(t_0,s)||x(s)|\, ds +$$
$$+ |f(t_0+\Delta t) - f(t_0)|.$$

Since K and f are continuous on the compact sets Q and $[a,b]$, respectively, they are uniformly continuous on these sets.

Therefore, for any given ϵ one can find a small number δ such that

$$|f(t_0+\Delta t) - f(t_0)| < \epsilon$$

and

$$|K(t_0+\Delta t, s) - K(t_0,s)| < \epsilon,$$

provided that $|\Delta t| < \delta$.

Thus, we have

$$|y(t_0 + \Delta t) - y(t_0)| \le \epsilon(|\lambda||b-a| \max_{s\in[a,b]} |x(s)| + 1) = C\epsilon,$$

and, consequently, $y(t_0 + \Delta t)$ tends to $y(t_0)$ as $|\Delta t| \to 0$.

$\mathfrak{T} : C[a,b] \to C[a,b]$ is a contractive mapping. Indeed,

$$d(\mathfrak{T}x, \mathfrak{T}y) = \max_{a\le t\le b} |\mathfrak{T}x(t) - \mathfrak{T}y(t)| =$$

$$= \max_{a\le t\le b} \left| \lambda \int_a^b K(t,s)(x(s) - y(s))\, ds \right| \le$$

$$\le |\lambda| M(b-a) \max_{a\le s\le b} |x(s) - y(s)| = |\lambda| M(b-a) d(x,y),$$

so that $\mathfrak{T}$ is a q-contractive operator with

$$q = |\lambda| M(b-a), \tag{2.25}$$

provided that

$$|\lambda| < \frac{1}{M(b-a)}. \tag{2.26}$$

2.4.9 Numerical procedure

An approximate solution of (2.23) can be found by the iteration method

$$x_{i+1}(t) = \lambda \int_a^b K(t,s) x_i(s)\, ds + f(t). \tag{2.27}$$

If (2.26) holds, then from the Banach theorem it follows that the sequence $\{x_i\}$ converges to the exact solution.

We apply the theory exposed above and find that the accuracy of x_i is subject to the estimate

$$\frac{1}{1+q} \int_a^b K(t,s)(x_{i+1}(s) - x_i(s))\, ds \le$$

$$\le \max_{a\le t\le b} |x_i(t) - x_\odot(t)| \le \frac{q}{1-q} \int_a^b K(t,s)(x_i(s) - x_{i-1}(s))\, ds. \tag{2.28}$$

2.4.10 Applications to Volterra type equations

Consider the fixed point problem

$$x_{\odot}(t) = \lambda \int_a^t K(t,s)\, x_{\odot}(s)\, ds + f(t), \tag{2.29}$$

where

$$|K(t,s)| \le M, \qquad \forall (t,s) \in Q$$

and $f \in C[a,b]$.

Define the operator T as follows:

$$Tx(t) = \lambda \int_a^t K(t,s)\, x(s)\, ds + f(t).$$

Similarly, to the previous case we establish that

$$d(Tx, Ty) \le |\lambda| M (t-a) d(x,y).$$

By the same arguments we find that

$$d(T^n x, T^n y) \le |\lambda|^n M^n \frac{(t-a)^n}{n!} d(x,y),$$

Thus, the operator $\mathfrak{T} := T^n$ is q-contractive with a certain $q < 1$, provided that n is large enough.

In view of Proposition 1, we conclude that the iteration method converges to $x_{\odot}$ and the errors are controlled by the two–sided error estimates.

2.4.11 Applications to ordinary differential equations

Let u be a solution of the simplest initial boundary-value problem

$$\frac{du}{dt} = \varphi(t, u(t)), \quad u(t_0) = a, \tag{2.30}$$

where the solution $u(t)$ is to be found on the interval $[t_0, t_1]$. Assume that the function $\varphi(t, p)$ is continuous on the set

$$Q = \{t_0 \le t \le t_1,\ a - \Delta \le p \le a + \Delta\}$$

and

$$|\varphi(t, p_1) - \varphi(t, p_2)| \le L|p_1 - p_2|, \quad \forall (t, p) \in Q. \tag{2.31}$$

Problem (2.30) can be reduced to the integral equation

$$u(t) = \int_{t_0}^{t} \varphi(s, u(s))\, ds + a \tag{2.32}$$

and it is natural to solve the latter problem by the iteration method

$$u_j(t) = \int_{t_0}^{t} \varphi(s, u_{j-1}(s))\, ds + a. \tag{2.33}$$

To justify this procedure, we must verify that the operator

$$\mathfrak{T}u := \int_{t_0}^{t} \varphi(s, u(s))\, ds + a$$

is q-contractive with respect to the norm

$$\|u\| := \max_{t \in [t_0, t_1]} |u(t)|. \tag{2.34}$$

We have

$$\begin{aligned} \|\mathfrak{T}z - \mathfrak{T}y\| = \max_{t \in [t_0, t_1]} \left| \int_{t_0}^{t} (\varphi(s, z(s)) - \varphi(s, y(s)))\, ds \right| \le \\ \le \max_{t \in [t_0, t_1]} L \int_{t_0}^{t} |z(s) - y(s)|\, ds \le L \int_{t_0}^{t_1} |z(s) - y(s)|\, ds \le \\ \le L(t_1 - t_0) \max_{s \in [t_0, t_1]} |z(s) - y(s)| = L(t_1 - t_0)\|z - y\|. \end{aligned}$$

We see that if

$$t_1 < t_0 + L^{-1}, \tag{2.35}$$

then the operator $\mathfrak{T}$ is q-contractive with

Remark 2.4.3 $q := L(t_1 - t_0) < 1$.

Therefore, if the interval $[t_0, t_1]$ is small enough(i.e., it satisfies the condition 2.35), then the existence and uniqueness of a continuous solution $u(t)$ follows from the Banach theorem. In this case, the solution can be found by the iteration procedure whose accuracy is explicitly controlled by the two–sided error estimates[5]

2.5 A posteriori methods based on monotonicity

The theory of monotone operatorsgives another way of constructing a posteriori estimates.

Monotone operators are defined on the so–called ordered(or partially ordered) spaces that introduce the relation $x \leq y$ for all (or almost all) elements x, y of the space.

Definition 2.5.1 *An operator $\mathfrak{T}$ is called monotone if $x \leq y$ implies $\mathfrak{T}x \leq \mathfrak{T}y$.*

Consider the fixed point problem

$$x_\odot = \mathfrak{T}x_\odot + f$$

on an ordered (partially ordered) space X. Assume that

$$\mathfrak{T} = \mathfrak{T}_\oplus + \mathfrak{T}_\ominus,$$

$\mathfrak{T}_\oplus$ is monotone,
$\mathfrak{T}_\ominus$ is antitone: $x \leq y$ implies $\mathfrak{T}x \geq \mathfrak{T}y$,
$\mathfrak{T}_\oplus$ and $\mathfrak{T}_\ominus$ have a common set of images D which is a convex subset of X.

[5] A. N. Kolmogorov and S. V. Fomin. *Introductory real analysis.* Dover Publications, Inc., New York, 1975, E. Zeidler. *Nonlinear functional analysis and its applications. I. Fixed-point theorems.* Springer-Verlag, New York, 1986.

Next, let $x_{\ominus 0}$, $x_{\ominus 1}$, $x_{\oplus 0}, x_{\oplus 1} \in D$ be such elements that

$$x_{\ominus 0} \le x_{\ominus 1} \le x_{\oplus 1} \le x_{\oplus 0},$$
$$x_{\ominus 1} = \mathfrak{T}_{\oplus} x_{\ominus 0} + \mathfrak{T}_{\ominus} x_{\oplus 0} + f,$$
$$x_{\oplus 1} = \mathfrak{T}_{\oplus} x_{\oplus 0} + \mathfrak{T}_{\ominus} x_{\ominus 0} + f,$$

Then, we observe that

$$x_{\ominus 2} = \mathfrak{T}_{\oplus} x_{\ominus 1} + \mathfrak{T}_{\ominus} x_{\oplus 1} + f \ge \mathfrak{T}_{\oplus} x_{\ominus 0} + \mathfrak{T}_{\ominus} x_{\oplus 0} + f = x_{\ominus 1}$$
$$x_{\oplus 2} = \mathfrak{T}_{\oplus} x_{\oplus 1} + \mathfrak{T}_{\ominus} x_{\ominus 1} + f \le \mathfrak{T}_{\oplus} x_{\oplus 0} + \mathfrak{T}_{\ominus} x_{\ominus 0} + f = x_{\oplus 1}.$$

By continuing the iterations we obtain elements such that

$$x_{\ominus k} \le x_{\ominus (k+1)} \le x_{\oplus (k+1)} \le x_{\oplus k}.$$

Then $x \to \mathfrak{T} x + f$ maps D to itself. If D is compact, then by the Schauder fixed point theorem $x_{\odot} \in D$ exists. Moreover, it is bounded from below and above by the sequences $\{x_{\ominus k}\}$ and $\{x_{\oplus k}\}$.

Applications of this method are mainly oriented towards systems of linear simultaneous equations and integral equations[6]. For example, consider a system of linear simultaneous equations

$$x = Ax + f$$

that is supposed to have a unique solution $x_{\odot}$. Assume that

$$A = A_{\oplus} - A_{\ominus}, \qquad A_{\ominus} = \{a^{\ominus}_{ij}\} \in \mathbb{M}^{d\times d},$$
$$A_{\oplus} = \{a^{\oplus}_{ij}\} \in \mathbb{M}^{d\times d}, \quad a^{\ominus}_{ij} \ge 0, \quad a^{\oplus}_{ij} \ge 0.$$

We may partially orderthe space $\mathbb{R}^d$ by saying that $x \le y$ if and only if $x_i \le y_i$ for $i = 1, 2, ...n$. Compute the vectors

$$x_{\ominus (k+1)} = A_{\oplus} x_{\ominus k} + A_{\ominus} x_{\oplus k} + f, \quad x_{\oplus (k+1)} = A_{\oplus} x_{\oplus k} + A_{\ominus} x_{\ominus k} + f.$$

If $x_{\ominus 0} \le x_{\ominus 1} \le x_{\odot} \le x_{\oplus 1} \le x_{\oplus 0}$, then for all the components of $x_{\odot}$ we obtain two–sided estimates

$$x^{(i)}_{\ominus k} \le x^{(i)}_{\ominus (k+1)} \le x^{(i)}_{\odot} \le x^{(i)}_{\oplus (k+1)} \le x^{(i)}_{\oplus k}, \quad i = 1, 2, ...n.$$

[6] L. Collatz. Funktionanalysis und numerische mathematik, Springer-Verlag, Berlin, 1964.

Task 2.5.1 *Apply the above method for finding two–sided bounds of the Euclid error norm and componentwise errors for a system of linear simultaneous equations*

$$Ax = f$$

where

$$a_{ij} = (-1)^{i+j}\kappa/ij \quad \text{if } i \neq j, \quad \kappa = 0.1$$
$$a_{ii} = i.$$

For the ith component of the solution determine the lower and upper bounds as follows:

$$\max_{j=0,1,\dots k+1} \left(x_j^{\ominus}\right)_i \leq (x_{\odot})_i \leq \min_{j=0,1,\dots k+1} \left(x_j^{\oplus}\right)_i .$$

It should be remarked that convergence of $x^{(i)}_{\ominus\, k}$ and $x^{(i)}_{\oplus\, k}$ to $x_{\odot}$ (and the convergence rate) requires a special investigation, which must use specific features of a particular problem.

Remark 2.5.1 *In principle, a posteriori error estimates based on monotonicity can provide the most informative POINTWISE a posteriori error estimates. Regrettably, the respective theory has not been yet properly investigated.*

Chapter 3

A POSTERIORI ERROR INDICATORS FOR FEM

The goal of this chapter is to give an overview of a posteriori error estimation methods developed for Finite Element approximations in 70th–80th.

Chapter plan

- **Mathematical background;**
- **Residual type error estimates;**
  - **Basic idea;**
  - **Estimates in 1D case;**
  - **Estimates in 2D case;**
  - **Comments;**
- **Methods based on post–processing;**
- **Methods using adjoint problems;**

3.1 Sobolev spaces with negative indices

Sobolev spaces with negative indices

Definition 3.1.1 *Linear functionals defined on the functions of the space* $\overset{\circ}{C}{}^{\infty}(\Omega)$ *are called* **distributions**. *They form the space* $\mathcal{D}'(\Omega)$

Value of a **distribution** g on a function ϕ is $\langle g, \phi \rangle$.

Distributions possess an important property: they have derivatives of any order.

Let $g \in \mathcal{D}'(\Omega)$, then the quantity $-\langle \mathbf{g}, \frac{\partial \phi}{\partial x_i} \rangle$ is another linear functional on $\mathcal{D}(\Omega)$. It is viewed as a generalized partial derivative of g taken over the i-th variable.

Derivatives of L^q–functions. Any function g from the space $L^q(\Omega)$ $(q \geq 1)$ defines a certain distribution as

$$\langle \mathbf{g}, \phi \rangle \; = \; \int_{\mathbf{\Omega}} \mathbf{g} \phi \, \mathbf{dx}$$

and, therefore, has generalized derivatives of any order. The sets of distributions, which are derivatives of q-integrable functions, are called Sobolev spaces with negative indices.

Definition 3.1.2 *The space* $W^{-\ell,q}(\Omega)$ *is the space of distributions* $g \in \mathcal{D}'(\Omega)$ *such that*

$$g = \sum_{|\alpha| \le \ell} D^{\alpha} g_{\alpha},$$

where $\mathbf{g}_{\alpha} \in L^{q}(\Omega)$.

Spaces $W^{-1,p}(\Omega)$

$W^{-1,p}(\Omega)$ contains distributions that can be viewed as generalized derivatives of L^{q}-functions.The functional

$$\left\langle \frac{\partial \mathbf{f}}{\partial \mathbf{x_i}}, \phi \right\rangle := - \int_{\mathbf{\Omega}} \mathbf{f} \frac{\partial \phi}{\partial \mathbf{x_i}} \, \mathbf{dx} \qquad \mathbf{f} \in \mathbf{L^q(\Omega)}$$

is linear and continuous not only for $\phi \in \overset{\circ}{C}{}^{\infty}(\Omega)$ but, also, for $\phi \in \overset{\circ}{W}{}^{1,p}(\Omega)$, where $1/p + 1/q = 1$ (density property). Hence, first generalized derivatives of f lie in the space dual to $\overset{\circ}{W}{}^{1,p}(\Omega)$ denoted by $W^{-1,p}(\Omega)$.

For $\overset{\circ}{W}{}^{1,2}(\Omega) = \overset{\circ}{\mathbf{H}}{}^{\mathbf{1}}(\mathbf{\Omega})$, the respective dual space is denoted by $H^{-1}(\Omega)$.

Norms in "negative spaces"

For $\mathbf{g} \in H^{-1}(\Omega)$ we may introduce two equivalent "negative norms".

$$\|g\|_{(-1),\Omega} := \sup_{\phi \in \overset{\circ}{H}{}^1(\Omega)} \frac{|\langle \mathbf{g}, \phi \rangle|}{\|\phi\|_{1,2,\Omega}} < +\infty$$

$$\boldsymbol{|}g\boldsymbol{|} := \sup_{\phi \in \overset{\circ}{H}{}^1(\Omega)} \frac{|\langle \mathbf{g}, \phi \rangle|}{\|\nabla \phi\|_{\Omega}} < +\infty$$

From the definitions, it follows that

$$\langle \mathbf{g}, \phi \rangle \leq \|\mathbf{g}\|_{(-1),\Omega} \|\varphi\|_{1,2,\Omega}$$

$$\boxed{\langle \mathbf{g}, \phi \rangle \leq \boldsymbol{|}\mathbf{g}\boldsymbol{|} \quad \|\nabla \varphi\|_{\Omega}}$$

3.2 Residual method

3.2.1 Errors and Residuals. First glance

If an analyst is not sure in the quality of an approximate solution computed, then the very first idea that comes to his mind is to substitute the approximate solution into the equation and look at the **equation residual**.

We begin by recalling basic relations between residuals and errors that hold for systems of linear simultaneous equations. Let $\mathcal{A} \in \mathbb{M}^{d\times d}$, $\det \mathcal{A} \neq 0$, consider the system

$$\mathcal{A}u + f = 0.$$

For any v we have the simplest *residualtype estimate*

$$\mathcal{A}(v-u) = \mathcal{A}v + f; \quad \Rightarrow \quad \|e\| \leq \|\mathcal{A}^{-1}\|\|r\|.$$

where $e = v - u$ *and* $r = \mathcal{A}v + f$.

Two–sided estimates Define the quantities

$$\lambda_{\min} = \min_{\substack{y\in\mathbb{R}^d \\ y\neq 0}} \frac{\|\mathcal{A}y\|}{\|y\|} \quad \text{and} \quad \lambda_{\max} = \max_{\substack{y\in\mathbb{R}^d \\ y\neq 0}} \frac{\|\mathcal{A}y\|}{\|y\|}$$

Since $\mathcal{A}e = r$, we see that

$$\lambda_{\min} \leq \frac{\|\mathcal{A}e\|}{\|e\|} = \frac{\|r\|}{\|e\|} \leq \lambda_{\max} \;\Rightarrow\; \lambda_{\max}^{-1}\|r\| \leq \|e\| \leq \lambda_{\min}^{-1}\|r\|.$$

Since u is a solution, we have

$$\lambda_{\min} \leq \frac{\|\mathcal{A}u\|}{\|u\|} = \frac{\|f\|}{\|u\|} \leq \lambda_{\max} \;\Rightarrow\; \lambda_{\max}^{-1}\|f\| \leq \|u\| \leq \lambda_{\min}^{-1}\|f\|$$

Thus,

$$\frac{\lambda_{\min}}{\lambda_{\max}}\frac{\|r\|}{\|f\|} \leq \frac{\|e\|}{\|u\|} \leq \frac{\lambda_{\max}}{\lambda_{\min}}\frac{\|r\|}{\|f\|}.$$

Key "residual–error" relation Since

$$\frac{\lambda_{\max}}{\lambda_{\min}} = \mathbf{Cond}\,\mathcal{A},$$

we arrive at the basic relation where the matrix condition number serves as an important factor

$$\boxed{(\mathbf{Cond}\,\mathcal{A})^{-1}\frac{\|r\|}{\|f\|} \leq \frac{\|e\|}{\|u\|} \leq \mathbf{Cond}\,\mathcal{A}\,\frac{\|r\|}{\|f\|}.} \tag{3.1}$$

Thus, the relative error is controlled by the relative value of the residual. However, the bounds deteriorates when the conditional number is large.

In principle, the above consideration can extended to a wider set of linear problems, where

$$\mathcal{A} \in \mathcal{L}(X, Y)$$

is a coercive linear operator acting from a Banach space X to another space Y and f is a given element of Y.

> However, if $\mathcal{A}$ is related to a boundary-value problem, then one should properly define the spaces X and Y and find a practically meaningful analog of the estimate (3.1).

3.2.2 Residual type estimates for elliptic equations

Elliptic equations Let $\mathcal{A} : X \to Y$ be a linear elliptic operator. Consider the boundary-value problem

$$\mathcal{A}u + f = 0 \quad \text{in } \Omega, \qquad u = u_0 \quad \text{on } \partial\Omega.$$

Assume that $v \in X$ is an approximation of u. Then, we should measure the error in X and the residual in Y, so that the principal form of the estimate is

$$\|v - u\|_X \leq C\|\mathcal{A}v + f\|_Y, \tag{3.2}$$

where the constant C is independent of v. The key question is as follows:

Which spaces X and Y should we choose for a particular boundary-value problem ?

Consider the problem

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

with $f \in L^2(\Omega)$. The generalized solution satisfies the relation

$$\int_\Omega \nabla u \cdot \nabla w \, dx = \int_\Omega f w \, dx \quad \forall w \in V_0 := \overset{\circ}{H}{}^1(\Omega),$$

which implies the energy estimate

$$\|\nabla u\|_{2,\Omega} \le C_\Omega \|f\|_{2,\Omega}.$$

Here C_Ω is a constant in the Friederichs-Steklov inequality. Assume that an approximation $v \in V_0$ and $\Delta v \in L^2(\Omega)$. Then,

$$\int_\Omega \nabla(u - v) \cdot \nabla w \, dx = \int_\Omega (f + \Delta v) w \, dx, \quad \forall w \in V_0.$$

Setting $w = u - v$, we obtain the estimate

$$\|\nabla(u - v)\|_{2,\Omega} \leq C_\Omega \|f + \Delta v\|_{2,\Omega}, \tag{3.3}$$

whose right-hand side of (3.3) is formed by the L^2-norm of the residual.

However, usually a sequence of approximations $\{v_k\}$ converges to u only in the energy space, i.e.,

$$\{v_k\} \to u \qquad \text{in } H^1(\Omega),$$

so that $\|\Delta v_k + f\|$ may not converge to zero !

This means that the consistency(the key property of any practically meaningful estimate) is lost.

Which norm of the residual leads to a consistent estimate of the error in the energy norm?

To find it, we should consider Δ not as $H^2 \to L^2$ mapping, but as $H^1 \to H^{-1}$ mapping. For this purpose we use the integral identity

$$\int_\Omega \nabla u \cdot \nabla w \, dx = \langle f, w \rangle, \quad \forall \, w \in V_0 := \overset{\circ}{H}{}^1(\Omega).$$

Here, $\nabla u \in L^2$, so that it has derivatives in H^{-1} and we consider the above as equivalence of two distributions on all trial functions $w \in V_0$.

By $\langle f, w \rangle \le \| f \| \|\nabla w\|_{2,\Omega}$, we obtain another "energy estimate"

$$\|\nabla u\|_{2,\Omega} \le \| f \|.$$

Consistent residual estimate Let $v \in V_0$ be an approximation of u. We have

$$\int_\Omega \nabla(u-v)\cdot\nabla w\,dx \;= \int_\Omega (fw - \nabla v\cdot\nabla w)\,dx = \\ = \langle \Delta v + f, w\rangle, \quad f+\Delta v \in H^{-1}(\Omega).$$

By setting $w = v - u$, we obtain

$$\|\nabla(u-v)\|_{2,\Omega} \le \left| f+\Delta v \right|. \tag{3.4}$$

where

$$\left| f+\Delta v \right| = \sup_{\varphi\in\mathring{H}^1(\Omega)} \frac{|\langle f+\Delta v,\varphi\rangle|}{\|\nabla\varphi\|} = \\ = \sup_{\varphi\in\mathring{H}^1(\Omega)} \frac{|\int_\Omega \nabla(u-v)\cdot\nabla\varphi|}{\|\nabla\varphi\|} \le \sup_{\varphi\in\mathring{H}^1(\Omega)} \frac{\|\nabla(u-v)\|\|\nabla\varphi\|}{\|\nabla\varphi\|} \le \|\nabla(u-v)\|$$

Thus, for the problem considered

$$\|\nabla(u-v)\|_{2,\Omega} = \| f + \Delta v \| \text{ !!!} \tag{3.5}$$

From (3.5), it readily follows that

$$\| f + \Delta v_k \| \to 0 \quad \text{as} \quad \{v_k\} \to u \text{ in } H^1.$$

We observe that the estimate (3.5) is consistent.

Diffusion equation Similar estimates can be derived for

$$\mathcal{A}u + f = 0, \quad \text{in}\,\Omega, \qquad u = 0 \text{ on } \partial\Omega,$$

where

$$\mathcal{A}u = \text{div } A\nabla u := \sum_{i,j=1}^{d} \frac{\partial}{\partial x_i}\left(a_{ij}(x)\frac{\partial u}{\partial x_j}\right),$$
$$a_{ij}(x) = a_{ji}(x) \in L^\infty(\Omega),$$
$$\lambda_{\min}|\eta|^2 \le a_{ij}(x)\eta_i\eta_j \le \lambda_{\max}|\eta|^2, \quad \forall \eta \in \mathbb{R}^d,\ x \in \Omega,$$
$$\lambda_{\max} \ge \lambda_{\min} \ge 0.$$

Let $v \in V_0$ be an approximation of u. Then,

$$\int_\Omega A\nabla(u-v)\cdot\nabla w\,dx = \int_\Omega (fw - A\nabla v\cdot\nabla w)\,dx, \quad \forall w \in V_0.$$

Again, the right-hand side of this relation is a bounded linear functional on V_0, i.e.,

$$f + \text{div }(A\nabla v) \in H^{-1}.$$

Hence, we have the relation

$$\int_\Omega A\nabla(u-v)\cdot\nabla w\,dx = \langle f + \text{div }(A\nabla v), w\rangle, \quad \forall w \in V_0.$$

Setting $w = u - v$, we derive the estimate

$$\|\nabla(u-v)\|_{2,\Omega} \le \lambda_{\min}^{-1} \| f + \text{div }(A\nabla v) \|. \tag{3.6}$$

Next,

$$\| f + \mathrm{div}\ (A\nabla v) \| = \sup_{\varphi \in \overset{\circ}{H}{}^1(\Omega)} \frac{| \langle f + \mathrm{div}\ (A\nabla v), \varphi \rangle |}{\|\nabla \varphi\|_{2,\Omega}} =$$

$$= \sup_{\varphi \in \overset{\circ}{H}{}^1(\Omega)} \frac{| \int_\Omega A\nabla(u-v) \cdot \nabla\varphi\, dx |}{\|\nabla \varphi\|_{2,\Omega}} \le \lambda_{\max} \|\nabla(u-v)\|_{2,\Omega}. \quad (3.7)$$

Combining (3.6) and (3.7) we obtain

$$\boxed{\lambda_{\max}^{-1} \| R(v) \| \le \|\nabla(u-v)\|_{2,\Omega} \le \lambda_{\min}^{-1} \| R(v) \|,} \quad (3.8)$$

where $R(v) = f + \mathrm{div}\ (A\nabla v) \in H^{-1}(\Omega)$. We see that upper and lower bounds of the error can be evaluated in terms of the negative norm of $R(v)$.

Main goal

We observe that to find guaranteed bounds of the error reliable estimates of $|R(v)|$ are required.

In essence, a posteriori error estimates derived in 70-90' for Finite Element Methods (FEM) offer several approaches to the evaluation of $|R(v)|$.

We consider them starting with the so–called **explicit residual method** where such estimates are obtained with help of two key points:

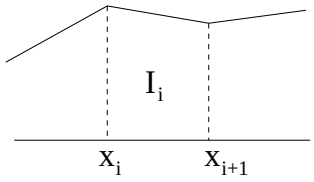
- Galerkin orthogonality property;
- $H^1 \to V_h$ interpolation estimates by Clément.

3.2.3 Explicit residual method in 1D case

Explicit residual method in 1D case Take the simplest model

$$(\alpha u')' + f = 0, \qquad u(0) = u(1).$$

Let $I := (0,1)$, $f \in L^2(I)$, $\alpha(x) \in C(\overline{I}) \geq \alpha_0 > 0$. Divide I into a number of subintervals $I_i = (x_i, x_{i+1})$, where $x_0 = 0$, $x_{N+1} = 1$, and $|x_{i+1} - x_i| = h_i$. Assume that $v \in \overset{\circ}{H}{}^1(I)$ and it is smooth on any interval I_i.



In this case,

$$\| R(v) \| = \sup_{w\in V_0(I),\, w\neq 0} \frac{\int_0^1(-\alpha v'w' + fw)dx}{\|w'\|_{2,I}} =$$

$$= \sup_{w\in \overset{\circ}{H}{}^1(I);w\neq 0} \frac{\sum_{i=0}^N \int_{I_i}(-\alpha v'w' + fw)dx}{\|w'\|_{2,I}} =$$

$$= \sup_{w\in V_0(I),\, w\neq 0} \frac{\sum_{i=0}^N \int_{I_i} r_i(v)w\,dx + \sum_{i=1}^N \alpha(x_i)w(x_i)j(v'(x_i))}{\|w'\|_{2,I}},$$

where $j(\phi(x)) := \phi(x+0) - \phi(x-0)$ is the "jump–function" and $r_i(v) = (\alpha v')' + f$ is the residualon I_i.

For arbitrary v we can hardly get an upper bound for this supremum.

Use Galerkin orthogonality Assume that $v = u_h$, i.e., it is the *Galerkin approximation* obtained on a finite–dimensional subspace V_{0h} formed by piecewise polynomial continuous functions. Since

$$\int_I \alpha u_h' w_h' \, dx - \int_I f w_h \, dx = 0 \quad \forall w_h \in V_{0h}.$$

we may add the left–hand side with any w_h to the numerator what gives

$$\| R(u_h) \| = \sup_{w \in V_0(I)} \frac{\int_0^1 (-\alpha u_h'(w - \pi_h w)' + f(w - \pi_h w)) \, dx}{\|w'\|_{2,I}},$$

where $\pi_h : V_0 \to V_{0h}$ is the interpolation operator defined by the conditions $\pi_h v \in V_{0h}$, $\pi_h v(0) = \pi_h v(1) = 0$ and

$$\pi_h v(x_i) = v(x_i), \quad \forall x_i, \ i = 1, 2, ..., N.$$

Integrating by parts Now, we have

$$\| R(u_h) \| = \sup_{w \in V_0(I)} \left\{ \frac{\sum_{i=0}^{N} \int_{I_i} r_i(u_h)(w - \pi_h w)\, dx}{\|w'\|_{2,I}} + \right.$$
$$\left. + \frac{\sum_{i=1}^{N} \alpha(x_i)(w(x_i) - \pi_h w(x_i)) j(u_h'(x_i))}{\|w'\|_{2,I}} \right\}.$$

Since $w(x_i) - \pi_h w(x_i) = 0$, the second sum vanishes. For first one we have

$$\sum_{i=0}^{N} \int_{I_i} r_i(u_h)(w - \pi_h w)\, dx \le \sum_{i=0}^{N} \|r_i(u_h)\|_{2,I_i} \|w - \pi_h w\|_{2,I_i}.$$

Since for $w \in \overset{\circ}{H}{}^1(I_i)$

$$\|w - \pi_h w\|_{2,I_i} \le c_i \|w'\|_{2,I_i},$$

we obtain for the numerator of the above quotient

$$\sum_{i=0}^{N} \int_{I_i} r_i(u_h)(w - \pi_h w)\, dx \le \sum_{i=0}^{N} c_i \|r_i(u_h)\|_{2,I_i} \|w'\|_{2,I_i} \le$$

$$\le \left(\sum_{i=0}^{N} c_i^2 \|r_i(u_h)\|^2_{2,I_i} \right)^{1/2} \|w'\|_{2,I},$$

which implies the desired upper bound

$$\boxed{| R(u_h) | \le \left(\sum_{i=0}^{N} c_i^2 \|r_i(u_h)\|^2_{2,I_i} \right)^{1/2}.} \tag{3.9}$$

This bound is the sum of local residuals $r_i(u_h)$ with weights given by the interpolation constants c_i.

Interpolation constants For piecewise affine approximations, the interpolation constants c_i are easy to find. Indeed, let γ_i be a constant that satisfies the condition

$$\inf_{w\in\overset{\circ}{H}{}^1(I_i)} \frac{\|w'\|^2_{2,I_i}}{\|w-\pi_h w\|^2_{2,I_i}} \geq \gamma_i.$$

Then, for all $w \in \overset{\circ}{H}{}^1(I_i)$, we have

$$\|w-\pi_h w\|_{2,I_i} \leq \gamma_i^{-1/2}\|w'\|_{2,I_i}$$

and one can set $c_i = \gamma_{I_i}^{-1/2}$.

Let us estimate γ_{I_i}.

Note that

$$\int_{x_i}^{x_{i+1}} |w'|^2\,dx = \int_{x_i}^{x_{i+1}} |(w-\pi_h w)' + (\pi_h w)'|^2\,dx,$$

where $(\pi_h w)'$ is constant on (x_i, x_{i+1}). Therefore,

$$\int_{x_i}^{x_{i+1}} (w-\pi_h w)'(\pi_h w)'\,dx = 0$$

and

$$\int_{x_i}^{x_{i+1}} |w'|^2\,dx = \int_{x_i}^{x_{i+1}} |(w-\pi_h w)'|^2\,dx + \int_{x_i}^{x_{i+1}} |(\pi_h w)'|^2\,dx \geq$$
$$\geq \int_{x_i}^{x_{i+1}} |(w-\pi_h w)'|^2\,dx.$$

Interpolation constants in 1D problem Thus, we have

$$\inf_{w\in\overset{\circ}{H}{}^1(I_i)} \frac{\int_{x_i}^{x_{i+1}} |w'|^2\,dx}{\int_{x_i}^{x_{i+1}} |w-\pi_h w|^2\,dx} \geq \inf_{w\in\overset{\circ}{H}{}^1(I_i)} \frac{\int_{x_i}^{x_{i+1}} |(w-\pi_h w)'|^2\,dx}{\int_{x_i}^{x_{i+1}} |w-\pi_h w|^2\,dx} \geq$$

$$\geq \inf_{\eta\in\overset{\circ}{H}{}^1(I_i)} \frac{\int_{x_i}^{x_{i+1}} |\eta'|^2\,dx}{\int_{x_i}^{x_{i+1}} |\eta|^2\,dx} = \frac{\pi^2}{h_i^2},$$

so that $\gamma_i = \pi^2/h_i^2$ and $c_i = h_i/\pi$.

Remark. To prove the very last relation we note that

$$\inf_{\eta\in\overset{\circ}{H}{}^1((0,h))} \frac{\int_0^h |\eta'|^2\,dx}{\int_0^h |\eta|^2\,dx} = \frac{\pi^2}{h^2}$$

is attained on the eigenfunction $sin\frac{\pi}{h}x$, of the problem $\phi'' + \lambda\phi = 0$ on $(0, h)$.

Task 3.2.1 *Solve a boundary–value problem*

$$(\alpha v')' = f,$$
$$v(0) = a, \quad v(1) = b$$

with certain $\alpha(x) > 0$, f, a, *and* b *by the finite element method with uniform elements (i.e.,* $h = 1/N$*). Apply the residual method and compare the errors computed with the true error distribution.*

3.2.4 Explicit residual method in 2D case

Residual method in 2D case Let Ω be represented as a union $\mathcal{T}_h$ of simplexes T_i. For the sake of simplicity, assume that $\overline{\Omega} = \cup_{i=1}^{N} \overline{T}_i$ and V_{0h} consists of piecewise affine continuous functions. Then the Galerkin approximation u_h satisfies the relation

$$\int_\Omega A\nabla u_h \cdot \nabla w_h \, dx = \int_\Omega f w_h \, dx, \quad \forall w_h \in V_{0h},$$

where

$$V_{0h} = \{w_h \in V_0 \mid w_h \in P^1(T_i),\ T_i \in \mathcal{F}_h\}.$$

In this case, negative norm of the residual is

$$\| R(u_h) \| = \sup_{w \in V_0} \frac{\int_\Omega (fw - A\nabla u_h \cdot \nabla w)\, dx}{\|\nabla w\|_{2,\Omega}}.$$

Let $pi : \overset{\circ}{H}{}^1 \to V_{0h}$ be a continuous interpolation operator. Then, for the **Galerkin approximation**

$$\| R(u_h) \| = \sup_{w \in V_0} \frac{\int_\Omega (f(w - \pi_h w) - A\nabla u_h \cdot \nabla (w - \pi_h w))\, dx}{\|\nabla w\|_{2,\Omega}}.$$

For finite element approximations such a type projection operators has been constructed. One of the most known was suggested by Ph. Clément[1] and is often called the *Clement's interpolation operator*. Its properties play an important role in the a posteriori error estimation method considered.

[1] Clément, Ph. Approximation by finite element functions using local regularization. (English) Revue Franc. Automat. Inform. Rech. Operat. 9, R-2, 77-84 (1975).

Clement's Interpolation operator

Let E_{ij} denote the common edge of the simplexes T_i and T_j. If s is an inner node of the triangulation $\mathcal{F}_h$, then ω_s denotes the set of all simplexes having this node.

For any s, we find a polynomial $p_s(x) \in P^1(\omega_s)$ such that

$$\int_{\omega_s} (v - p_s) q \, dx = 0 \quad \forall q \in P^1(\omega_s).$$

Now, the interpolation operator π_h is defined by setting

$$\pi_h v(x_s) = p(x_s), \quad \forall x_s \in \Omega,$$
$$\pi_h v(x_s) = 0, \quad \forall x_s \in \partial\Omega.$$

It is a linear and continuous mapping of $\overset{\circ}{H}{}^1(\Omega)$ to the space of piecewise affine continuous functions.

Interpolation estimates in 2D

Moreover, it is subject to the relations

$$\|v - \pi_h v\|_{2,T_i} \leq c_i^T \operatorname{diam}(T_i) \|v\|_{1,2,\omega_N(T_i)}, \tag{3.10}$$

$$\|v - \pi_h v\|_{2,E_{ij}} \leq c_{ij}^E |E_{ij}|^{1/2} \|v\|_{1,2,\omega_E(T_i)}, \tag{3.11}$$

where $\omega_N(T_i)$ is the union of all simplexes having at least *one common node* with T_i and $\omega_E(T_i)$ is the union of all simplexes having *a common edge* with T_i.

Interpolation constants c_i^T and c_{ij}^E are LOCAL and depend on the shape of patches $\omega_N(T_i)$ and $\omega_E(T_i)$.

Quotient relations for the constants

Evaluation of c_i^T and c_{ij}^E requires finding *exact lower bounds* of the following variational problems:

$$\gamma_i^T := \inf_{w \in V_0} \frac{\|w\|_{1,2,\omega_N(T_i)}}{\|w - \pi_h w\|_{2,T_i}} \operatorname{diam}(T_i)$$

and

$$\gamma_{ij}^E := \inf_{w \in V_0} \frac{\|w\|_{1,2,\omega_E(T_i)}}{\|w - \pi_h w\|_{2,E_{ij}}} |E_{ij}|^{1/2}.$$

Certainly, we can replace V_0 be $H^1(\omega_N(T_i))$ and $H^1(\omega_E(T_i))$, respectively, but, anyway finding the constants amounts solving functional eigenvalue type problems !

Let $\sigma_h = A\nabla u_h$. Then,

$$| R(u_h) | = \sup_{w\in V_0} \frac{\int_\Omega (f(w-\pi_h w) - \sigma_h \cdot \nabla(w-\pi_h w))\, dx}{\|\nabla w\|_{2,\Omega}}.$$

If ν_{ij} is the unit outward normal to E_{ij}, then

$$\int_{T_i} \sigma_h \cdot \nabla(w-\pi_h w)\, dx = \\ = \sum_{E_{ij}\subset\partial T_i} \int_{E_{ij}} (\sigma_h\cdot\nu)(w-\pi_h w) ds - \int_{T_i} \operatorname{div}\, \sigma_h (w-\pi_h w)\, dx,$$

Since on the boundary $w-\pi_h w = 0$, we obtain

$$| R(u_h) | = \sup_{w\in V_0} \left\{ \frac{\sum_{i=1}^N \int_{T_i} (\operatorname{div}\, \sigma_h + f)(w-\pi_h w)\, dx}{\|\nabla w\|_{2,\Omega}} + \right. \\ \left. + \frac{\sum_{i=1}^N \sum_{j>i}^N \int_{E_{ij}} j(\sigma_h\cdot\nu_{ij})(w-\pi_h w)\, ds}{\|\nabla w\|_{2,\Omega}} \right\}.$$

First term in sup

$$\int_{T_i}(\operatorname{div}\sigma_h + f)(w - \pi_h w)dx \le \|\operatorname{div}\sigma_h + f\|_{2,T_i}\|w - \pi_h w\|_{2,T_i}$$
$$\le c_i^T \|\operatorname{div}\sigma_h + f\|_{2,T_i}\mathbf{diam}\,(T_i)\|w\|_{1,2,\omega_N(T_i)},$$

Then, the first sum is estimated as follows:

$$\sum_{i=1}^{N}\int_{T_i}(\operatorname{div}\,\sigma_h + f)(w - \pi_h w)dx \le$$
$$\le d_1\left(\sum_{i=1}^{N}\left(c_i^T\right)^2\mathbf{diam}\,(T_i)^2\|\operatorname{div}\,\sigma_h + f\|^2_{2,T_i}\right)^{1/2}\|w\|_{1,2,\Omega},$$

where the constant d_1 depends on the maximal number of elements in the set $\omega_N(T_i)$.

Second term in sup For the second one, we have

$$\sum_{i=1}^{N}\sum_{j>i}^{N}\int_{E_{ij}} j(\sigma_h\cdot\nu_{ij})(w-\pi_h w)\,dx \le$$
$$\le \sum_{i=1}^{N}\sum_{j>i}^{N} \|j(\sigma_h\cdot\nu_{ij})\|_{2,E_{ij}}\; c_{ij}^{E}\;|E_{ij}|^{1/2}\,\|w\|_{1,2,\omega_E(T_i)} \le$$
$$\le d_2\Bigg(\sum_{i=1}^{N}\sum_{j>i}^{N}\left(c_{ij}^{E}\right)^2|E_{ij}|\|j(\sigma_h\cdot\nu_{ij})\|^2_{2,E_{ij}}\Bigg)^{1/2}\|w\|_{1,2,\Omega},$$

where d_2 depends on the maximal number of elements in the set $\omega_E(T_i)$.

Residual type error estimate

By the above estimates we obtain

$$\| R(u_h) \| \le C_0 \Bigg(\Bigg(\sum_{i=1}^{N} \left(c_i^T\right)^2 \operatorname{diam}(T_i)^2 \|\operatorname{div}\, \sigma_h + f\|^2_{2,T_i} \Bigg)^{1/2} + \\ + \Bigg(\sum_{i=1}^{N} \sum_{j>i}^{N} \left(c_{ij}^E\right)^2 |E_{ij}| \, \|j(\sigma_h \cdot \nu_{ij})\|^2_{2,E_{ij}} \Bigg)^{1/2} \Bigg). \quad (3.12)$$

Here $C_0 = C_0(d_1, d_2)$. We observe that the right-hand side is the sum of local quantities (usually denoted by $eta(T_i)$) multiplied by constants depending on properties of the chosen splitting $\mathcal{F}_h$.

Error indicator for quasi-uniform meshes For quasi–uniform meshes all generic constants c_i^T have approximately the same value and can be replaced by a single constant c_1. If the constants c_{ij}^E are also estimated by a single constant c_2, then we have

$$| R(u_h) | \le C \left(\sum_{i=1}^{N} \eta^2(T_i) \right)^{1/2}, \tag{3.13}$$

where $C = C(c_1, c_2, C_0)$ and

$$\eta^2(T_i) = c_1^2 \mathrm{diam}\,(T_i)^2 \| \mathrm{div}\ \sigma_h + f \|^2_{2,T_i} + \frac{c_2^2}{2} \sum_{E_{ij} \subset \partial T_i} |E_{ij}| \| j(\sigma_h \cdot \nu_{ij}) \|^2_{2,E_{ij}}.$$

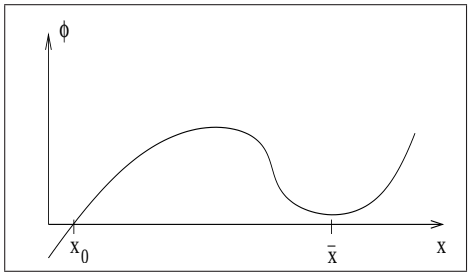
The multiplier $1/2$ arises, because any interior edge is common for two elements.

Comment 1 General form of the residual type a posteriori error estimates is as follows:

$$\|u - u_h\| \leq M(u_k, c_1, c_2, ...c_N, \mathcal{D}),$$

where $\mathcal{D}$ is the data set, u_h is the Galerkin approximation, and $c_i, i = 1, 2, ...N$ are the interpolation constants. The constants depend on the mesh and properties of the special type interpolation operator. The number N depends on the dimension of V_h and may be rather large. If the constants are not sharply defined, then this functional is not more than a certain error indicator. However, in many cases it successfully works and was used in numerous researches.

Comment 2 It is worth noting that for nonlinear problems the dependence between the error and the respective residual is much more complicated. A simple example below shows that the value of the residual may fail to control the distance to the exact solution.



References It is commonly accepted that this approach brings its origin from the papers

I. Babuska and W. C. Rheinboldt. A-posteriori error estimates for the finite element method. Internat. J. Numer. Meth. Engrg., 12(1978).
I. Babuska and W. C. Rheinboldt. Error estimates for adaptive finite element computations. SIAM J.Numer. Anal., 15(1978).

Detailed mathematical analysis of this error estimation method can be found in

R. Verfürth. *A review of a posteriori error estimation and adaptive mesh-refinement techniques* Wiley and Sons, Teubner, New-York, 1996.
M. Ainsworth and T. Oden. *A posteriori error estimation in finite element analysis*, Wiley and Sons, New York, 2000.
K. Eriksson, D. Estep, P. Hansbo, and C. Johnson, *Computational differential equations*, Cambridge University Press, Cambridge, 1996
I. Babuska and T. Strouboulis, *The finite element method and its reliability*, Oxford University Press, New York, 2001.

3.3 A posteriori error indicators based on post–processing of computed solutions

Post–processingof approximate solutions is a numerical procedure intended to modify already computed solution in such a way that the post–processed function would fit some a priori known propertiesmuch better than the original one.

3.3.1 Preliminaries

Let e denotes the *error* of an approximate solution $v \in V$ and $\mathcal{E}(v) : V \to R_+$ denotes the value of an *error estimator* computed on v.

Definition 3.3.1 The estimator is said to be *equivalent to the error* for the approximations v from a certain subset $\widetilde{V}$ if

$$c_1\mathcal{E}(v) \leq \|e\| \leq c_2\mathcal{E}(v) \qquad \forall v \in \widetilde{V}$$

Definition 3.3.2 *The ratio*

$$i_{eff} := 1 + \frac{\mathcal{E}(v) - \|e\|}{\|e\|}$$

is called the effectivity indexof the estimator $\mathcal{E}$.

Ideal estimator has $i_{eff} = 1$. However, in real life situations it is hardly possible, so that values i_{eff} in the diapason from 1 to 2-3 are considered as quite good.

In FEM methods with mesh size h one other term is often used:

Definition 3.3.3 *The estimator $\mathcal{E}$ is called asymptotically equivalent to the error if for a sequence of approximate solutions $\{u_h\}$ obtained on consequently refined meshes there holds the relation*

$$\inf_{h\to 0} \frac{\mathcal{E}(u_h)}{\|u - u_h\|} = 1$$

It is clear that an estimator may be asymptotically exact for one sequence of approximate solutions (e.g. computed on regular meshes) and not exact for another one.

General outlook. Typically, the function $\mathrm{T}u_h$ (where T is a certain linear operator, e.g., ∇) lies in a space U that is wider than the space $\bar{U}$ that contains $\mathrm{T}u$. If we have a computationally inexpensive continuous mapping $\mathbf{G}$ such that $\mathbf{G}(\mathrm{T}v_h) \in \overline{U},\ \forall v_h \in V_h$. then, probably, the function $\mathbf{G}(\mathrm{T}u_h)$ is much closer to $\mathrm{T}u$ than $\mathrm{T}u_h$.

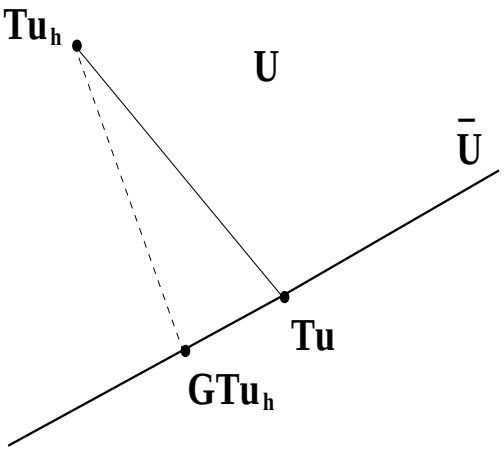


Figure 3.1: Mapping to the set $\overline{U}$

These arguments form the basis of various *post-processing algorithms* that change a computed solution in accordance with some a priori knowledge of properties of the exact solution.

If the error caused by violations of a priori known properties is dominant and the post-processing operator $\mathbf{G}$ is properly constructed, then

$$\|\mathbf{G}\mathrm{T}u_h - \mathrm{T}u\| << \|\mathrm{T}u_h - \mathrm{T}u\|\,.$$

In this case, the explicitly computable norm $\|\mathbf{G}\mathrm{T}u_h - \mathrm{T}u_h\|$ can be used to evaluate upper and lower bounds of the error.

Indeed, assume that there is a positive number $\alpha < 1$ such that

$$\|\mathbf{G}\mathrm{T}u_h - \mathrm{T}u\| \ \leq\ \alpha\,\|\mathrm{T}u_h - \mathrm{T}u\|\,.$$

Then, for $e = u_h - u$ we have

$$\begin{aligned}(1-\alpha)\,\|\mathrm{T}e\| = (1-\alpha)\,\|\mathrm{T}u_h - \mathrm{T}u\| \leq \\ \leq \|\mathrm{T}u_h - \mathrm{T}u\| - \|\mathbf{G}\mathrm{T}u_h - \mathrm{T}u\| \leq \\ \leq \|G\mathrm{T}u_h - \mathrm{T}u_h\| \leq \\ \leq \|\mathbf{G}\mathrm{T}u_h - \mathrm{T}u\| + \|\mathrm{T}u_h - \mathrm{T}u\| \leq \\ \leq (1+\alpha)\,\|\mathrm{T}u_h - \mathrm{T}u\| = (1+\alpha)\,\|\mathrm{T}e\|\,.\end{aligned}$$

Thus, if $\alpha << 1$, then

$$\|\mathrm{T}u_h - \mathrm{T}u\| \simeq \|\mathbf{G}\mathrm{T}u_h - \mathrm{T}u_h\|\,.$$

and the right-hand can be used as an *error indicator*.

3.3.2 Post-processing by averaging

Post-processing operators are often constructed by averaging $\mathrm{T}u_h$ on finite element patches or on the entire domain.

Integral averaging on patches

If $\mathrm{T}u_h \in L^2$, then post-processing operators are obtained by various averaging procedures. Let Ω_i be a *patch* of M_i elements, i.e.,

$$\overline{\Omega_i} = \bigcup T_{ij}, \quad j = 1, 2, ...M_i.$$

Let $P^k(\Omega_i, \mathbb{R}^d)$ be a subspace of $\overline{U}$ that consists of vector-valued polynomial functions of degrees less than or equal to k. Define $g_i \in P^k(\Omega_i, \mathbb{R}^d)$ as the minimizer of the problem:

$$\inf_{g \in P^k(\Omega_i, \mathbb{R}^d)} \int_{\Omega_i} |g - \mathrm{T}u_h|^2\, dx.$$

The minimizer g_i is used to define the values of an averaged function at some points (nodes). Further, these values are utilized by an extension procedure that defines an averaged function

$$\mathbf{G}\mathrm{T}u_h : \Omega \to \mathbb{R}.$$

Consider the simplest case. Let T be the operator ∇ and u_h be a piecewise affine continuous function. Then,

$$\nabla u_h \in P^0(T_{ij}, \mathbb{R}^d) \quad \text{on each } T_{ij} \subset \Omega_i.$$

We denote the values of ∇u_h on T_{ij} by $(\nabla u_h)_{ij}$.

Set $k = 0$ and find $g_i \in P^0$ such that

$$\int\limits_{\Omega_i} |g_i - \nabla u_h|^2 \, dx = \inf_{g \in P^0(\Omega_i)} \int\limits_{\Omega_i} |g - \nabla u_h|^2 \, dx =$$

$$= \inf_{g \in P^0(\Omega_i)} \left\{ |g|^2 |\Omega_i| - 2g \cdot \sum_{j=1}^{M_i} (\nabla u_h)_{ij} |T_{ij}| + \sum_{j=1}^{M_i} |(\nabla u_h)_{ij}|^2 |T_{ij}| \right\}.$$

It is easy to see that g_i is given by a weighted sum of $(\nabla u_h)_{ij}$, namely,

$$g_i = \sum_{j=1}^{M_i} \frac{|T_{ij}|}{|\Omega_i|} (\nabla u_h)_{ij}.$$

Set

$$\mathbf{G}(\nabla u_h)(x_i) = g_i.$$

Repeat this procedure for all nodes and define the vector-valued function $\mathbf{G}\nabla(u_h)$ by the piecewise affine prolongation of these values. For regular meshes with equal $|T_{ij}|$, we have

$$g_i = \sum_{j=1}^{M_i} \frac{1}{M_i} (\nabla u_h)_{ij}.$$

Various averaging formulas of this type are represented in the form

$$g_i = \sum_{j=1}^{M_i} \lambda_{ij} (\nabla u_h)_{ij}, \quad \sum_{j=1}^{M_i} \lambda_{ij} = 1,$$

where λ_{ij} are the weight factors. For internal nodes, they may be taken, e.g., as follows

$$\lambda_{ij} = \frac{|\gamma_{ij}|}{2\pi}, \quad |\gamma_{ij}| \text{ is the angle.}$$

However, if a node belongs to the boundary, then it is better to choose special weights. Their values depend on the mesh and on the type of the boundary[2]

[2]see I. Hlavácek and M. Krizek. On a superconvergence finite element scheme for elliptic systems. I. Dirichlet boundary conditions. *Aplikace Matematiky*, 32(1987), No.2, 131-154.

Discrete averaging on patches Consider the problem

$$\inf_{g\in\mathbb{P}^k(\Omega_i)} \sum_{s=1}^{m_i} \left|g(x_s) - \mathrm{T}u_h(x_s)\right|^2,$$

where the points x_s are specially selected in Ω_i.

Usually, the points x_s are the so–called *superconvergent points*.

Let $g_i \in \mathbb{P}^k(\Omega_i)$ be the minimizer of this problem.

If $k = 0$, and $\mathrm{T} = \nabla$ then

$$g_i = \frac{1}{m_i} \sum_{s=1}^{m_i} \nabla u_h(x_s).$$

Global averaging

Global averaging makes the post–processing not on patches, but on the whole domain.

Assume that $\mathrm{T}u_h \in L^2$ and find $\bar{g}_h \in V_h(\Omega) \subset \overline{U}$ such that

$$\|\bar{g}_h - \mathrm{T}u_h\|^2_\Omega = \inf_{g_h \in V_h(\Omega)} \|g_h - \mathrm{T}u_h\|^2_\Omega \,.$$

The function $\bar{g}_h$ can be viewed as $\mathbf{G}\mathrm{T}u_h$. Very often $\bar{g}_h$ is a better image of $\mathrm{T}u$ than the functions obtained by local procedures.

Moreover, mathematical justifications of the methods based on global averaging procedures can be performed under weaker assumptions what makes them applicable to a wider class of problems[3]

Task 3.3.1 *Solve the boundary–value problem*

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

by h-version FEM (use Matlab or another code). Apply the simplest gradient–averaging error indicator to indicate the error distribution. Compare it with the distribution of true error (the latter can be extracted from a solution on a much finer mesh).

[3]see, e.g., Carstensen, C.; Bartels, S. Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids. I: Low order conforming, nonconforming, and mixed FEM, *Math. Comp.*, 71(2002)

3.3.3 Superconvergence

Justifications of the gradient averaging method. Let u_h be a Galerkin approximation of u computed on V_h. For piecewise affine approximations of the diffusion problem we have the estimate

$$\|\nabla(u - u_h)\|_{2,\Omega} \leq c_1 h, \quad \|u - u_h\|_{2,\Omega} \leq c_2 h^2$$

However, it was discovered[4] that in certain cases this rate may be higher.

For example it may happen that

$$|u(x_s) - u_h(x_s)| \leq Ch^{2+\sigma} \qquad \sigma > 0$$

at a superconvergent point x_s.

Certainly, existence and location of superconvergent points strongly depends on the structure of $\mathcal{T}_h$.

Superconvergence in terms of integral type norms. For example, approximate solutions of the problem

$$\Delta u + f = 0 \quad \text{in} \quad \Omega$$

are said to be supercomverging and an operator $\mathbf{G}$ possesses a *superconvergence* property in $\omega \subset \Omega$ if

$$\|\nabla u - \mathbf{G}\nabla u_h\|_{2,\omega} \leq c_2 h^{1+\sigma},$$

where the constant c_2 may depend on higher norms of u and the structure of $\mathcal{T}_h$.

[4] see, e.g., L. A. Oganesjan and L. A. Ruchovec. *Z. Vyčisl. Mat. i Mat. Fiz.*,9(1969);
M. Zlámal. Lecture Notes. Springer, 1977;
L. B. Wahlbin. Lecture Notes. Springer, 1969.

By exploiting the superconvergence properties, e.g.,

$$\|\nabla u - \mathbf{G}\nabla u_h\|_{2,\omega} \leq c_2 h^{1+\sigma},$$

while

$$\|\nabla u - \nabla u_h\|_{2,\omega} \leq c_2 h,$$

one can usually construct a simple post-processing operator $\mathbf{G}$ satisfying the condition

$$\|\mathbf{G}\nabla u_h - \nabla u\| \leq \alpha \|\nabla u_h - \nabla u\|.$$

where the value of α decreases as h tends to zero.

Since

$$\|\mathbf{G}\nabla u_h - \nabla u_h\| \leq \|\nabla u_h - \nabla u\| + \|\mathbf{G}\nabla u_h - \nabla u\|,$$
$$\|\mathbf{G}\nabla u_h - \nabla u_h\| \geq \|\nabla u_h - \nabla u\| - \|\mathbf{G}\nabla u_h - \nabla u\|.$$

where the first term in the right–hand side is of the order h and the second one is of $h^{1+\delta}$. We see that

$$\|\mathbf{G}\nabla u_h - \nabla u_h\| \sim h$$

Therefore, we observe that in the decomposition

$$\|\nabla(u_h - u)\| \leq \|\nabla u_h - \mathbf{G}\nabla u_h\| + \|\mathbf{G}\nabla u_h - \nabla u\|$$

asymptotically dominates the second directly computable term.

Thus, we obtain a simple error indicator:

$$\|\nabla(u_h - u)\| \approx \|\nabla u_h - \mathbf{G}\nabla u_h\|.$$

Note that

$$i_{eff} = \frac{\|\nabla(u_h - u)\|}{\|\nabla u_h - \mathbf{G}\nabla u_h\|} \approx 1 + ch^{\delta}$$

so that this error indicator is *asymptotically exact* provided that u_h is a Galerkin approximation, u is sufficiently regular and h is small enough.

Such type error indicators (often called ZZ indicatorsby the names of Zienkiewicz and Zhu) are widely used as cheap error indicators in engineering computations [5]

[5]see, e.g., M. Ainsworth, J. Z. Zhu, A. W. Craig and O. C. Zienkiewicz. Analysis of the Zienkiewicz-Zhu a posteriori

3.3.4 Post-processing by equilibration

For a solution of the diffusion problem we know that

$$\operatorname{div} \sigma + f = 0,$$

where $\sigma = A\nabla u$. This suggests an idea to construct an operator $\mathbf{G}$ such that

$$\operatorname{div} (\mathbf{G}(A\nabla u_h)) + f = 0.$$

If $\mathbf{G}$ possesses additional properties (linearity, boundedness), then we may hope that the function $\mathbf{G}A\nabla u_h$ is closer to sig than $A\nabla u_h$ and use the quantity $\|A\nabla u_h - \mathbf{G}A\nabla u_h\|$ as an error indicator.

---

error estimator in the finite element method, *Int. J. Numer. Methods Engrg.*, 28(1989). I. Babuska and R. Rodriguez. The problem of the selection of an a posteriori error indicator based on smoothing techniques, *Internat. J. Numer. Meth. Engrg.*, 36(1993). O. C. Zienkiewicz and J. Z. Zhu. A simple error estimator and adaptive procedure for practical engineering analysis, *Internat. J. Numer. Meth. Engrg.*, 24(1987)

This idea can be applied to an important class of problems

$$\Lambda^\star \mathrm{T} u + f = 0, \qquad \mathrm{Tu} = \mathcal{A}\Lambda u, \tag{3.14}$$

where $\mathcal{A}$ is a positive definite operator, Λ is a linear continuous operator, and $\Lambda^\star$ is the adjoint operator.

In continuum mechanics, equations of the type (3.14) are referred to as the equilibrium equations. Therefore, it is natural to call an operator $\mathbf{G}$ an *equilibration* operator.

> *If the equilibration has been performed exactly then it is not difficult to get an upper error bound. However, in general, this task is either cannot be fulfilled or lead to complicated and expensive procedures. Known methods are usually end with approximately equilibrated fluxes.*

3.4 A posteriori error estimates constructed with help of adjoint problems

3.4.1 Goal–oriented error estimates

Global error estimates give a general idea on the quality of an approximate solution and stopping criteria. However, often it is useful to estimate the errors in terms of specially selected linear functionals ℓ_s, $s = 1, 2, ...M$, e.g.,

$$< \ell, v - u >= \int_\Omega \varphi_0 \, (v - u) \, dx,$$

where ϕ is a locally supported function. Since

$$| < \ell, u - u_h > | \leq \|\ell\| \|u - u_h\|_V,$$

we can obtain such an estimate throughout the global a posteriori estimate. However, in many cases, such a method will strongly overestimate the quantity.

3.4.2 Adjoint problem

A posteriori estimates of the errors evaluated in terms of linear functionals are derived by attracting the *adjoint* boundary-value problem whose right-hand side is formed by the functional ℓ.

Let us represent this idea in the simplest form. Consider a system

$$Au = f,$$

where A is a positive definite matrix and f is a given vector. Let v be an approximate solution. Define u_ℓ by the relation

$$A^\star u_\ell = \ell,$$

where $A^\star$ is the matrix adjoint to A. Then,

$$\ell \cdot (u - v) = A^\star u_\ell \cdot u - \ell \cdot v = f \cdot u_\ell - \ell \cdot v = (f - Av) \cdot u_\ell$$

Certainly, the above consideration holds in a more general (operator) sense, so that for a pair of operators A and $A^\star$ we have

$$< \ell, u - v > = < f - Av, u_\ell > . \tag{3.15}$$

and find the error with respect to a linear functional by the product of the residual and the exact solution of the adjoint problem:

$$A^\star u_\ell = \ell.$$

Practical application of this principle depends on the ability to find either u_ℓ or its sharp approximation Methods using adjoint problems has been investigated in the works of R. Becker, C. Johnson, R. Rannacher and others [6].

3.4.3 Application to FEM. Dual-weighted residual method

Consider again the diffusion problem. Now, it is convenient to denote the solution of the original problem by u_f, i.e

$$\int_\Omega A\nabla u_f \cdot \nabla w \, dx = \int_\Omega f w \, dx, \quad \forall w \in V_0(\Omega).$$

Since in our case $A = A^\star$, the *adjoint* problem is to find $u_\ell \in V_0(\Omega)$ such that

$$\int_\Omega A\nabla u_\ell \cdot \nabla w \, dx = \int_\Omega \ell w \, dx, \quad \forall w \in V_0(\Omega).$$

Let Ω be divided into a number of elements T_i, $i = 1, 2, ...N$. Given approximations on the elements, we define a finite-dimensional subspace $V_{0h} \in V_0(\Omega)$ and the Galerkin

[6] A more detailed exposition of these works can be found in W. Bangerth and R. Rannacher. *Adaptive finite element methods for differential equations.* Birkhäuser, Berlin, 2003.
R. Becker and R. Rannacher. A feed–back approach to error control in finite element methods: Basic approach and examples, *East–West J. Numer. Math.*, 4(1996), 237-264.
Concerning error estimation in goal–oriented quantities we refer, e.g., to J. T. Oden, S. Prudhomme. Goal-oriented error estimation and adaptivity for the finite element method, *Comput. Math. Appl.*, 41, 735-756, 2001.
P. Neittaanmaki and S. Repin. Reliable Methods for Computer Simulation. Error Control and A Posteriori Estimates. Elsevier. 2004.

approximations u_{fh} and $u_{\ell h}$:

$$\int_\Omega A\nabla u_{fh}\cdot\nabla w_h\,dx=\int_\Omega f w_h dx,\qquad \forall w_h\in V_{0h},$$
$$\int_\Omega A\nabla u_{\ell h}\cdot\nabla w_h\,dx=\int_\Omega \ell w_h dx,\qquad \forall w_h\in V_{0h}.$$

Since

$$\int_\Omega \ell(u_f-u_{fh})dx=\int_\Omega A\nabla u_\ell\cdot\nabla(u_f-u_{fh})dx$$

and

$$\int_\Omega A\nabla u_{\ell h}\cdot\nabla(u_f-u_{fh})dx=0,$$

We arrive at the relation

$$\int_\Omega \ell(u_f-u_{fh})dx=\int_\Omega A\nabla(u_\ell-u_{\ell h})\cdot\nabla(u_f-u_{fh})dx \tag{3.16}$$

whose right-hand side is expressed in the form

$$\sum_{i=1}^N\int_{T_i} A\nabla(u_f-u_{fh})\cdot\nabla(u_\ell-u_{\ell h})\,dx=$$
$$\sum_{i=1}^N\left\{-\int_{T_i}\mathrm{div}\,\left(A\nabla(u_f-u_{fh})\right)(u_\ell-u_{\ell h})\,dx+\right.$$
$$\left.+\frac{1}{2}\int_{\partial T_i} j\left(\nu_i\cdot A\nabla(u_f-u_{fh})\right)(u_\ell-u_{\ell h})\,ds\right\}.$$

This relation implies the estimate

$$\int_\Omega \ell(u_f - u_{fh})dx = \sum_{i=1}^N \Big\{ \|\mathrm{div}\, A\nabla(u_f - u_{fh})\|_{2,T_i} \|u_\ell - u_{\ell h}\|_{2,T_i} +$$
$$+\tfrac{1}{2} \|j(\nu_i \cdot A\nabla(u_f - u_{fh}))\|_{2,\partial T_i} \|u_\ell - u_{\ell h}\|_{2,\partial T_i} \Big\} =$$
$$= \sum_{i=1}^N \Big\{ \|f + \mathrm{div}\, A\nabla u_{fh}\|_{2,T_i} \|u_\ell - u_{\ell h}\|_{2,T_i} +$$
$$+\tfrac{1}{2} \|j(\nu_i \cdot A\nabla u_{fh})\|_{2,\partial T_i} \|u_\ell - u_{\ell h}\|_{2,\partial T_i} \Big\}.$$

Here, the principal terms are the same as in the explicit residual method, but the weights are given by the norms of $u_\ell - u_{\ell h}$.

Assume that $u_\ell \in H^2$ and $u_{\ell h}$ is constructed by piecewise affine continuous approximations. Then the norms

$\|u_\ell - u_{\ell h}\|_{T_i}$and $\|u_\ell - u_{\ell h}\|_{2,\partial T_i}$

are estimated by the quantities $h^\alpha |u_\ell|_{2,2,T_i}$ with $\alpha = 1$ and $1/2$ and the multipliers $\hat{c}_i$ and $\hat{c}_{ij}$, respectively.

In this case, we obtain an estimate with constants defined by the standard

$$H^2 \to V_{0h}$$

interpolation operator whose evaluation is much simpler than that of the constants arising in the

$$H^1 \to V_{0h}$$

interpolation.

This is the advantage of the dual-weighted residual method.

3.4.4 A posteriori estimates in L^2–norm.

In principle, this technology can be exploited to evaluate estimates in L^2–norm. Indeed,

$$\|u_f - u_{fh}\| = \sup_{\boldsymbol{\ell}\in L^2} \frac{(\boldsymbol{\ell}, u_f - u_{fh})}{\|\boldsymbol{\ell}\|} = \sup_{\boldsymbol{\ell}\in L^2} \frac{(A\nabla u_{\boldsymbol{\ell}}, \nabla(u_f - u_{fh}))}{\|\boldsymbol{\ell}\|} =$$

$$= \sup_{\boldsymbol{\ell}\in L^2} \frac{(A\nabla(u_{\boldsymbol{\ell}} - \pi_h(u_{\boldsymbol{\ell}})), \nabla(u_f - u_{fh}))}{\|\boldsymbol{\ell}\|} =$$

$$= \sup_{\boldsymbol{\ell}\in L^2} \frac{(\nabla(u_{\boldsymbol{\ell}} - \pi_h(u_{\boldsymbol{\ell}})), A\nabla(u_f - u_{fh}))}{\|\boldsymbol{\ell}\|} =$$

$$= \sup_{\boldsymbol{\ell}\in L^2} \frac{\sum\limits_{i=1}^{N} \left\{ \int\limits_{T_i} \nabla(u_{\boldsymbol{\ell}} - \pi_h(u_{\boldsymbol{\ell}})), A\nabla(u_f - u_{fh})\, dx \right\}}{\|\boldsymbol{\ell}\|}$$

Integrating by parts, we obtain

$$\frac{\sum_{i=1}^{N}\left\{\|f+\operatorname{div} A\nabla u_{fh}\|_{T_i}\ \|u_{\boldsymbol{\ell}}-\pi_h(u_{\boldsymbol{\ell}})\|_{T_i}+\frac{1}{2}\|j(\nu_i\cdot A\nabla u_{fh})\|_{\partial T_i}\|u_{\boldsymbol{\ell}}-\pi_h(u_{\boldsymbol{\ell}})\|_{\partial T_i}\right\}}{\|\boldsymbol{\ell}\|}$$

If for *any* $\boldsymbol{\ell}\in L^2$ the adjoint problem has a regular solution (e.g., $u_{\boldsymbol{\ell}}\in H^2$), so that we could combine the standard interpolation estimate for the interpolant of $u_{\boldsymbol{\ell}}$ with the regularity estimate for the PDE (e.g., $\|u_{\boldsymbol{\ell}}\|\le C_1\|\boldsymbol{\ell}\|$), then we obtain

$$\|u_{\boldsymbol{\ell}}-\pi_h(u_{\boldsymbol{\ell}})\|_{T_i}\le C_1h^{\alpha_1}\|\boldsymbol{\ell}\|,\qquad \|u_{\boldsymbol{\ell}}-\pi_h(u_{\boldsymbol{\ell}})\|_{\partial T_i}\le C_1h^{\alpha_2}\|\boldsymbol{\ell}\|$$

with certain α_k.

Under the above conditions $\|\boldsymbol{\ell}\|$ is reduced and we arrive at the estimate, in which the element residuals and interelement jumps are weighted with factors $C_1h^{\alpha_1}$ and $C_2h^{\alpha_2}$.

3.4.5 Comment

We end up this lecture with a "terminological" comment. In the literature devoted to a posteriori error analysis one can find often find terms like
"duality approach to a posteriori error estimation" or
"dual-based error estimates".
However, the essence behind such a terminology may be quite different because the word *"duality"* is used in 3 different meanings:
(a) Duality in the sense of functional spaces. We have seen that if for the equation $\mathcal{L}u = f$ errors are measured in the original (energy) norm then a consistent upper bound is given by the residual in the norm of the space **topologically dual** to a subspace of the energy space (e.g., H^{-1}).
(b) Duality in the sense of using the Adjoint Problem.
(c) Duality in the sense of the Theory of the Calculus of Variations.

In the next lecture
we will proceed to the detailed exposition
of the approach (c).

Chapter 4

FUNCTIONAL A POSTERIORI ESTIMATES FOR A MODEL ELLIPTIC PROBLEM

4.1 Introduction

In this chapter, we derive Functional A Posteriori Estimate for the problem

$$\Delta u + f = 0,\ \Omega \qquad u = 0\ \partial\Omega.$$

and discuss their meaning, principal features and practical implementation.

Functional A Posteriori Estimates. Functional A Posteriori Estimate is a *computable majorant* of the difference between exact solution u and any conforming approximation v having the general form:

$$\mathbf{\Phi}(u - v) \leq M\left(\mathcal{D}, v\right) \qquad \forall v \in V\,! \tag{4.1}$$

$\mathcal{D}$ is the data set (coefficients, domain, parameters, etc.),
$\Phi : V \to \mathrm{R}_+$ is a given functional.
M must be computable and continuous in the sense that

$$M\left(\mathcal{D},\, v\right) \to 0, \quad \text{if } v \to u$$

Types of error measure functionals Φ.

- Energy norm $\quad \boldsymbol{\Phi}(u-v) = \|u-v\|_{\Omega}$
- Local norm $\quad \boldsymbol{\Phi}(u-v) = \|u-v\|_{\omega}$
- Goal–oriented quantity $\quad \boldsymbol{\Phi}(u-v) = (\ell, u-v)$

We will see that

Functional a posteriori estimates provide guaranteed bounds for all above-introduced error measures.

Derivation methods.

These estimates are derived by purely functional methods using the analysis of variational problems or integral identities.

Variational method 96'-97'[1] Variational method exploits variational structure of the original problem and Duality Theory in the Calculus of Variations.

Nonvariational method 2000'Derives a posteriori estimates by certain transformations of integral identities [2]).

Let us consider both methods in application to our basic problem

4.2 Deriving functional a posteriori estimates by the variational method

Let u be a (generalized) solution of the problem

$$\Delta u + f = 0, \ \Omega \qquad u = 0 \ \partial\Omega.$$

[1] S. Repin *Mathematics of Computation*, 69(230), pp. 2000, 481-500.
A systematic exposition of the variational approach to deriving Functional a Posteriori Estimates can be found in P. Neittaanmaki and S. Repin. *Reliable methods for computer simulation. Error control and a posteriori estimates.* Elsevier, NY, 2004

[2] Basic idea of the method is presented in S. Repin. *Proc. St.-Petersburg Math. Society*, 2001 pp. 148-179 (in Russian, translated in American Mathematical Translations Series 2, 9(2003)

As we have seen in Lecture 1, this problem is equivalent to the following variational problem:

Problem $\mathcal{P}$. Find $u \in V_0 := \overset{\circ}{H}{}^1(\Omega)$ such that

$$J(u) = \inf_{v \in V_0} J(v),$$

where

$$J(v) = \frac{1}{2}\|\nabla v\|^2 - (f, v).$$

By the reasons that we discussed earlier this problem has a unique solution.

Lagrangian. Note that

$$J(v) = \sup_{y \in Y} L(\nabla v, y), \quad L(\nabla v, y) = \int_\Omega \left(\nabla v \cdot y - \frac{1}{2}|y|^2 - fv \right) dx$$

where $Y = L^2(\Omega, \mathbb{R}^d)$. Indeed, the value of the above supremum cannot exceed the one we obtain if for almost all $x \in \Omega$ solve the pointwise problems

$$\sup_{y(x)} (\nabla v)(x) \cdot y(x) - \frac{1}{2}|y(x)|^2 \qquad x \in \Omega$$

whose upper bound is attained if set $y(x) = (\nabla v)(x)$. Since $\nabla v \in Y$, we observe that the respective maximizer belongs to Y and, therefore

$$\sup_{y \in Y} L(\nabla v, y) = L(\nabla v, \nabla v) = J(v).$$

Minimax Formulations. Then, the original problem comes in the minimaxform:

$$(\mathcal{P}) \qquad \inf_{v \in V_0} \sup_{y \in Y} L(\nabla v, y)$$

If the order of inf and sup is changed, then we arrive at the so-called dual problem

$$(\mathcal{P}^*) \qquad \sup_{y \in Y} \inf_{v \in V_0} L(\nabla v, y)$$

Note that

$$\inf_{v\in V_0}\int_\Omega\left(\nabla v\cdot y-\frac{1}{2}|y|^2-fv\right)dx=-\frac{1}{2}\|y\|^2+\inf_{v\in V_0}\int_\Omega(\nabla v\cdot y-fv)dx=$$

$$=\begin{cases}-\frac{1}{2}\|y\|^2 & \text{if } y\in Q_f:=\{y\in Y\mid divy+f=0\}\\ -\infty & \text{if } y\notin Q_f\end{cases}$$

Dual Problem. Thus, we observe that the dual problem has the form: find $p \in Q_f$ such that

$$-I^*(p) = \sup_{y \in Q_f} -I^*(y)$$

where

$$I^*(q) = \frac{1}{2}\|q\|^2$$

How these two problems are connected?

First, we establish one relation that holds regardless of the structure of the Lagrangian.

Sup Inf and Inf Sup

Lemma 4.2.1 *Let $L(x, y)$ be a functional defined on the elements of two nonempty sets X and Y. Then*

$$\sup_{y \in Y} \inf_{x \in X} L(x, y) \le \inf_{x \in X} \sup_{y \in Y} L(x, y). \tag{4.2}$$

Proof. It is easy to see that

$$L(x, y) \ge \inf_{\xi \in X} L(\xi, y), \quad \forall x \in X,\ y \in Y.$$

Taking the supremum over $y \in Y$, we obtain

$$\sup_{y \in Y} L(x, y) \ge \sup_{y \in Y} \inf_{\xi \in X} L(\xi, y), \quad \forall x \in X.$$

The left-hand side depends on x, while the right-hand side is a number. Thus, we may take infimum over $x \in X$ and obtain the inequality

$$\inf_{x \in X} \sup_{y \in Y} L(x, y) \ge \sup_{y \in Y} \inf_{\xi \in X} L(\xi, y).$$

Therefore, we always have

$$\sup \mathcal{P}^* \le \inf \mathcal{P}$$

Duality relations. However, in our case we have a stronger relation, namely

$$\sup \mathcal{P}^* = \inf \mathcal{P}$$

To prove this fact, we note that

$$\int_\Omega \nabla u \cdot \nabla v \, dx = \int_\Omega f v \, dx \quad \forall v \in V_0.$$

From here, we conclude that $p = \nabla u \in Q_f$ and

$$-I^*(p) = -\frac{1}{2}\|\nabla u\|^2 = \int_\Omega (\frac{1}{2}|\nabla u|^2 - |\nabla u|^2)dx = \int_\Omega (\frac{1}{2}|\nabla u|^2 - fu)dx = J(u).$$

Now, we use Mikhlin's estimate:

$$\frac{1}{2}\|\nabla(u-v)\|^2 \le J(v) - J(u).$$

Since $J(u) = -I^*(p)$, we have

$$\frac{1}{2}\|\nabla(u-v)\|^2 \le J(v) + I^*(p) \le J(v) + I^*(q) \qquad \forall q \in Q_f.$$

Reform this estimate by using the fact that $q \in Q_f$.

$$\begin{aligned} J(v) + I^*(q) &= \frac{1}{2}\|\nabla v\|^2 - (f,v) + \frac{1}{2}\|q\|^2 \\ &= \frac{1}{2}\|\nabla v\|^2 + \frac{1}{2}\|q\|^2 - (\nabla v, q) = \\ &= \frac{1}{2}\|\nabla v - q\|^2 \end{aligned}$$

We find that

$$\|\nabla(v-u)\| \le \|\nabla v - q\| \qquad \forall q \in Q_f.$$

Take arbitrary $y \in L^2(\Omega)$. Then,

$$\|\nabla(v-u)\| \le \|\nabla v - y\| + \inf_{q\in Q_f}\|y-q\|.$$

How to estimate the above infimum?

Various methods give one and the same answer:

$$\inf_{q\in Q_f}\|y-q\| \le \big|\operatorname{div} y + f\big| \qquad y \in L^2(\Omega), \tag{4.3}$$

$$\inf_{q\in Q_f}\|y-q\| \le C_\Omega\|\operatorname{div} y + f\| \qquad y \in H(\Omega, \operatorname{div}), \tag{4.4}$$

Proof. To prove these estimates we consider an auxiliary problem

$$\Delta\eta + f + \operatorname{div} y = 0 \ \ \Omega \quad \eta = 0 \ \ \partial\Omega.$$

$$\int_\Omega \nabla\eta \cdot \nabla w dx = \int_\Omega (fw - y \cdot \nabla w) dx$$

$$\int_\Omega \overbrace{(\nabla\eta + y)}^{\overline{q}} \cdot \nabla w \, dx = \int_\Omega fw \, dx \quad \forall w \in V_0$$

Thus, $\bar{q} \in Q_f$!!!

Since η is a solution of the boundary–value problem with right–hand side div $y + f \in H^{-1}$, we have

$$\|\nabla\eta\| \leq \left|\!\left| \text{div } y + f \right|\!\right|,$$

Then

$$\inf_{q \in Q_f} \|y - q\| \leq \|y - \bar{q}\| = \|\nabla\eta\| \leq \left|\!\left| \text{div } y + f \right|\!\right|.$$

Here

$$\left|\!\left| \text{div } y + f \right|\!\right| = \sup_{w \in V_0} \frac{\int_\Omega (y \cdot \nabla w - fw) dx}{\|\nabla w\|}$$

Let $y \in H(\Omega, \operatorname{div})$. Then we have

$$\left| \operatorname{div} y + f \right| = \sup_{w \in V_0} \frac{\int_\Omega (\operatorname{div} y + f) w dx}{\|\nabla w\|} \leq C_\Omega \|\operatorname{div} y + f\|,$$

where C_Ω is the constant in the Friederichs–Steklov inequality for the domain Ω.

We observe that a "noncomputable" negative norm has been estimated by a "computable" one without an attraction of Galerkin orthogonality and local (mesh–dependent) constants.

Thus, for any $y \in H(\Omega, \operatorname{div})$ we obtain

$$\begin{aligned} \|\nabla(v - u)\| \leq \|\nabla v - y\| + \inf_{q \in Q_f} \|y - q\| \leq \\ \|\nabla v - y\| + C_\Omega \|\operatorname{div} y + f\| := \overline{\mathfrak{M}}_\Delta(v, y). \end{aligned}$$

Above presented *modus operandi* can be viewed as a simplest version of the variational approach to the derivation of Functional Error Majorants.

Error Minorant. A lower bound of the error is given in the theorem below.

Theorem 4.2.1 *For any $v \in V_0$,*

$$\|\nabla(u-v)\|^2 \geq \underline{\mathfrak{M}}\,{}^2_{\Delta}(v,w), \tag{4.5}$$

where

$$\underline{\mathfrak{M}}\,{}^2_{\Delta}(v,w) := 2\mathcal{F}_v(w) - \|\nabla w\|^2,$$

w is an arbitrary function in V_0, and

$$\mathcal{F}_v(w) = \int_\Omega (f\cdot w - \nabla v\cdot\nabla w)\,dx$$

is the residual functional.

Proof. From the relation

$$2(J(v)-J(u)) = \|\nabla(u-v)\|^2,$$

it follows that

$$\|\nabla(u-v)\|^2 \geq 2(J(v)-J(v+w)),$$

where w is an arbitrary function in V_0. Therefore,

$$\|\nabla(u-v)\|^2 \geq \int\limits_\Omega \left(-|\nabla w|^2 - 2\nabla v\cdot\nabla w\right)dx + 2\int_\Omega f\cdot w\,dx,$$

and we arrive at (4.5).

4.3 Deriving functional a posteriori estimates by the non–variational method

For many problems the variational techniques cannot be applied (e.g., because they may have no variational formulation).

It was suggested another method [3], which is *based on certain transformations of integral identities.*

Non–variational method in the simplest case. Let us expose its simplest version adapted to our model problem.

We have

$$\int_\Omega \nabla(u-v)\nabla w dx = \int_\Omega (fw - \nabla v \cdot \nabla w)\, dx$$

In order to get an upper bound of $\|\nabla(u-v)\|$ we use the relation

$$\int_\Omega (\operatorname{div} yw + \nabla w \cdot y)\, dx = 0 \qquad \forall w \in V_0$$

valid for any $y \in H(\Omega, \operatorname{div})$.

We have

$$\begin{aligned}
&\int_\Omega (\nabla v \cdot \nabla w - fw) dx = \\
&\int_\Omega (\nabla v \cdot \nabla w - fw - (\operatorname{div} y\, w + \nabla w \cdot y)) dx = \\
&\int_\Omega ((\nabla v - y) \cdot \nabla w - (f + \operatorname{div} y) w) dx \leq \\
&\|\nabla v - y\| \|\nabla w\| + \|f + \operatorname{div} y\| \|w\| \leq \\
&\leq (\|\nabla v - y\| + C_\Omega \|f + \operatorname{div} y\|) \|\nabla w\|.
\end{aligned}$$

[3] S. Repin. Two-sided estimates for deviation from an exact solution to uniformly elliptic equation. *Trudi St.-Petersburg Math. Society*, 9(2001), translated in *American Mathematical Translations Series 2, 9(2003)*

Later this method was applied to parabolic problems: S.Repin. Estimates of deviation from exact solutions of initial-boundary value problems for the heat equation, *Rend. Mat. Acc. Lincei*, 13(2002).

Set $w = u - v$.

$$\int_\Omega |\nabla(u - v)|^2 dx \le (\|\nabla v - y\| + C_\Omega \|f + \operatorname{div} y\|)\|\nabla(u - v)\|.$$

Thus, we find that

$$\|\nabla(u - v)\| \le \|\nabla v - y\| + C_\Omega \|f + \operatorname{div} y\|.$$

4.4 Properties of functional a posteriori estimates

For the problem

$$\Delta u + f = 0, \qquad u = 0 \,\text{on}\, \partial\Omega$$

we have obtained the estimate

$$\boxed{\|\nabla(\mathbf{u} - \mathbf{v})\| \le \|\nabla \mathbf{v} - \mathbf{y}\| + \mathbf{C_\Omega} \|\operatorname{div} \mathbf{y} + \mathbf{f}\|} \tag{4.6}$$

1. The estimate is valid for any $v \in V_0$ and $y \in H(\Omega, \operatorname{div})$.
2. Two terms in the right–hand side have a clear sense:
they present measures of the errors in two basic relations

$$p = \nabla u, \qquad \operatorname{div} p + f = 0 \quad \text{in}\, \Omega$$

that jointly form the equation.

3. The estimate is sharp. If set $v = 0$ and $y = 0$, we obtain the energy estimate for the generalized solution

$$\|\nabla u\| \le C_\Omega \|f\|.$$

Therefore, no constant less than C_Ω can be stated in the second term.

If set $y = \nabla u$, than the inequality holds as the equality.

Thus, we see that the estimate (4.6) is sharp in the sense that the multipliers of both terms *cannot be taken smaller* and in the set of admissible y there exists a function that makes the inequality hold as equality.

The estimate as a quadratic functional. By means of the algebraic Young's inequality

$$2ab \le \beta a^2 + \frac{1}{\beta} b^2, \qquad \beta > 0$$

we rewrite this estimate in the form

$$\|\nabla(u-v)\|^2 \le (1+\beta)\|\nabla v - y\|^2 + \frac{1+\beta}{\beta} C_\Omega^2 \|\operatorname{div} y + f\|^2 \tag{4.7}$$

For any β the right–hand side is a quadratic functional. This property makes it possible to apply well known methods for the minimization with respect to y.

Denote the right–hand side of (4.7) by $\overline{\mathfrak{M}}$, i.e.,

$$\overline{\mathfrak{M}}(v, y, \beta, C_\Omega, f) := (1+\beta)\|\nabla v - y\|^2 + \frac{1+\beta}{\beta} C_\Omega^2 \|\operatorname{div} y + f\|^2.$$

This functional provides an upper bound for the norm of the deviation of v from u. Therefore, it is natural to call it the Deviation Majorant.

BVP $\Delta u + f = 0$ has another variational formulation

$$\inf_{\substack{v\in V_0,\\ \beta>0,\\ y\in H(\Omega,\operatorname{div}),}} \overline{\mathfrak{M}}(v, y, \beta, C_\Omega, f)$$

- Minimum of this functional is zero;
- it is attained if and only if $v = u$ and $y = A\nabla u$!;
- $\overline{\mathfrak{M}}$ contains only one global constant C_Ω, which is problem independent;

In principle, one can select certain sequences of subspaces $\{V_{hk}\} \in V_0$ and $\{Y_{hk}\} \in H(\Omega, \operatorname{div})$ and minimize the Error Majorant with respect to these subspaces

$$\inf_{\substack{v\in V_{hk},\\ \beta>0,\\ y\in Y_{hk},}} \overline{\mathfrak{M}}(v, y, \beta, C_\Omega, f)$$

If the subspaces are limit dense, then we would obtain a sequence of approximate solutions (v_k, y_k) and the sequence of numbers

$$\gamma_k := \inf_{\beta>0} \overline{\mathfrak{M}}\,(\,v_k, y_k, \beta, C_\Omega, f\,) \to 0$$

4.5 How to use the estimates in practice?

We discuss practical aspects with the paradigm of conforming finite element approximations.

We have 3 basic ways to use the functional error majorant (deviation estimate):
(a) **Direct** (via flux averaging on the mesh $\mathcal{T}_h$);
(b) **One step delay** (via flux averaging on the mesh h_{ref});
(c) **Minimization** (minimization via y).

(a) Use recovered gradients Let $u_h \in V_h$, then

$$p_h := \nabla u_h \in L_2(\Omega, \mathbb{R}^d), \quad p_h \notin H(\Omega, \mathrm{div}\,).$$

Use an averaging operator $G_h : L_2(\Omega, \mathbb{R}^d) \to H(\Omega, \mathrm{div}\,)$ and have a directly computable estimate

$$\|\nabla(u - u_h)\| \le \|\nabla u_h - G_h p_h\| + C_\Omega \,\|\mathrm{div}\, G_h p_h + f\|$$

(b) Use recovered gradients from $\mathcal{T}_{h_{ref}}$ Let $u_1, u_2, ..., u_k, ...$ be a sequence of approximations on meshes $\mathcal{T}_{h_k}$. Compute $p_k := \nabla u_k$, average it by G_k and for u_{k-1} use the estimate

$$\|u - u_{k-1}\| \le \|\nabla u_{k-1} - G_k p_k\| + C_\Omega \,\|\mathrm{div}\, G_k p_k + f\|$$

This estimate gives a quantitative form of the Runge's rule.

(c) Minimize $\overline{\mathfrak{M}}$ with respect to y. Select a certain subspace Y_τ in $H(\Omega, \mathrm{div}\,)$. Generally, Y_τ may be constructed on another mesh $\mathcal{T}_\tau$ and with help of different trial functions. Then

$$\|\nabla(u - u_h)\| \le \inf_{y_h \in Y_h} \{\|\nabla u_h - y_h\| + C_\Omega \,\|\mathrm{div}\, y_h + f\|\}$$

The wider $Y_h \subset H(\Omega, \mathrm{div}\,)$ the sharper is the upper bound.

Quadratic type functional. From the technical point of view it is better to square both parts of the estimate and apply minimization to a quadratic functional, namely

$$\|\nabla(u-u_h)\|^2 \le \inf_{y_h \in Y_h} \Big\{ (1+\beta)\|\nabla u_h - y_h\| + \\ + C_\Omega \left(1+\frac{1}{\beta}\right) \|\mathrm{div}\, y_h + f\|^2 \Big\}$$

Here, the positive parameter β can be also used to minimize the right–hand side.

Before going to more complicated problems where Deviation Majorants are derived by a more sophisticated theory, we observe several simple examples that nevertheless reflect key points of the above method.

Simple 1-D problem.

$$\begin{aligned} (\,\alpha(x)\;\; u'\,)' &= f(x), \\ u(a) = 0, &\quad u(b) = u_b. \end{aligned}$$

It is equivalent to the variational problem

$$J(v) = \int_a^b \left(\frac{1}{2}\alpha(x) \mid v' \mid^2 + f(x)v \right) dx.$$

Assume that the coefficient α belongs to $\in L^\infty$ and bounded from below by a positive constant. Now

$$V_0 + u_0 = \{ v \in H^1(a,b) \;\mid\; v(a) = 0,\, v(b) = u_b \}.$$

Deviation Majorant.

$$\overline{\mathfrak{M}}\,(v,\beta,y) = (1+\beta) \left(\int_a^b \mid \alpha v' - y \mid^2 dx + \frac{C^2_{(a,b)}}{\beta} \int_a^b |y' - f|^2 \right) dx. \qquad (4.8)$$

In this simple model, u can be presented in the form

$$u(x) = \int_a^x \frac{1}{\alpha(t)} \int_a^t f(z)dzdt + \frac{x}{b}\left(u_b - \int_a^b \frac{1}{\alpha(t)} \int_a^t f(z)dzdt \right).$$

what gives an opportunity to verify how error estimation methods work.

Approximations. Let V_h be made of piecewise–P^1 continuous functions on uniform splitting of the interval and consider approximations of the following types:

- Galerkin approximations;
- Approximations very close to Galerkin (sharp);
- Approximations which are "good" but not Galerkin;
- Coarse (rough) approximations.

Our aim is to show that the Deviation Majorant can be effectively used as an error estimation instrument in all the above cases.

Computation of the Majorant. To find a sharp upper bound, we minimize $\overline{\mathfrak{M}}$ with respect to y and β starting from the function $y_0 = G(v')$, where G is a simple averaging operator, e.g, defined by the relations

$$G(v')(x_i) = \frac{1}{2}(v'(x_i - 0) + v'(x_i + 0)),$$

By the quantity

$$\inf_{\beta>0} \overline{\mathfrak{M}}\,(\,v, \beta, y_0),$$

we obtain a coarse upper bound of the error. It is further improved by minimizing $\overline{\mathfrak{M}}$ with respect to y.

Example. Let $\alpha(x) = 1$, $f(x) = c$, $a = 0$, $b = 1$, $u_b = 1$, e.g., we consider the problem

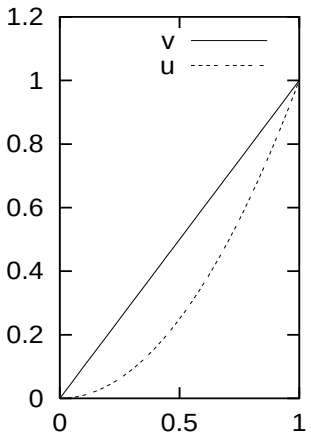
$$u'' = 2, \quad u(0) = 0, \ u(1) = 1.$$

In this case, $C_{(a,b)} = 1/\pi$ and

$$u = \frac{c}{2}\, x^2 \ + \ (1 \ - \frac{c}{2})x, \quad u' = cx + 1 - \frac{c}{2}.$$

Take a rough approximation $v = x$. Then

$$\|(u - v)'\|^2 = \int_0^1 c^2(x - 0.5)^2 dx = c^2/12 \approx 0.083c^2.$$



Exact solution and an approximation.

Various y give different upper bounds. (a) Take $y = v' = 1$, then the first term in

$$\overline{\mathfrak{M}}\,(v,\beta,y) = (1+\beta)\left(\int_0^1 |\,v'-y\,|^2\,dx + \frac{1}{\pi^2\beta}\int_0^1 |y'-f|^2\right)dx.$$

vanishes and we have

$$\overline{\mathfrak{M}} \;\to c^2/\pi^2 \approx 0.101c^2; \text{as } \beta \to +\infty.$$

We see that this upper bound overestimates true error. Note that in this case, all sensible averagings of $v' = 1$ give exactly the same function: $G(1) = 1$! Therefore,

$$G(v') \;-\; v' \;\equiv\; 0$$

and formally ZZ indicator "does not see the error".

For the choice $y = v'$ the Majorant give a certain upper bound of the error (which is not so bad), but the integrand cannot indicate the distribution of local errors. Indeed, we have

$$\overline{\mathfrak{M}} \;= \frac{1}{\pi^2}\int_0^1 c^2 dx.$$

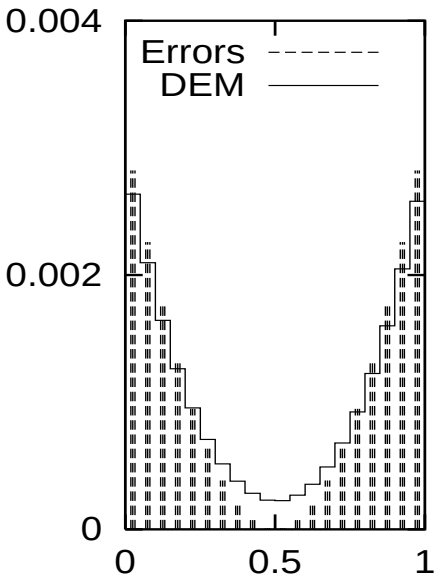
However, the integrand of the Majorant is a constant function, but the error is distributed in accordance with a parabolic law:

$$(u - v)' = c(x - 0.5)^2.$$

(b). Take $y = cx + 1 - c/2$.Then, $y' = c$ and the second term of the majorant vanishes. We have (for $\beta = 0$)

$$\overline{\mathfrak{M}} = \int_0^1 c^2(x - 1/2)^2 dx = c^2/12.$$

We observe that both the global error and the error distribution are exactly reproduced. In real life computations such an "ideal" function y may be unattainable. However, the minimization makes the Majorant close to the sharp value. In this elementary example, we have minimized the Majorant on using piecewise affine approximations of y on 20 subintervals. The elementwise error distribution obtained as the result of this procedure is exposed on the next picture.



True errors and those computed by the Majorant.

To give further illustrations, we consider the functions

$$u_\delta = u + \delta\phi,$$

where δ is a number and ϕ is a certain function (perturbation).

Table 4.1: Errors and two-sided estimates.

δ	e	$2\overline{\mathfrak{M}}$	$2\underline{\mathfrak{M}}$	$\mathrm{i}_{\mathrm{eff}}$	$\mathrm{i}_{\mathrm{esh}}$
0.1	0.019692	0.019743	0.019683	1.003	1.018
0.01	0.001022	0.001025	0.001013	1.003	1.011
0.001	0.000835	0.000839	0.000827	1.005	1.002
0	0.000833	0.000836	0.000825	1.004	1.002

Approximate solutions are piecewise affine continuous interpolants of u_δ defined on a uniform mesh with 20 subintervals.

We take $\phi = x\sin(\pi x)$ and $\delta = 0.1$, 0.01, 0.001, and 0[4].

Task 4.5.1 *Apply the above theory to the problem*

$$(\alpha u')' = f,$$
$$u(0) = 0, \quad u(1) = b$$

with your own α, f, and b. Compute approximate solutions and verify their accuracy along the same lines as in the example above.

[4]In this experiment the Majorant was computed for $\frac{1}{2}\|e\|^2$.

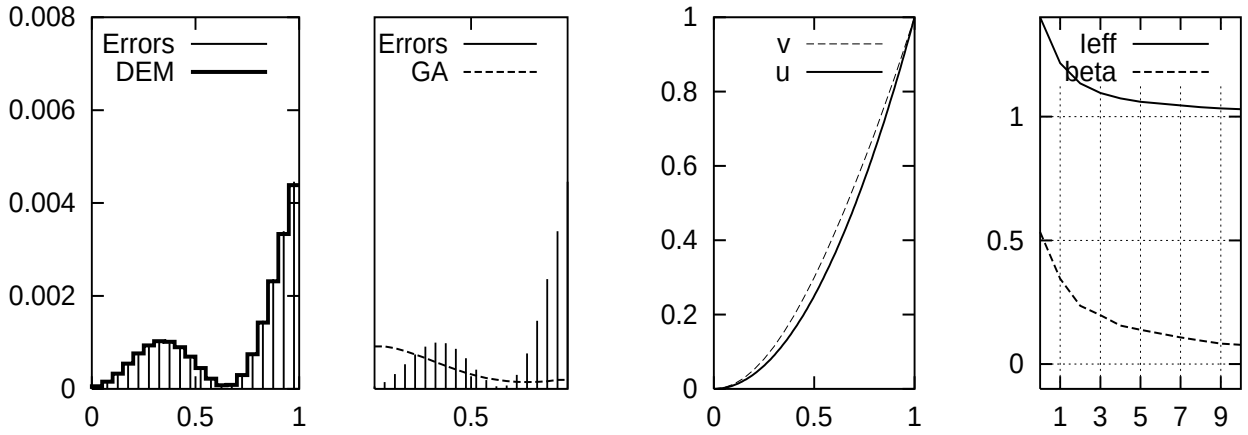


Figure 4.1: $\delta = 0.1$

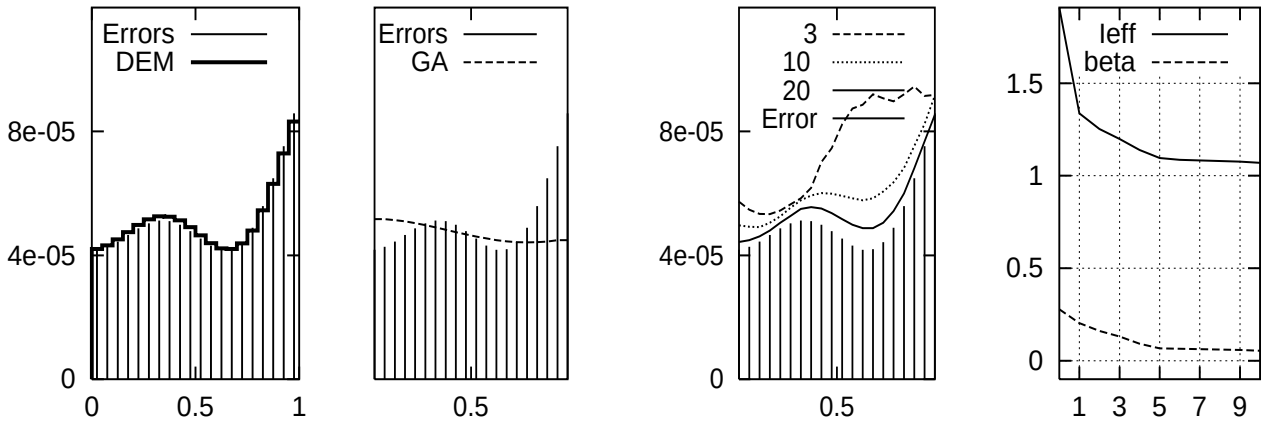


Figure 4.2: $\delta = 0.01$

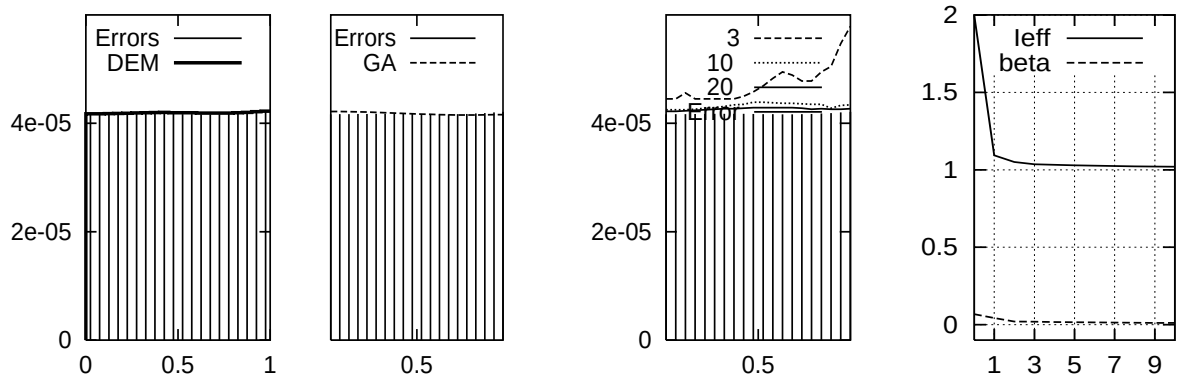


Figure 4.3: $\delta = 0.001$

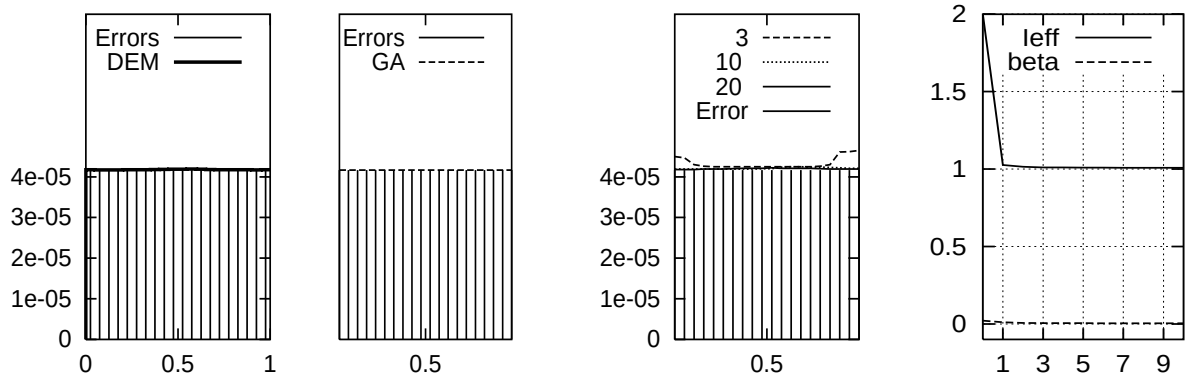


Figure 4.4: $\delta = 0$

4.6 Estimates without Friedrichs constants

For problems with lower terms it is easy to obtain estimates without C_Ω.

$$\begin{aligned} &\Delta u - \varrho u + f = 0, \qquad \varrho > 0, \\ &u = u_0 \quad \text{on } \partial\Omega. \end{aligned}$$

Such estimates can be derived by both *variational* and *non-variational* method. Let $w \in V_0 := \overset{\circ}{H}{}^1(\Omega)$. We have

$$\begin{aligned} &\int_\Omega \nabla(u-v)\cdot\nabla w\, dx + \varrho \int_\Omega (u-v)w dx = \\ &= \int_\Omega (fw - \nabla v\cdot\nabla w) dx - \varrho\int_\Omega vw\, dx. \end{aligned}$$

Use the integral identity for $y \in H(\Omega, \operatorname{div})$:

$\int_\Omega (\nabla w \cdot y + w \operatorname{div} y)\, dx = 0 \quad \forall w \in V_0.$

$$\begin{aligned} \int_\Omega \nabla(u-v)\cdot\nabla w\, dx + \varrho\int_\Omega (u-v)w dx = \\ \int_\Omega (f + \operatorname{div} y - \varrho v) w\, dx + \int_\Omega (y - \nabla v)\cdot\nabla w)\, dx \le \\ \le \|f + \operatorname{div} y - \varrho v\|\|w\| + \|\nabla v - y\|\|\nabla w\|. \end{aligned}$$

Set $w = u - v$ and note that

$$\begin{aligned} &\|f + \operatorname{div} y - \varrho v\|\|u-v\| + \|\nabla v - y\|\|\nabla(u-v)\| = \\ &= \frac{1}{\varrho}\|f + \operatorname{div} y - \varrho v\|\varrho\|u-v\| + \|\nabla v - y\|\|\nabla(u-v)\| \le \\ &\le \left(\frac{1}{\varrho^2}\|f + \operatorname{div} y - \varrho v\|^2 + \|\nabla v - y\|^2\right)^{1/2} \|\!|\!| u - v |\!|\!| \end{aligned}$$

where

$$|\!|\!| u - v |\!|\!|^2 = \int_\Omega (|\nabla(u-v)|^2 + \varrho|u-v|^2) dx.$$

Then, we obtain the estimate

$$\| u - v \|^2 \leq \frac{1}{\varrho^2} \| f + \operatorname{div} y - \varrho v \|^2 + \| \nabla v - y \|^2$$

By the variational method this estimate was derived in 97'. Also, it readily follows from the general a posteriori framework[5].

This estimate has no C_Ω. However, in practice, it may give *big overestimation* if ϱ is small due to large penalty at the first term.

4.7 Estimates in the primal-dual norm

In *mixed* formulations (see Chapter 2), the solution is defined as a pair of functions (u, p). $\overline{\mathfrak{M}}_\Delta(v, q)$ also considers v and q as independent functions. Therefore, it is natural to measure the respective error in terms of combined (primal-dual) norms of the product space

$$\mathcal{W} := V_0 \times H(\Omega, \operatorname{div}),$$

with the norm

$$\|(v, y)\|_{\mathcal{W}} := \|\nabla v\| + \|y\| + \|\operatorname{div} y\| = \|\nabla v\| + \|y\|_{\operatorname{div}}.$$

Other equivalent norms are

$$\begin{aligned} \|(v, y)\|^{(1)}_{\mathcal{W}} &:= \|\nabla v\| + \|y\| + C_{F\Omega} \|\operatorname{div} y\|, \\ \|(v, y)\|^{(2)}_{\mathcal{W}} &:= \left(\|\nabla v\|^2 + \|y\|^2 + \|\operatorname{div} y\|^2 \right)^{1/2}. \end{aligned}$$

It is easy to see that

$$\gamma_1 \|(v, y)\|_{\mathcal{W}} \leq \|(v, y)\|^{(1)}_{\mathcal{W}} \leq \gamma_2 \|(v, y)\|_{\mathcal{W}}, \tag{4.9}$$

$$\frac{1}{\sqrt{3}} \|(v, y)\|_{\mathcal{W}} \leq \|(v, y)\|^{(2)}_{\mathcal{W}} \leq \|(v, y)\|_{\mathcal{W}}, \tag{4.10}$$

where $\gamma_1 = \min\{1, C_{F\Omega}\}$ and $\gamma_2 = \max\{1, C_{F\Omega}\}$.

[5] see, e.g., S.R. Math. Comp. 2000

Let us show that the majorant $\overline{\mathfrak{M}}_{\Delta}(v,y)$ is *equivalent to the error* in the combined norm $\|(v,y)\|^{(1)}_{\mathcal{W}}$. Since

$$\|p-y\| = \|\nabla u - y\| \le \|\nabla(u-v)\| + \|\nabla v - y\|$$

and $\|\operatorname{div}(p-y)\| = \|\operatorname{div} y + f\|$, we find that

$$\begin{aligned}\|(u-v,p-y)\|^{(1)}_{\mathcal{W}} := \;& \|\nabla(u-v)\| + \|p-y\| + C_{F\Omega}\|\operatorname{div} y + f\| \\ & \le 2\|\nabla(u-v)\| + \|\nabla v - y\| + C_{F\Omega}\|\operatorname{div} y + f\| \le \\ & \le 3\overline{\mathfrak{M}}_{\Delta}(v,y).\end{aligned}$$

On the other hand,

$$\overline{\mathfrak{M}}_{\Delta}(v,y) \le \|\nabla(v-u)\| + \|p-y\| + C_{F\Omega}\|\operatorname{div} y + f\|. \tag{4.11}$$

Thus, we note that the following two-sided estimate holds:

$$\overline{\mathfrak{M}}_{\Delta}(v,y) \le \|(u-v,p-y)\|^{(1)}_{\mathcal{W}} \le 3\overline{\mathfrak{M}}_{\Delta}(v,y). \tag{4.12}$$

By (4.12) we conclude that $\overline{\mathfrak{M}}_{\Delta}$ is an *efficient* and *reliable* measure of the error in the combined norm $\|(u-v,p-y)\|^{(1)}_{\mathcal{W}}$.

In view of (4.9) and (4.10), the majorant is also equivalent to two other combined norms, namely,

$$\tfrac{1}{\gamma_2}\overline{\mathfrak{M}}_{\Delta}(v,y) \le \|(u-v,p-y)\|_{\mathcal{W}} \le \tfrac{3}{\gamma_1}\overline{\mathfrak{M}}_{\Delta}(v,y), \tag{4.13}$$

$$\tfrac{1}{\sqrt{3}\gamma_2}\overline{\mathfrak{M}}_{\Delta}(v,y) \le \|(u-v,p-y)\|^{(2)}_{\mathcal{W}} \le \tfrac{3}{\gamma_1}\overline{\mathfrak{M}}_{\Delta}(v,y). \tag{4.14}$$

Also, we can define lower and upper bounds for the norm $\|(u-v,p-y)\|_{\mathcal{W}}$ with the help of the functionals

$$\overline{\mathbb{M}}_{\Delta}(v,y) = 3\|\nabla v - y\| + (1+2C_{F\Omega})\,\|\operatorname{div} y + f\|$$

and

$$\underline{\mathbb{M}}_{\Delta}(v,y) := \|\nabla v - y\| + \|\operatorname{div} y + f\|,$$

which consist of the same terms as those in $\overline{\mathfrak{M}}_{\vartriangle}$ but with different weights. We have

$$\begin{aligned}\|(u-v,p-y)\|_{\mathcal{W}} := \; & \|\nabla(u-v)\| + \|p-y\| + \|\operatorname{div} y + f\| \\ & \le 2\|\nabla(u-v)\| + \|\nabla v - y\| + \|\operatorname{div} y + f\| \le \\ & \le 3\|\nabla v - y\| + (1+2C_{F\Omega})\|\operatorname{div} y + f\| := \overline{\mathbb{M}}_{\vartriangle}(v,y).\end{aligned}$$

Hence, we find that

$$\underline{\mathbb{M}}_{\vartriangle}(v,y) \le \|(u-v,p-y)\|_{\mathcal{W}} \le \overline{\mathbb{M}}_{\Delta}(v,y). \tag{4.15}$$

Similarly,

$$\gamma_1 \underline{\mathbb{M}}_{\vartriangle}(v,y) \le \|(u-v,p-y)\|^{(1)}_{\mathcal{W}} \le \gamma_2 \overline{\mathbb{M}}_{\Delta}(v,y), \tag{4.16}$$

$$\frac{1}{\sqrt{3}} \underline{\mathbb{M}}_{\vartriangle}(v,y) \le \|(u-v,p-y)\|^{(2)}_{\mathcal{W}} \le \overline{\mathbb{M}}_{\Delta}(v,y). \tag{4.17}$$

Finally, we note that

$$\overline{\mathfrak{M}}_{\vartriangle}(v,p) = \|\nabla(u-v)\|, \tag{4.18}$$

$$\overline{\mathfrak{M}}_{\vartriangle}(u,y) = \|y - \nabla u\| + C_{F\Omega}\|\operatorname{div}(y-p)\|. \tag{4.19}$$

Therefore,

$$\|(u-v,p-y)\|^{(1)}_{\mathcal{W}} := \; \overline{\mathfrak{M}}_{\vartriangle}(v,p) + \overline{\mathfrak{M}}_{\vartriangle}(u,y). \tag{4.20}$$

4.8 Error indicators generated by error majorants

Error majorants imply easily computable functions that furnish information on the overall error and adequately reproduce the error function

$$|e(x)| := |\nabla(u-v)|.$$

Such functions are called *error indicators*. We have discussed some error indicators of them in the context of finite element approximations.

Let y_τ be a vector-valued function found by minimization of $\overline{\mathfrak{M}}_{\vartriangle}(v,y)$ with respect to y on a certain finite-dimensional space Y_τ. Then the function

$$\eta(x) := y_\tau - \nabla v$$

is a simple indicator. Experiments have shown that it efficiently reproduces the error distribution, namely:

$$\mathcal{E}_1(v, y_\tau) = |\eta(x)|^2 \approx |e(x)|^2. \tag{4.21}$$

Since

$$\|e - \eta\| = \|\nabla(u - v) - y_\tau + \nabla v\| = \|p - y_\tau\|, \tag{4.22}$$

we see that the indicator $\mathcal{E}_1(v, y_\tau)$ is sharp (i.e., the computable function η is close to $|e|$), if y_τ is close to p.

Let us show that
$\overline{\mathfrak{M}}_{\Delta}(v, y_\tau)$ is sufficiently close to the error, then y_τ is a good representative of the true flux p.

Assume that $v = u_h$, where u_h is a finite element approximation computed on $\mathcal{T}_h$ and $\{y_{\tau_k}\}$ is a sequence of fluxes computed by minimization of $\overline{\mathfrak{M}}_{\Delta}(v, y)$ on expanding spaces $\{Y_{\tau_k}\}$, which are limit dense in $H(\Omega, \operatorname{div})$.

1. It is easy to prove that the exact lower bound of $\overline{\mathfrak{M}}_{\Delta}(v, y)$ (and of $\overline{\mathfrak{M}}_{\beta,\Delta}(v, y)$) with respect to y is attained on a subspace of $H(\Omega, \operatorname{div})$. Indeed, for any $v \in V_0$ (and any $\beta > 0$) the majorant is convex, continuous, and coercive on $H(\Omega, \operatorname{div})$. By known results in the calculus of variations [6], we conclude that a minimizer $\bar{y}(v)$ exists. Since $\overline{\mathfrak{M}}_{\beta,\Delta}(v, y)$ is a quadratic functional, the corresponding minimizer $\bar{y}(v, \beta)$ is unique. We note that it depends on β.

2. Property of the minimizer.

Lemma 4.8.1 *Let $\bar{y}$ be such that*

$$\overline{\mathfrak{M}}_{\Delta}(v, \bar{y}) = \inf_{y \in H(\Omega, \operatorname{div})} \overline{\mathfrak{M}}_{\Delta}(v, y). \tag{4.23}$$

There exists $\bar{w} \in V_0$ such that $\bar{y} = \nabla \bar{w}$.

Proof. For any $y_0 \in \mathrm{S}(\Omega)$ we have

$$\|\nabla v - \bar{y}\| + C_{F\Omega}\|\operatorname{div} \bar{y} + f\| \le \|\nabla v - y_0 - \bar{y}\| + C_{F\Omega}\|\operatorname{div} \bar{y} + f\|.$$

[6] e.g., see I. Ekeland and R. Temam. Convex Analysis and Variational Problems. North-Holland, New York, 1976.

From the above we conclude that for any y_0,

$$\int_\Omega \bar{y}\cdot y_0\,dx + \frac{1}{2}\|y_0\|^2 \geq 0.$$

This inequality holds if and only if

$$\int_\Omega \bar{y}\cdot y_0\,dx = 0, \qquad \forall y_0 \in \mathrm{S}(\Omega). \tag{4.24}$$

Recall that $\bar{y}\in L^2(\Omega,\mathbb{R}^d)$ admits the decomposition $\bar{y}=\nabla\bar{w}+\tau_0$, where $\bar{w}\in V_0$ and τ_0 is a solenoidal field. Set $y_0=\tau_0$. From (4.24), it follows that $\|\tau_0\|=0$. Thus, $\bar{y}=\nabla\bar{w}$.

Task 4.8.1 *Prove that the minimizer of* $\overline{\mathfrak{M}}_{\beta,\Delta}(v,y)$ *has a similar property.*

3. FEM approximations.

By the construction of y_{τ_k}, we know that

$$\overline{\mathfrak{M}}_{\Delta}(v,y_{\tau_k}) \to \|\nabla(u-v)\|. \tag{4.25}$$

Hence, the sequence $\{y_{\tau_k}\}$ is bounded in $H(\Omega,\mathrm{div}\,)$ and a weak limit $\widetilde{y}$ of this sequence exists. Since $\overline{\mathfrak{M}}_{\Delta}(u_h,y)$ is convex and continuous with respect to y, we know that

$$\begin{aligned}\|\nabla(u-u_h)\| = \lim_{k\to+\infty}\overline{\mathfrak{M}}_{\Delta}(u_h,y_{\tau_k}) &\geq \overline{\mathfrak{M}}_{\Delta}(u_h,\widetilde{y})\\ &= \|\nabla u_h-\widetilde{y}\|+C_{F\Omega}\|\mathrm{div}\,\widetilde{y}+f\| \geq \|\nabla(u-u_h)\|. \end{aligned}\tag{4.26}$$

Thus, we conclude that

$$\|\nabla u_h-\widetilde{y}\| + C_{F\Omega}\|\mathrm{div}\,\widetilde{y}+f\| = \|\nabla(u-u_h)\|$$

and, therefore, $\widetilde{y}$ minimizes the functional $\overline{\mathfrak{M}}_{\Delta}(u_h,y)$.

If $\nabla u_h \notin H(\Omega,\mathrm{div}\,)$ (which is typical of FEM approximations), then one can prove that $\widetilde{y}=\nabla u$. Indeed, by Lemma 4.8.1, we know that $\widetilde{y}=\nabla\bar{u}\in H(\Omega,\mathrm{div}\,)$, where $\bar{u}\in V_0$. Then,

$$\|\nabla(u_h-\bar{u})\| + C_{F\Omega}\|\Delta\bar{u}+f\| = \|e\|, \tag{4.27}$$

where $e = \nabla(u - u_h)$. On the other hand,

$$\|\nabla(u - \bar{u})\| \le \|\nabla\bar{u} - y\| + C_{F\Omega}\|\operatorname{div} y + f\|$$

and, therefore,

$$C_{F\Omega}\|\Delta\bar{u} + f\| \ \ge \ \|\nabla(u - \bar{u})\|. \tag{4.28}$$

From (4.27) and (4.28) we conclude that

$$\|e\| \ge \|\nabla(u - \bar{u})\| + \|\nabla(u_h - \bar{u})\|. \tag{4.29}$$

By the triangle inequality,

$$\|e\| \le \|\nabla(u - \bar{u})\| + \|\nabla(u_h - \bar{u})\|, \tag{4.30}$$

and, consequently, (4.29) and (4.30) result in the relation

$$\|e\| = \|\nabla(u - \bar{u})\| + \|\nabla(u_h - \bar{u})\|, \tag{4.31}$$

which implies

$$\int_\Omega \nabla(u - \bar{u}) \cdot \nabla(\bar{u} - u_h)\, dx = \|\nabla(u - \bar{u})\|\|\nabla(\bar{u} - u_h)\|. \tag{4.32}$$

Such a relation is true if (a) $\nabla(u - \bar{u}) = 0$, (b) $\nabla(u_h - \bar{u}) = 0$, or (c)

$$\nabla(\bar{u} - u_h) = \mu\nabla(u - \bar{u}) \qquad \text{for some } \mu \in \mathbb{R},\ (\mu \ne 0). \tag{4.33}$$

In view of the boundary conditions, the case (a) means that $u = \bar{u}$. Since $\nabla u_h \notin H(\Omega, \operatorname{div})$, the case (b) is impossible. From (4.33), it follows that

$$\nabla u_h = (1 + \mu)\nabla\bar{u} - \mu\nabla u \in H(\Omega, \operatorname{div}),$$

so that if $\nabla u_h \notin H(\Omega, \operatorname{div})$, then this relation does not hold and (c) cannot be true. It remains to conclude that $\widetilde{y} = \nabla u$.

Then,

$$\|\nabla(u - u_h)\| = \lim_{k\to+\infty} \overline{\mathfrak{M}}_{\vartriangle}(u_h, y_{\tau_k}) \ge \lim_{k\to+\infty} \|\nabla u_h - y_{\tau_k}\| \ge$$
$$\ge \|\nabla u_h - \widetilde{y}\| = \|\nabla(u - u_h)\|,$$

so that

$$\|\nabla u_h - y_{\tau_k}\| \to \|\nabla(u_h - u)\| \quad \text{as } k \to +\infty.$$

From here, it follows that $\|y_{\tau_k}\| \to \|\nabla u\|$ and, consequently, y_{τ_k} tends to ∇u in $L^2(\Omega)$. Hence, $\|p - y_{\tau_k}\| \to 0$. By (4.22) we then conclude that the indicator $\eta_k := y_{\tau_k} - \nabla u_h$ tends to e as $k \to +\infty$.

The indicator $\mathcal{E}_1$ was verified in numerous tests not only for the Poisson's equation but also for diffusion, linear elasticity, Stokes, and Maxwell's problems (where analogs of this indicator were used). Experiments confirmed its efficiency and stability with respect to approximations of different types.