JYVÄSKYLÄ STUDIES IN COMPUTING
219

Svetlana Matculevich

Fully Reliable A Posteriori Error Control for Evolutionary Problems

JYVÄSKYLÄN YLIOPISTO

JYVÄSKYLÄ STUDIES IN COMPUTING 219

Svetlana Matculevich

Fully Reliable A Posteriori Error Control for Evolutionary Problems

Esitetään Jyväskylän yliopiston informaatioteknologian tiedekunnan suostumuksella
julkisesti tarkastettavaksi yliopiston Agora-rakennuksen auditoriossa 1
lokakuun 30. päivänä 2015 kello 12.

Academic dissertation to be publicly discussed, by permission of
the Faculty of Information Technology of the University of Jyväskylä,
in building Agora, auditorium 1, on October 30, 2015 at 12 o'clock noon.

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2015

Fully Reliable
A Posteriori Error Control
for Evolutionary Problems

JYVÄSKYLÄ STUDIES IN COMPUTING 219

Svetlana Matculevich

Fully Reliable A Posteriori Error Control for Evolutionary Problems

UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2015

Editors
Timo Männikkö
Department of Mathematical Information Technology, University of Jyväskylä
Pekka Olsbo, Ville Korkiakangas
Publishing Unit, University Library of Jyväskylä

URN:ISBN:978-951-39-6291-3
ISBN 978-951-39-6291-3 (PDF)

ISBN 978-951-39-6290-6 (nid.)
ISSN 1456-5390

Jyväskylä University Printing House, Jyväskylä 2015

ABSTRACT

Matculevich, Svetlana
Fully reliable a posteriori error control for evolutionary problems
Jyväskylä: University of Jyväskylä, 2015, 75 p. (+included articles)
(Jyväskylä Studies in Computing
ISSN 1456-5390; **219**)
ISBN 978-951-39-6290-6 (nid.)
ISBN 978-951-39-6291-3 (PDF)
Finnish summary

This work is devoted to fully reliable a posteriori error analysis for a class of evolutionary problems and some questions emerging in relation to it. The first articles in this collection are concerned with theoretical and numerical analysis, efficient and robust implementation of the functional type a posteriori error estimates and indicators for the nonlinear Cauchy problem, and time-dependent reaction-diffusion initial-boundary value problems of parabolic type. The last part of the study is dedicated to computable and sharp upper bounds of constants in Poincaré-type inequalities for functions with zero mean on the boundary (or a measurable part of it) on non-degenerate triangles and tetrahedrons. These sharp upper bounds are crucial for quantitative analysis of problems generated by differential equations, where numerical approximations are typically constructed with the help of simplicial meshes and become particularly useful in implementation of the functional error majorants applied for the problems with a decomposed domain.

The error estimates presented in this thesis are explicitly computable and guaranteed. The two-sided functional type error bounds hold for all conforming approximations, do not depend on any mesh discretization parameters, and only contain global and local constants in Poincaré inequalities. Extensive numerical experiments, performed alongside with theoretical findings, provide results, which confirm the efficiency and reliability of the error estimates and robustness of the indicators they comprise. For numerical implementation we use MATLAB and The FEniCS Project (with Python).

Keywords: Cauchy problem, Picard–Lindelöf method, Ostrowski estimates, evolutionary problem of parabolic type, reaction-diffusion equation, functional type a posteriori error estimates, error indicators, Poincaré-type estimates

Author *M.Sc. Svetlana Matculevich*
Department of Mathematical Information Technology
University of Jyväskylä
Finland

Supervisors *Prof. Dr. Pekka Neittaanmäki*
Department of Mathematical Information Technology
University of Jyväskylä
Finland

Prof. Dr. Sergey Repin
St. Petersburg Department of V.A. Steklov Institute of
Mathematics of the Russian Academy of Sciences
Russia
Department of Mathematical Information Technology
University of Jyväskylä
Finland

Reviewers *Prof. Dr. Roland Glowinski*
University of Houston
Department of Mathematics
Houston, TX
USA

Prof. Dr. Ulrich Langer
Institute of Computational Mathematics
Johann Radon Institute for Computational and
Applied Mathematics (RICAM)
Austrian Academy of Sciences (ÖAW)
Austria

Opponent *Prof. Dr. Stefan Sauter*
Institute of Mathematics
Zürich University
Switzerland

ACKNOWLEDGEMENTS

I am greatly indebted to my supervisors Prof. Pekka Neittaanmäki and Prof. Sergey Repin for their excellent supervision, which started in 2011 during my Master thesis, for fruitful scientific discussions with them, for their ongoing support to my attempts to make scientific contributions, and for their constant mentoring outside of scientific work.

I would like to express my appreciation and admiration to the reviewers Prof. Dr. Roland Glowinski and Prof. Dr. Ulrich Langer for their valuable comments and the significant improvement of the quality of my thesis during the review process.

I am deeply grateful to all the local and international members of our research group 'Reliable Methods for Computer Simulation' for interesting discussions we had about scientific and non-scientific matters. I want to especially thank Dr. Olli Mali and Dr. Immanuel Anjam for their help and guidance during my PhD studies and detailed corrections for the thesis. My deep appreciation also goes to Dr. Monika Wolfmayr for thorough proof-reading and revision of this work.

My kind acknowledgments go the GETA Postgraduate School, COMAS Graduate School, Ella and Georg Ehrnrooth Foundation, and Finnish Academy of Science and Letters Foundation for funding my PhD research. I also thank the Department of Mathematical Informational Technology and its head Prof. Tuomo Rossi for financial support to allow me to participate in several conferences and make research visits during my studies.

Lastly, I thank my family for their love and support despite the distance between us, to my boyfriend Philipp, who significantly contributed to this journey by dragging me away from work every so often, showed me care and encouragement, and yet was tolerant to my endless hours in the office, and last but not the least my dearest friends scattered around many countries for all the wonderful experiences shared throughout the past years.

I dedicate this work to the loving memory of my mother Elena (1966–2008).

Svetlana Matculevich
Jyväskylä, September 21, 2015

NOTATION

$:=$	equals by definition
$\hookrightarrow$	compact embedding
$\equiv$	logical equivalence
$\forall$	for all
$a \cdot b$	scalar product of vectors
$\mathbb{N}$	space of natural numbers
$\mathbb{R}$	space of real numbers
$\mathbb{R}^d$	space of real d-vectors
Ω	open bounded connected domain in $\mathbb{R}^d$ with Lipschitz continuous boundary
$\overline{\Omega}$	closure of Ω
$\partial\Omega$	Lipschitz continuous boundary of Ω
Γ	part of $\partial\Omega$ such that $\mathrm{meas}_{d-1}\Gamma > 0$
Q_T	space-time cylinder $Q_T := \Omega \times (0, T)$, where T is given time
S_T	lateral surface of Q_T, i.e., $S_T := \partial\Omega \times [0, T]$
$\mathrm{diam}\Omega$	diameter of the set Ω
$\mathrm{meas}\Omega$	Lebesgue measure of the set Ω
$D^\alpha v$	derivative of order $\lvert\alpha\rvert$
$C^k(\Omega)$	space of k-times differentiable scalar-valued functions
$C_0^k(\Omega)$	subspace of $C^k(\Omega)$ that contains functions with compact support in Ω
$C_0^\infty(\Omega)$	space of smooth functions with compact support in Ω
$L^p(\Omega)$	space of scalar-valued functions in Ω summable with power p
$L^p(\Omega, \mathbb{R}^d)$	space of vector-valued functions with components summable with power p in Ω
X	Banach space
V	Hilbert space
V^*	space dual to V
$W^{l,p}(\Omega)$	Sobolev space of functions w summable with power p and possessing derivatives $D^\alpha w \in L^p(\Omega)$, $\lvert\alpha\rvert \le l$
$H^l(\Omega)$	Sobolev space $W^{l,p}$ with $p = 2$
$H_0^l(\Omega)$	subspace of $H^l(\Omega)$ formed by functions vanishing on Γ
$H^{-1}(\Omega)$	space dual to $H_0^1(\Omega)$
$H(\Omega, \mathrm{div})$	subspace of $L^2(\Omega, \mathbb{R}^d)$ that contains vector-valued functions with square-summable divergence
$\lVert\cdot\rVert_X$	norm in space X
$\lVert\cdot\rVert$	norm in $L^2(\Omega)$
$\lvert\!\lvert\!\lvert\cdot\rvert\!\rvert\!\rvert$	energy norm
$\lVert w\rVert_A$	a weighed norm in $L^2(\Omega)$, i.e., $\left(\int_\Omega Aw \cdot w \,\mathrm{d}x\right)^{1/2}$

$\|w\|_{A^{-1}}$	$\big(\int_\Omega A^{-1}w \cdot w \,\mathrm{d}x\big)^{1/2}$
w_t	partial derivative with respect to time coordinate
$w_{,i}$	partial derivative with respect to space i^{th} coordinate
∇	gradient of a scalar-valued function $\nabla w = (w_{,1}, \ldots, w_{,d})$
Δ	Laplace operator $\Delta w := \operatorname{div}\nabla w$
div	divergence of a vector-valued function $\operatorname{div} w = \sum_{i=1}^{d} w_{i,i}$
$\{w\}_\Omega$	mean value of w on Ω, i.e., $\{u\}_\Omega := \frac{1}{\|\Omega\|}\int_\Omega w \,\mathrm{d}x$
π	projection operator
e	error
$\overline{\mathrm{M}}$	majorant functional
$\underline{\mathrm{M}}$	minorant functional
I_{eff}	efficiency index $I_{\text{eff}} := \frac{\overline{\mathrm{M}}}{\|\|\|e\|\|\|}$
$\mathbb{M}$	marker
$\mathbb{M}_{\text{AVR}}$	marker defined by the level of the average error
$\mathbb{M}_\theta$	marker with bulk parameter θ
P_k	Lagrangian finite element space of order k
RT_k	Raviart-Tomas finite element space of order k

ACRONYMS

APL	adaptive Picard–Lindelöf
BVP	boundary value problem
BC	boundary condition
DD	domain decomposition
DOF	degrees of freedom
EL	elements
ND	nodes
FDM	finite difference method
FE	finite element
FEM	finite element method
I-BVP	initial-boundary value problem
LHS	left-hand side
PDE	partial differential equation
RHS	right-hand side
REF	refinement iteration
SLE	system of linear equations

LIST OF TABLES

CONTENTS

LIST OF INCLUDED ARTICLES

PI S. Matculevich, P. Neittaanmäki, and S. Repin. Guaranteed error bounds for a class of Picard-Lindelöf iteration methods. *Numerical methods for differential equations, optimization, and technological problems, Comput. Methods Appl. Sci.*, **27**: 175–189, 2013.

PII S. Matculevich and S. Repin. Computable estimates of the distance to the exact solution of the evolutionary reaction-diffusion equation. *Applied Mathematics and Computation*, **247**: 329–347, 2014.

PIII S. Matculevich, P. Neittaanmäki, and S. Repin. A posteriori error estimates for time-dependent reaction-diffusion problems based on the Payne–Weinberger inequality. *Discrete and Continuous Dynamical Systems - Series A, AIMS*, **35**(6): 2659–2677, 2015.

PIV S. Matculevich and S. Repin. Estimates of the distance to the exact solution of evolutionary reaction-diffusion problems based on local Poincaré type inequalities. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov (POMI)*, **425**(1): 7–34, 2014.

PV S. Matculevich and S. Repin. Sharp bounds of constants in Poincaré type inequalities for polygonal domains. *arXiv*, math/1504.031662, 2015.

1 INTRODUCTION

Nowadays, *mathematical models* are widely used to describe processes in different branches of natural sciences, medicine, engineering, and economics. Evolutionary problems, in particular, are fundamental components in simulations of real-life processes such as heat conduction and thermal radiation models in thermodynamics, global climate prediction, forecasting and understanding the weather, and estimation of forest growth, among others. Later examples basically testify the fact that questions arising in mathematical modeling originate from and are highly motivated by the phenomena surrounding us.

Most of the models mentioned above are governed by time-dependent *partial differential equations* (PDEs) or systems of PDEs, which in combination with initial (IC) and boundary conditions (BCs) produce so-called *initial-boundary value problems* (I-BVPs). The current study is focused on evolutionary problems of *parabolic type*, the systematic mathematical analysis of which is presented in monographs [79, 80, 148, 151, 152]. The numerical analysis and study of the practical application are exposed in works [138, 81] and partially in classical books on finite element method (FEM) on PDEs and saddle problems (see, e.g., [21, 58, 69, 56, 57]). The multiharmonic analysis of a distributed parabolic and optimal control problem in a time-periodic BVPs setting has been studied in [73, 84].

Let $Q_T := \Omega\times]0,T[$ denote the space-time cylinder, where $\Omega \subset \mathbb{R}^d$, $d \in \{1,2,3\}$, is an bounded domain with Lipschitz boundary $\partial\Omega$, and $]0,T[$ is a given time interval, $0 < T < +\infty$. The cylindrical surface is denoted by S_T, i.e., $S_T := \partial\Omega \times [0,T]$. A general form of a *linear parabolic I-BVP problem* reads as follows:

$$\partial_t u + \mathcal{L}u = f \quad \text{in} \quad Q_T, \tag{1.1}$$
$$u = u_D \quad \text{on} \quad S_T, \tag{1.2}$$
$$u(x,0) = u_0 \quad \text{on} \quad \Omega. \tag{1.3}$$

Here, depending on the application, u might describe the temperature alteration in heat conduction or the concentration of certain substance in chemical diffusion. The given data includes the source term f, Dirichlet BC u_D (Neumann or Robin

can be considered instead), and IC u_0. The elliptic operator $\mathcal{L}$ has the general form

$$\mathcal{L}u := -\mathrm{div}(A(x,t)\nabla u(x,t)) + b(x)\cdot\nabla u(x,t) + c(x)\,u(x,t), \quad (x,t)\in Q_T,$$

where A is the material characteristics matrix, and b and c stand for convection and reaction, respectively. If any of the latter forms depend on u (or ∇u), we arrive at a nonlinear problem.

For $b \equiv 0$, $c \equiv 0$, and $A = \nu I$, we obtain a heat equation which governs diffusion processes. For instance, in heat conduction applications the parameter $\nu = \frac{k}{c_p \varrho}$ stands for thermal diffusivity [50, 24, 147, 23], in electromagnetics it illustrates resistivity $\nu = \frac{1}{\sigma}$. Moreover, the heat equation is used in propagation of action potential in nerve cells, phenomena arising in finance, e.g., the Black–Scholes [20] or Ornstein-Uhlenbeck processes, probability, and description of random walks [109]. The nonlinear analogs of the heat equation have also been used in image processing and modeling of porous media [141].

The subject of our interest, i.e., evolutionary systems of PDEs, in majority of cases can only be solved in the generalized sense by one of two discretization techniques described below. In the first, the so-called *incremental time-stepping method*, the time is discretized by ordinary differentiation (OD) and the obtained reduced problem (in space coordinates) is approximated by FEM ([35, 66, 157, 32, 69]) or the finite difference method (FDM [87, 127, 98, 58, 37]) on successive time sub-intervals (the detailed study of such an approach can be found in the monographs [138, 21, 69]). In the second method the time is considered as an additional spatial variable [60, 149, 140, 63]. It is usually referred to as the *space-time discretization* technique. Regardless of the method used, the obtained approximation contains an *error*. Therefore, it is of high importance to construct a proper numerical tool to analyze the obtained results and to provide reliable information on the *approximation error* encompassed in it in order to avoid the risk of drawing the wrong conclusion form obtained numerical information.

There exist two approaches for evaluating the approximation error. The *a priori* approach is used for the qualitative verification of theoretical properties of the numerical method, e.g., rate of convergence and asymptotic behavior of the approximation with respect to mesh size parameters (see, e.g., [22, 32, 133] and references cited therein). However, the high regularity requirements, which must be satisfied in order to apply estimates from the latter group, are quite unrealistic.

In the second, the so-called *a posteriori* approach, the error is measured after computation of the approximation. Unlike in a priori error analysis, the alternative estimates exploit only the given data, e.g., domain characteristics, source function together with IC and BC, and the approximation itself. The upper bound of the gap between the approximate and exact solution measured in terms of relevant energy norm is called an *error estimate* or *majorant*. The quantity replicating the distribution of the true error over the domain is called an *error indicator*. There are three principal ways classifying existing error indicators. The first, the so-called *residual method*, is based on the estimation of

the residual functional introduced in [11, 8] and various modifications of them covered in a wealth of publications [47, 70, 3, 4, 142, 44, 25, 30, 5, 26, 10, 12]). The second approach is based on the approximation of latter functional or so-called *post-processing*, e.g., gradient averaging [155, 156] and expanded in various works [3, 9, 142, 154, 145, 10, 15, 146, 62, 153]). Its mathematical justification relies on the *superconvergence* phenomenon [104, 158] and actively studied in [75, 76, 77, 77, 144]. Other techniques from the second group are based on partial equilibration [78, 5, 21], global averaging [26, 15, 62], and solution of local sub-problems [2, 5, 6]. Finally, the third method is dependent on the solution of the auxiliary problem, e.g., *hierarchically* based error indicators [38, 1, 45, 43] and *goal-oriented* error [16, 131, 110, 65, 116, 103, 132, 96, 117, 19]. The concept of a posteriori error estimation jointly with mesh-adaptive methods, which are focused on the optimization of computing resources, have become a well-established approach in the numerical analysis of PDEs.

The guaranteed error bounds for evolutionary models considered in this thesis are based on two different mathematical approaches. One of these follows from the theory of contraction mappings and the Banach fixed point theorem. The other approach pursues the theory of functional a posteriori estimates.

The first part of the study, in particular, is dedicated to the investigation of numerical treatment of the Cauchy problem with non-linearity (see, e.g., [33, 61, 136]), which can be obtained from (1.1)–(1.3) by assuming that Ω coincides with $\mathbb{R}^d$. The so-called Picard–Lindelöf method suggests one possible way to treat nonlinear ordinary differentiation equations (ODEs). It belongs to a class of iteration method and can be found in [89, 108, 17, 88, 111]. A similar idea is used for PDEs in [111] and analyzed thoroughly in [112, Vol.II]. The combination of the Picard–Lindelöf method with Ostrowski a posteriori estimates provides a fully guaranteed Adaptive Picard–Lindelöf (APL) algorithm for solving ODEs. Moreover, the algorithm takes into account information about discretization errors related to the numerical integration and interpolation. The results obtained during the investigation of the APL method confirmed that it can be applied for the treatment of nonlinear evolutionary models, which belong to my main research topics for the future. Nonlinear PDEs exhibit multiple properties which do not appear in linear theory but are often related to important features of the real world phenomena. In this work, we concentrate only on the linear models. The application of Ostrowski estimates is also extended to classical iteration schemes, the obtained results are exposed in [93, Section 6.7].

Generally, the Picard–Lindelöf method can be used not only for ODEs but also for time-dependent algebraic and functional equations (see, e.g., [101, 102], where it is shown that the speed of convergence is independent of the step sizes). Numerical methods based on Picard–Lindelöf iterations for dynamical processes (the so-called waveform relaxation in the context of electrical networks) are discussed in [46]. A posteriori estimates and nodal superconvergence for time stepping methods are studied in [7, 92] for linear and nonlinear problems.

The second and main part of this work is devoted to the *functional type* a posteriori error estimates and indicators initially introduced by Repin in [119, 118,

123, 120] and thoroughly studied for various classes of problems (see, e.g., [100, 121, 93] and references therein). Unlike the above-listed error indicators, functional type error estimates are guaranteed, they do not contain mesh-dependent local interpolation constants (contrary to residual estimates), and they are valid for any function from the class of conforming approximations (not restricted by the Galerkin orthogonality assumption). The detailed comparison of the above-described approaches can be found in monograph by Mali, Neittaanmäki, and Repin [93].

Our main goal is to develop a fully reliable tool to quantitatively control the error in approximate solutions of evolutionary problems. The numerical treatment of this class of I-BVPs produces approximations, which alongside with the progress of simulations accumulate the error. This error may eventually 'blow up' if it is not controlled. Therefore, the appropriate error estimates are crucial for monitoring its possible dramatic growth. Once the error in the approximation has been controlled reliably, it is possible to detect areas with excessively high local errors and calculate an essentially more accurate approximation.

In the framework of the a posteriori error estimates studied in this work, we highlight the paper [124], where a method of deriving functional error estimates for parabolic I-BVPs is suggested. The first attempt on their numerical analysis is presented in [54]. In [125], the authors study the extension of error estimates for evolutionary convection-diffusion problems with possible discontinuity of approximations in time. A posteriori error analysis of parabolic time-periodic BVPs in connection with their multiharmonic FE discretization is presented in [83]. The residual estimates are also extended to evolutionary PDEs in [143, 14, 128, 95, 19, 126] and the reference cited therein. Lastly, *hp*-Galerkin time-stepping for the same class of problems is addressed in [68, 129, 130] and references cited therein.

In order to make functional estimates applicable to a wider class of problems with Ω of complicated geometry, the domain decomposition (DD) technique in combination with local Poincaré inequalities is discussed in [121, Section 3.5.3] for elliptic PDEs. The current work extends the latter estimates to time-dependent PDEs and suggests a method to omit both Friedrichs' and trace global constants, which are included into the basic form of the majorant.

Suggested in [PIII] and [PIV] method applies Poincaré-type inequalities that in addition to quantitative analysis of PDEs is also used in various problems of numerical analysis, e.g., discontinuous Galerkin, mortar and DD methods. The exact values of respective constants (or sharp and guaranteed bounds of them) are interesting from both analytical and computational points of view. Results related to constants in extension and projection type estimates related to FE approximations can be found in, e.g., [97, 32]. Constants in the trace inequalities associated with polygonal domains are discussed in [29]; in FETI, FETI-DP DD methods the application of constants is highlighted in [72, 39] and [139]. Functional inequalities and respective constants play an important role in analysis of problems described in terms of vector-valued functions (see, e.g., [53, 105]). In [71, 90], the analysis of error constants for piecewise constant and linear interpo-

lations over triangular finite elements can be found. And finally, [27] introduces fully computable two-sided bounds on the eigenvalues of the Laplace operator based on the approximation of the corresponding eigenfunction in the nonconforming Crouzeix-Raviart FE space.

The last part of this study is dedicated to sharp bounds of the constants in classical Poincaré and Poincaré-type inequalities for arbitrary non-degenerate triangles and tetrahedrons, which are typical objects in various discretization methods. These computable estimates are based on the mapping of the reference simplices to arbitrary one using the exact values of the respective constants derived in [113, 64, 71] for some triangles and [99] for parallelepipeds, rectangles, and right triangles. Knowledge about the sharp upper bounds for the above-mentioned constants is particularly useful for quantitative analysis of problems generated by differential equations and implementation of the functional error majorants applied for the problems with decomposed domain.

Below, we sketch the structure of the thesis. Chapter 2 is dedicated to the overview of the mathematical framework, including definitions and theorems in the field of functional analysis as well as results on solvability parabolic I-BVPs, which provide fundamental results required in the subsequent chapters. Chapter 3 is focused on the main results achieved in this study, i.e., a fully guaranteed APL method for ODEs, functional a posteriori error estimates for the distance to the exact solution of parabolic I-BVPs, and sharp bounds of the constants in classical Poincaré and Poincaré-type inequalities for functions with zero mean traces on the faces of arbitrary simplexes in $\mathbb{R}^2$ and $\mathbb{R}^3$. In Chapter 4, we draw some conclusions and give an outlook on future work in connection to efficient and fully guaranteed solvers for nonlinear evolutionary problems. The results presented in the included papers, or in other publications, will be highlighted accordingly. The connections between the topics are presented in Figure 1.

Author's contribution to the included articles

[PI]: The estimates studied in this paper were discussed originally in monograph of Neittaanmäki and Repin [100, Section 3.1]. The goal of this work is to implement the adaptive iterative Picard–Lindelöf method and combine it with Ostrowski estimates. The computations of the numerical part are carried out in `MATLAB` [94] by the author. Application of Ostrowski estimates to classical iteration schemes is also presented in [93, Section 6.7.6] together with a guaranteed APL method.

[PII]: This article studies functional type a posteriori error estimates for evolutionary reaction-diffusion I-BVP with a reaction function, which drastically changes its values on different parts of the domain. The method suggested for derivation of the majorant combines ideas presented in original work of Repin [124] on bounds of the distance to the exact solution of heat equation and join paper of Repin and Sauter [122], which is concerned with state reaction-diffusion BVP. The minorant of the error in the approximate solution for the evolutionary class of problems derived in the paper is the original result. Its efficiency is

confirmed by numerical tests. All experiments presented in the paper are implemented by the author in MATLAB.

[PIII]: The focus of this paper is on error estimates for an approximate solution of the evolutionary reaction-diffusion problem in case of decomposed domains. The method suggested in the paper is based on the idea originally introduced for the elliptic problems in [121] and [122]. The main goal of the work is to overcome the complications arising with the calculation of Friedrichs' constant included in the majorant presented in [PII] once it is applied to problems with a domain of a complicated shape. By exploiting the idea of DD and classical Poincaré inequalities [114, 115], we exclude global constants from the majorant. The proofs and technicalities in the paper are the work of the author.

[PIV]: This work is another generalization of the error estimates presented in [PII] to the problems formulated on complicated domains with nontrivial mixed Dirichlet–Robin BC. Again, by using the method of domain decomposition and application of local Poincaré-type inequalities for functions with zero mean trace, we omit global trace and Friedrichs' constants included into the basic majorant. Besides that, we demonstrate the equivalence of errors measured in primal and combined norms to advanced and basic forms of majorants, respectively.

[PV]: The technique suggested in [PIV] and [PIII] is based on local Poincaré and Poincaré-type inequalities for functions with zero mean trace on the whole boundary or measurable part of it. We suggest explicit relations (based on exact constants from [113, 64, 71, 99]) that serve as sharp and easily computable (independent of any discretization parameters) bounds of the respective constants. Moreover, we compare obtained bounds of the constants in the classical Poincaré inequalities with known analytical estimates and investigate, numerically, the behavior of minimizers of Rayleigh quotients, corresponding the constants. The numerical experiments in the paper are carried out by the author, using both MATLAB and The FEniCS Project [137, 91].

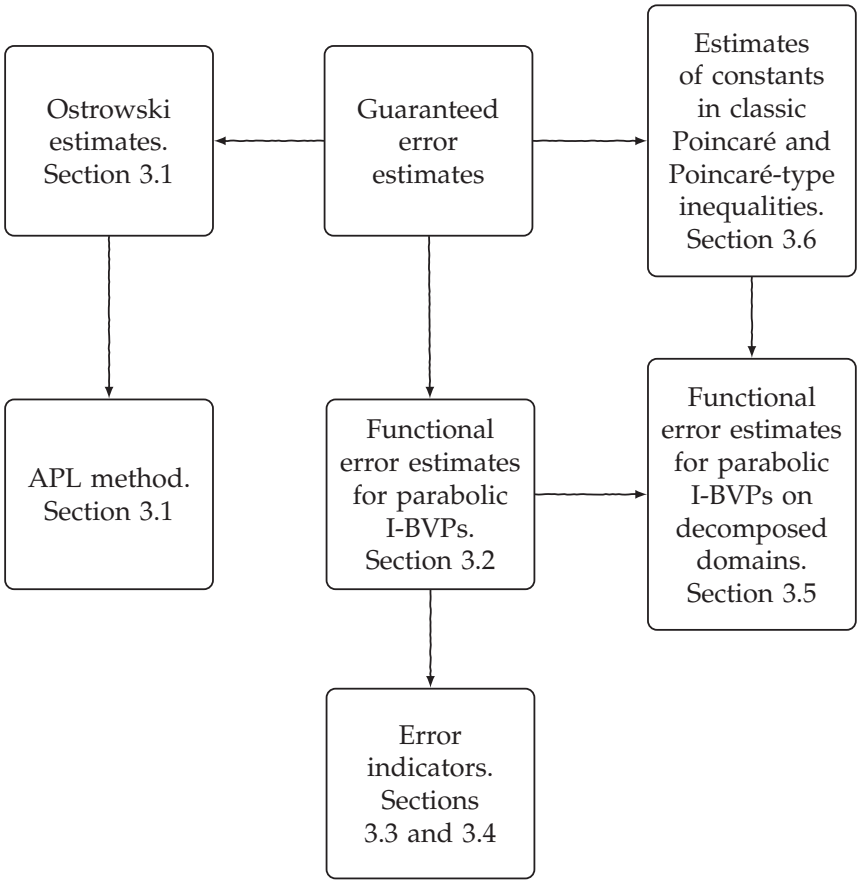


FIGURE 1 Structure of the thesis.

2 MATHEMATICAL BACKGROUND

In this chapter, we concisely introduce the notation, mathematical framework, and fundamental results that form the basis for the further investigations and findings presented in this thesis. For detailed expositions, we refer the reader to monographs [100, 121, 93].

2.1 Function spaces and inequalities

The sections below present the definitions and main results for Sobolev spaces, which are used for the treatment of elliptic BVPs and parabolic I-BVPs. Although, some results are quite well-known, we discuss them to keep the work as self-content as possible. In addition, for more detailed and fundamental presentations of the results highlighted below, we refer the reader to [48, 151, 152, 148].

2.1.1 Spaces of integrable functions

Let $\Omega \subset \mathbb{R}^d$, $d = \{1, 2, 3\}$, be a bounded domain with Lipschitz boundary $\partial\Omega$, where $\overline{\Omega}$ is the closure of Ω, and Γ be a part of $\partial\Omega$ such that $\text{meas}_{d-1}\Gamma > 0$ (or in particular case may coincide with it). We note that throughout the thesis discussions will be restricted to real spaces. Let $\{X, \|\cdot\|_X\}$ denote a *Banach space*, i.e., a vector space X equipped with a norm $\|\cdot\|_X$, such that X is complete with respect to it. Let $\{V, \|\cdot\|_V\}$ denote a *Hilbert space*, where the norm is induced by the inner product $(\cdot,\cdot)_V : V \times V \to \mathbb{R}$, i.e., $\|\cdot\|_V := (\cdot,\cdot)_V^{1/2}$. The space V^* denotes the dual to V, consists of linear continuous functionals on V, and is equipped with norm $\|f\|_{V^*} := \sup\limits_{v\in V, v\neq 0} \frac{f(v)}{\|v\|_V}$. The so-called duality product $\langle\cdot,\cdot\rangle_{V^*\times V} : V^* \times V \to \mathbb{R}$ is defined as

$$\langle f, v\rangle_{V^*\times V} := f(v), \quad \forall v \in V. \tag{2.1}$$

The totality of all measurable in the Lebesgue sense functions u with finite norm

$$\|u\|_{L^p} := \Big(\int\limits_{\Omega} |u(x)|^p \, \mathrm{d}x \Big)^{1/p}.$$

forms a separable Banach space and is denoted by $L^p(\Omega)$, $p \in [1, +\infty[$. For spaces of essentially bounded functions with $p = \infty$, the norm is defined as

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

Further, we are mainly interested in the Hilbert space of square-integrable functions $L^2(\Omega)$ eqquiped with the norm $\| \cdot \|_{L^2(\Omega)} := (\cdot, \cdot)^{1/2}_{L^2(\Omega)}$ induced by

$$(u, v)_{L^2(\Omega)} = (u, v) := \int\limits_{\Omega} u \, v \, \mathrm{d}x, \quad \forall u, v \in L^2(\Omega).$$

For the purpose of shortening the notation, in the cases of discussing L^2-measures on Ω, the L^2-norm is denoted $\| \cdot \|_\Omega$.

2.1.2 Differentiability classes

Let $\alpha = (\alpha_1, \ldots, \alpha_d)$, $\alpha_i \in \mathbb{N} \cup 0$, $i = 1, \ldots, d$, be a multi-index; then $D^\alpha u := \frac{\partial^{|\alpha|}}{\partial x^\alpha} u = \frac{\partial^{\alpha_1}}{\partial x^{\alpha_1}} \ldots \frac{\partial^{\alpha_d}}{\partial x^{\alpha_d}} u$, where x^α is the monomial ${x_1}^{\alpha_1} \ldots {x_d}^{\alpha_d}$ with degree $|\alpha| = \sum\limits_{i=1}^{d} \alpha_i$. Functions in $C^l(\Omega)$ possess continuous and bounded derivatives D^α up to order l. The space $C^l(\overline{\Omega})$ is equipped with the norm

$$\|u\|_{C^l(\overline{\Omega})} := \max_{0 \le |\alpha| \le l} \sup_{x \in \overline{\Omega}} |D^\alpha u(x)|.$$

The norm for continuous functions ($l = 0$) is defined by $\| \cdot \|_{C(\overline{\Omega})}$. The space $C^\infty(\Omega)$ consists of infinitely differentiable (smooth) functions, and elements of $C_0^\infty(\Omega) \subset C^\infty(\Omega)$ have compact support in Ω. Smooth functions vanishing on Γ are denoted by

$$C^\infty_{0,\Gamma}(\Omega) := \Big\{ \varphi \in C^\infty(\Omega) \,|\, \operatorname{dist}(\operatorname{supp}\varphi, \Gamma) > 0 \Big\}. \tag{2.2}$$

2.1.3 Sobolev spaces

The α^{th} weak (or generalized) derivative of $u \in L^2(\Omega)$ is denoted by $w = D^\alpha u \in L^2(\Omega)$ such that

$$\int\limits_{\Omega} w \, v \, \mathrm{d}x = (-1)^{|\alpha|} \int\limits_{\Omega} u \, D^\alpha v \, \mathrm{d}x, \quad \forall v \in C_0^\infty(\Omega).$$

The separable space of Banach type $W^{l,p}(\Omega)$, $p \in [1, +\infty[$ and $l \in \mathbb{N}$, is called the *Sobolev space* $W^{l,p}(\Omega) := \left\{ u \in L^p(\Omega) \mid D^\alpha u \in L^p(\Omega), |\alpha| \leq l \right\}$ and equipped with the norm

$$\|u\|_{W^{l,p}} := \Big(\sum_{|\alpha| \leq l} \|D^\alpha u\|_{L^p}^p \Big)^{1/p}. \tag{2.3}$$

If the boundary of Ω is smooth enough, latter space coincide with a clouser of $C^l(\overline{\Omega})$ under the norm (2.3), i.e., $\mathcal{W} := \overline{C^l(\overline{\Omega})}^{\|\cdot\|_{W^{l,p}}}$ (in general, $\mathcal{W} \subset W^{l,p}$).

The Hilbert spaces with $p = 2$ are traditionally denoted as $H^l(\Omega) = W^{l,2}(\Omega)$. Later in the thesis, we use the spaces

$$H^1(\Omega) := \left\{ u \in L^2(\Omega) \mid \nabla u \in L^2(\Omega, \mathbb{R}^d) \right\}, \quad \text{and}$$
$$H(\mathrm{div}, \Omega) := \left\{ u \in L^2(\Omega, \mathbb{R}^d) \mid \mathrm{div} u \in L^2(\Omega) \right\},$$

with the corresponding norms $\|\cdot\|_{H^1(\Omega)}$ and $\|\cdot\|_{H(\mathrm{div},\Omega)}$ induced by

$$(u, v)_{H^1} := (u, v) + (\nabla u, \nabla v) \quad \text{and} \quad (u, v)_{H(\mathrm{div})} := (u, v) + (\mathrm{div} u, \mathrm{div} v),$$

respectively. Spaces with homogenous boundary conditions on $\Gamma \subset \partial\Omega$ are defined as closures of (2.2):

$$H^1_{0,\Gamma}(\Omega) := \overline{C^\infty_{0,\Gamma}(\Omega)}^{H^1(\Omega)} \quad \text{and} \quad H_{0,\Gamma}(\mathrm{div}, \Omega) := \overline{C^\infty_{0,\Gamma}(\Omega)}^{H(\mathrm{div},\Omega)}.$$

If $\Gamma = \partial\Omega$, then $H^1_{0,\partial\Omega}(\Omega) = H^1_0(\Omega)$. Lastly, let $\gamma_\Gamma u \in C(\Gamma)$ denote the restriction of $u \in C(\overline{\Omega})$ to Γ, i.e., $\gamma_\Gamma u(x) := u(x)$, $\forall x \in \Gamma$. The latter one is called *trace operator* $\gamma_\Gamma : H^s(\Omega) \to H^{s-1/2}(\Gamma)$, $s \in (\frac{1}{2}, \frac{3}{2})$.

2.1.4 Inequalities

We list several algebraic and functional inequalities frequently used in the thesis. For $a, b \in \mathbb{R}$ and any positive β, we have general Young inequality

$$ab \leq \tfrac{1}{p}(\beta a)^p + \tfrac{1}{q}(\tfrac{b}{\beta})^q, \quad \tfrac{1}{p} + \tfrac{1}{q} = 1. \tag{2.4}$$

Next, for any functional $\mathcal{F}$ and its convex conjugate $\mathcal{F}^*$, the Fenchel inequality holds

$$\langle v^*, v \rangle_{V^* \times V} \leq \mathcal{F}^*(v^*) + \mathcal{F}(v), \quad \forall v^* \in V^*, \quad \forall v \in V. \tag{2.5}$$

When last two are used in combination, they are referred as Young-Fenchel inequality.

The Hölder inequality for integrable functions reads as

$$\int_\Omega u\, v \, \mathrm{d}x \leq \|u\|_{L^p} \|v\|_{L^q}, \quad \forall u \in L^p(\Omega),\ \forall v \in L^q(\Omega), \quad \tfrac{1}{p} + \tfrac{1}{q} = 1. \tag{2.6}$$

For $p = q = 2$, it is referred as the Cauchy-Bunyakowski-Schwarz inequality.

We recall the main inequalities from the embedding theory. First, Friedrichs' inequality [52] has the form

$$\|u\|_\Omega \le C_{\mathrm{F}\Omega}\|\nabla u\|_\Omega, \quad \forall u \in H_0^1(\Omega).$$

The Poincaré inequality [114, 115] reads as

$$\|u\|_\Omega \le C_{\mathrm{P}\Omega}\,\|\nabla u\|_\Omega, \quad \forall u \in \widetilde{H}^1(\Omega), \tag{2.7}$$

where $\widetilde{H}^1(\Omega) := \{u \in H^1(\Omega) \,|\, \{u\}_\Omega = 0\}$, where $\{u\}_\Omega := \frac{1}{|\Omega|}\int\limits_\Omega w\,\mathrm{d}x$. The above-introduced constants $0 < C_{\mathrm{F}\Omega} := \frac{1}{\sqrt{\lambda_1^D}} < +\infty$ and $0 < C_{\mathrm{P}} := \frac{1}{\sqrt{\lambda_2^N}} < +\infty$, where λ_1^D is the first eigenvalue of the Dirichlet-Laplacian and λ_2^N is the second eigenvalue of the Neumann-Laplacian. Due to inequality $0 < \lambda_{n+1}^N < \lambda_n^D$ for all $n \in \mathbb{N}$ (see [49]), the relation $C_{\mathrm{F}\Omega} < C_{\mathrm{P}\Omega}$ holds. According to [97], for simple bounded domain in $\mathbb{R}^d$ encompassed inside a rectangle with edges of length l_i, $i = 1, \dots, d$, we have the estimate $C_{\mathrm{F}\Omega} \le \frac{1}{\pi}\Big(\sum\limits_{i=1}^{d} l_i^{-2}\Big)^{-1/2}$. The Poincaré constant $C_{\mathrm{P}\Omega}$ can be estimated as $C_{\mathrm{P}\Omega} \le \frac{\mathrm{diam}\Omega}{\pi}$ for convex domain Ω (see [106]). For simplexes in $\mathbb{R}^2$, this estimate was improved in [86], where it was shown that $C_{\mathrm{P}\Omega} \le \frac{\mathrm{diam}\Omega}{j_{1,1}}$ for all nondegenerated triangles, and

$$C_{\mathrm{P}\Omega} \le \overline{C}_{\mathrm{T}}^{LS} := \mathrm{diam}\Omega \cdot \begin{cases} \frac{1}{j_{1,1}} & \alpha \in (0, \frac{\pi}{3}], \\ \min\Big\{\frac{1}{j_{1,1}}, \frac{1}{j_{0,1}}\big(2(\pi-\alpha)\tan(\alpha/2)\big)^{-1/2}\Big\} & \alpha \in (\frac{\pi}{3}, \frac{\pi}{2}], \\ \frac{1}{j_{0,1}}\big(2(\pi-\alpha)\tan(\alpha/2)\big)^{-1/2} & \alpha \in (\frac{\pi}{2}, \pi] \end{cases}$$

for isosceles one. Here, $j_{0,1} \approx 2.4048$ and $j_{1,1} \approx 3.8317$ are the smallest positive roots of the Bessel functions J_0 and J_1, respectively.

Exact value of constant in (2.7) on equilateral triangle with unit side is derived in [113], i.e., $C_\Gamma^{\mathrm{P}} = \frac{3}{4\pi}$. Constants for the right isosceles triangles with legs $\frac{\sqrt{2}}{2}$ and 1 are $C_\Gamma^{\mathrm{P}} = \frac{1}{\sqrt{2}\pi}$ and $C_\Gamma^{\mathrm{P}} = \frac{1}{\pi}$, respectively. The latter one can be found from [64] and [71]. Explicit formulas of the same constants for some three-dimensional domains can be found in papers [18] and [64].

The Poincaré-type inequalities also hold for functions $w \in \widetilde{H}^1(\Omega, \Gamma) := \{u \in H^1(\Omega) \,|\, \{u\}_\Gamma = 0\}$, where $\{u\}_\Gamma := \frac{1}{|\Gamma|}\int\limits_\Gamma w\,\mathrm{d}s$, i.e.,

$$\|u\|_{L^2(\Omega)} \le C_\Gamma^{\mathrm{P}}\|\nabla u\|_{L^2(\Omega)}, \tag{2.8}$$

$$\|u\|_{L^2(\Gamma)} \le C_\Gamma^{\mathrm{Tr}}\|\nabla u\|_{L^2(\Omega)}. \tag{2.9}$$

The exact values of C_Γ^{P} and C_Γ^{Tr} on right triangles, rectangles, and parallelepipeds can be found in [99]. We consider below mainly two reference cases in $\mathbb{R}^2$: triangle $\mathrm{T} := \mathrm{conv}\{(0,0),(0,h),(h,0)\}$ and $\Gamma := \{x_2 = 0,\, x_1 \in [0,h]\}$, and cor-

responding constants $C_\Gamma^{\mathrm{P}} := \frac{h}{\zeta_0}$, and $C_\Gamma^{\mathrm{Tr}} := \left(\frac{h}{\hat{\zeta}_0 \tanh(\hat{\zeta}_0)}\right)^{1/2}$, where ζ_0 and $\hat{\zeta}_0$ are the unique roots of the equations $z\cot(z)+1=0$ and $\tan(z)+\tanh(z)=0$ in $(0,\pi)$, respectively, and simplex $\mathrm{T} := \mathrm{conv}\{(0,0),(0,h),(\frac{h}{2},\frac{h}{2})\}$, with $\Gamma := \{x_2 = 0,\, x_1 \in [0,h]\}$, which are characterized by $C_\Gamma^{\mathrm{P}} := \frac{h}{2\zeta_0}$ and $C_\Gamma^{\mathrm{Tr}} := \left(\frac{h}{2}\right)^{1/2}$.

Finally, the classic trace inequality reads as follows

$$\|u\|_{L^2(\Gamma)} \le C_{\mathrm{TT}} \| u \|_{H^1(\Omega)}, \qquad \forall u \in C^1(\overline{\Omega}). \tag{2.10}$$

2.1.5 Sobolev spaces in the space-time cylinder

Let $Q_T := \Omega \times]0,T[$ denote the space-time cylinder with given Ω and time interval $]0,T[$, $0 < T < +\infty$. We denote $S_T = \partial\Omega \times [0,T]$ as a lateral surface of Q_T. Below, we introduce the Sobolev spaces of functions defined on Q_T as they are presented in [79, 80]. The space $L^2(Q_T)$ contains square-integrable functions in the cylinder Q_T and it is equipped with the norm $\|\cdot\|_{L^2(Q_T)} := (\cdot,\cdot)^{1/2}_{L^2(Q_T)}$. We generalize the notation by denoting the space $H^{s,k}(Q_T)$ as

$$H^{s,k}(Q_T) := \Big\{ u \in L^2(Q_T) \;\mid\; D^\alpha u \in L^2(Q_T),\ |\alpha| \le s,\ \partial_t^\beta u \in L^2(Q_T),\ 1 \le \beta \le k \Big\}$$

equipped with the norm

$$\|u\|^2_{H^{s,k}(Q_T)} := \int\limits_{Q_T} \left(\sum_{|\alpha|\le s} |D^\alpha u(x,t)|^2 + \sum_{1\le\beta\le k} |\partial_t^\beta u(x,t)|^2 \right) \mathrm{d}x\mathrm{d}t.$$

The most typical examples are $H^{1,0}(Q_T)$ and $H^{1,1}(Q_T)$. In [79], the same spaces are denoted by, e.g., $W_2^{1,0}(Q_T)$ and $W_2^{1,1}(Q_T)$. Furthermore, the Sobolev spaces with Dirichlet boundary $S_D \subset S_T$ (with assigned load u_D on it) are denoted by

$$H^{s,k}_{u_D}(Q_T) := \Big\{ u \in H^{s,k}(Q_T) \;\mid\; u = u_D \quad \text{on} \quad S_D \Big\}. \tag{2.11}$$

2.2 Bochner spaces

Consider the Bochner spaces as an alternative tool for the analysis of parabolic I-BVPs. Let $\{H,(\cdot,\cdot)_H\}$ and $\{V,(\cdot,\cdot)_V\}$ be a Hilbert space. The Bochner spaces $L^p(a,b;V)$, $p \in [1,+\infty[$, are the most regularly used. They consist of measurable functions $u :]a,b[\to V$ for which norm reads as

$$\|u\|_{L^p(a,b;V)} := \Big(\int\limits_a^b \|u(\cdot,t)\|_V^p \,\mathrm{d}t \Big)^{1/p} < +\infty.$$

For $p = \infty$, we obtain the Bochner space equipped with the norm

$$\|u\|_{L^\infty(a,b;V)} := \operatorname*{ess\,sup}_{t\in(a,b)} \|u(\cdot,t)\|_V < +\infty.$$

Furthermore, we define $C([a,b];H)$ as the space of functions $u : [a,b] \to H$ continuous at every $t \in [a,b]$ with the norm

$$\|u\|_{C([a,b];H)} := \max_{t\in[a,b]} \|u(\cdot,t)\|_H.$$

Infinitely differentiable functions are denoted by $C^\infty([a,b];H)$ and $C_0^\infty([a,b];H)$ (in case the functions have compact support on (a,b)).

For the treatment of parabolic I-BVPs, we consider $L^2(0,T;V)$ with $V = H^1(\Omega)$ (or $V = H_0^1(\Omega)$). Since V is a Hilbert space, then $L^2(0,T;V)$ is also a Hilbert space. The generalized weak derivative of $u \in L^2(0,T;V)$ with respect to time is denoted by $\partial_t u \in L^2(0,T;V^*)$, satisfying

$$\int_0^T u(t)\partial_t\varphi(t)\,\mathrm{d}t = -\int_0^T \partial_t u(t)\varphi(t)\,\mathrm{d}t, \quad \forall\varphi \in C_0^\infty([0,T];H).$$

For separable V and H, the Gelfand triple (or evolution triple) $V \hookrightarrow H \hookrightarrow V^*$ holds. Then, V^* is a Hilbert space. The most commonly used triples are $H^1(\Omega) \hookrightarrow L^2(\Omega) \hookrightarrow (H^1(\Omega))^*$ and $H_0^1(\Omega) \hookrightarrow L^2(\Omega) \hookrightarrow H^{-1}(\Omega)$.

To study the solvability of the parabolic I-BVPs, we define the Bochner space $W(0,T) := \left\{u(t) \in L^2(0,T;V) \,|\, \partial_t u(t) \in L^2(0,T;V^*)\right\}$ equipped with the norm

$$\|u\|_{W(0,T)} := \left(\int_0^T \left(\|u(\cdot,t)\|_V^2 + \|\partial_t u(\cdot,t)\|_{V^*}^2\right)\mathrm{d}t\right)^{1/2} < \infty.$$

The Gelfand triple implies that $W(0,T)$ is a Hilbert space. Moreover, we have the continuous embedding $W(0,T) \hookrightarrow C(0,T;H)$ (see, e.g., [148] and [151]). The formula of integration by parts reads as

$$\int_0^T \langle\partial_t u(t),\varphi(t)\rangle_{V^*,V}\,\mathrm{d}t = -\int_0^T \langle\partial_t\varphi(t),u(t)\rangle_{V^*,V}\,\mathrm{d}t + (u(T),\varphi(T)) - (u(0),\varphi(0)),$$

where $\partial_t u(t) \in L^2(0,T;V^*)$, $\varphi(t) \in L^2(0,T;V)$, and $\partial_t\varphi(t) \in L^2(0,T;V^*)$.

By comparing the norms of the spaces discussed above, one can see how Bochner spaces correlate with Sobolev spaces, e.g.,

$$H^{1,0}(Q_T) \cong L^2(0,T;H^1(\Omega)), \quad H_0^{1,0}(Q_T) \cong L^2(0,T;H_0^1(\Omega)).$$

If, in addition, we consider the space $H^1(0,T;L^2(\Omega))$ with finite norm

$$\|u\|_{H^1(0,T;L^2(\Omega))} := \left(\int_0^T \left(\|u(\cdot,t)\|^2_{L^2(\Omega)} + \|\partial_t u(\cdot,t)\|^2_{L^2(\Omega)} \right) \mathrm{d}t \right)^{1/2},$$

then the combination of the norms corresponding to spaces $H^{1,1}(Q_T)$ and $H_0^{1,1}(Q_T)$ (in some literature denoted by $H^1(Q_T)$ and $H_0^1(Q_T)$) provides the equivalences

$$H^{1,1}(Q_T) \cong L^2(0,T;H^1(\Omega)) \cap H^1(0,T;L^2(\Omega)),$$
$$H_0^{1,1}(Q_T) \cong L^2(0,T;H_0^1(\Omega)) \cap H^1(0,T;L^2(\Omega)).$$

Bochner space $W(0,T)$ with $V = H^1(\Omega)$ ($V = H_0^1(\Omega)$) is clearly wider than $H^{1,1}(Q_T)$ ($H_0^{1,1}(Q_T)$) based on evolution triple. Finally, we introduce, in general form, $V^{s,k}(Q_T)$ and $V_0^{s,k}(Q_T)$ (following the notation in [79]) such that

$$V^{s,k}(Q_T) := H^{s,k}(Q_T) \cap C([0,T];L^2(\Omega)),$$

and

$$V_0^{s,k}(Q_T) := H_0^{s,k}(Q_T) \cap C([0,T];L^2(\Omega)).$$

respectively, where $s \geq 0$, $k \geq 0$, equipped with the norm

$$\|u\|_{V^{s,k}(Q_T)} := \max_{t\in[0,T]} \|u(t)\|_{L^2(\Omega)} + \|u\|_{H^{s,k}(Q_T)} < +\infty.$$

2.3 Parabolic initial-boundary value problem

In the current section, we present fundamental results on solvability of linear parabolic PDEs, which have been thoroughly studied in monographs [79, 51, 151, 148]. The nonlinear class is considered in the monographs [80, 152]. Below, we present the variational formulation of a parabolic I-BVP and discuss the main requirements that provide the existence and uniqueness results.

Let Q_T be a space-time cylinder with boundary surface S_T as defined in Section 2.1. Assume that $\partial\Omega$ consists of two measurable non-intersecting parts Γ_D and Γ_R associated with mixed Dirichlet–Robin BC. Therefore, $S_T := \partial\Omega \times [0,T] = \left(\Gamma_D \cup \Gamma_R\right) \times [0,T] = S_D \cup S_R$. The general parabolic I-BVP reads as follows

$$u_t - \operatorname{div} p + a(x) \cdot \nabla u + \lambda^2(x)\, u = f, \qquad (x,t) \in Q_T, \tag{2.12}$$
$$p = A\nabla u, \qquad (x,t) \in Q_T, \tag{2.13}$$
$$u(x,0) = u_0, \qquad x \in \Omega, \tag{2.14}$$
$$u = 0, \qquad (x,t) \in S_D, \tag{2.15}$$
$$\sigma^2(x) u + p \cdot n = 0, \qquad (x,t) \in S_R, \tag{2.16}$$

where n denotes the vector of unit outward normal to $\partial\Omega$,

$$f \in L^2(Q_T), \quad u_0 \in L^2(\Omega). \tag{2.17}$$

We assume that, for almost all $x \in \Omega$ and $t \in]0,T[$, the operator A is symmetric and satisfies condition of uniform parabolicity

$$\underline{\nu}_A|\xi|^2 \le A(x,t)\,\xi\cdot\xi \le \overline{\nu}_A|\xi|^2, \quad \xi \in \mathbb{R}^d, \quad 0 < \underline{\nu}_A \le \overline{\nu}_A < \infty. \tag{2.18}$$

Henceforth, we use the notation

$$\|\,\tau\,\|_A^2 := \int\limits_\Omega A\tau\cdot\tau\,\mathrm{d}x, \quad \|\,\tau\,\|_{A^{-1}}^2 := \int\limits_\Omega A^{-1}\tau\cdot\tau\,\mathrm{d}x.$$

The functions a and λ, presenting the convection and reaction, respectively, as well as σ satisfy the following conditions for a.a $t \in]0,T[$

$$\begin{aligned} a \in L^\infty(\Omega,\mathbb{R}^d), \quad \operatorname{div} a \in L^\infty(\Omega), \quad |a| \le \overline{a},\\ \lambda \in L^\infty(\Omega), \quad |\lambda| \le \overline{\lambda},\\ \sigma \in L^\infty(\Omega), \quad |\sigma| \le \overline{\sigma}. \end{aligned} \tag{2.19}$$

After multiplying (2.12) by a test function $\eta \in H_0^{1,1}(Q_T)$, we arrive at the generalized formulation of (2.12)–(2.16): find $u(x,t) \in V_0^{1,0}(Q_T)$ (cf. (2.11)) satisfying the integral identity

$$\begin{aligned} \int\limits_{Q_T}\Big(A\nabla u\cdot\nabla\eta + a\cdot\nabla u\eta + \lambda^2 u\eta - u\eta_t\Big)\,\mathrm{d}x\mathrm{d}t + \int\limits_{S_R}\sigma^2 u\eta\,\mathrm{d}s\mathrm{d}t\\ + \int\limits_\Omega\big((u\eta)(x,T) - (u\eta)(x,0)\big)\,\mathrm{d}x = \int\limits_{Q_T} f\eta\,\mathrm{d}x\mathrm{d}t, \quad \forall\eta \in H_0^{1,1}(Q_T). \end{aligned} \tag{2.20}$$

According to [79, Theorem 3.2], the generalized problem (2.20) has a solution in $V_0^{1,0}(Q_T)$ and it is unique in $H_0^{1,0}(Q_T)$, provided that conditions (2.17), (2.18), and (2.19) hold. In the problem with only Robin BC, in order to provide the uniqueness of the solution additional conditions on coefficients

$$|\partial_t a| \le \widetilde{a}, \quad |\partial_t\lambda| \le \widetilde{\lambda}, \quad |\partial_t\sigma| \le \widetilde{\sigma},$$

must be imposed. The a priori stability estimate

$$\|u\|_{V_0^{1,0}(Q_T)} \le C\Big(\|f\|_{L^2(Q_T)} + \|u_0\|_{L^2(\Omega)}\Big) \tag{2.21}$$

holds with a positive constant C dependent only on characteristics of Q_T but independent of f and u. The estimate (2.21) provides the continuity of the mapping $\mathcal{M} : \{f, u_0\} \mapsto u$, where $\mathcal{M} : L^2(Q_T) \times L^2(\Omega) \mapsto V_0^{1,0}(Q_T)$.

The solvability results can be formulated in Bochner spaces. We consider

the simplest case, where $A = I$, $a(x) = 0$, $\lambda(x) = 0$, and $S_T = S_D$. According to [148, 151], if H and V are given separable Hilbert spaces satisfying evolution triple $V \hookrightarrow H \hookrightarrow V^*$, $f \in L^2(0,T,V^*)$, $u_0 \in H$, then the generalized problem

$$\int_0^T \langle u_t(t), v\rangle_{V^*,V}\,\mathrm{d}t + \int_0^T \nabla u(t)\cdot\nabla v\,\mathrm{d}x\mathrm{d}t = \int_0^T \langle f(t), v\rangle_{V^*,V}\,\mathrm{d}x\mathrm{d}t,$$

that holds for all $v \in V$ and a.a.$t \in]0,T[$, has a unique solution in $W(0,T)$, which depends continuously on f and u_0. By increasing the regularity on u_0 and f, one can get higher regularity of the exact solution. The problems with inhomogeneous BCs, e.g., u_D in (2.15) and g in (2.16), can be treated in the same manner, following the spirit of [148].

2.4 Fixed point iterations

First, we present the main idea of the fixed-point iterations approach. Consider the following general problem: find u in a Hilbert space V such that

$$u = \mathcal{L}u + b, \tag{2.22}$$

where $\mathcal{L} : V \to V$ is a bounded operator and $b \in V$. One of the ways to solve (2.22) is to apply the iteration procedure

$$u_k = \mathcal{L}u_{k-1} + b, \quad u_0 \in V, \quad k = 1,\ldots$$

which generates an infinite sequence $\{u_k\}_{k=1}^{\infty}$. If $\mathcal{L}$ is the *q-contractive operator* on a closed nonempty set $S \subset V$, i.e.,

$$\|\mathcal{L}w - \mathcal{L}v\|_V \le q\,\|w - v\|_V, \quad q \in (0,1), \quad \forall w, v \in S, \tag{2.23}$$

then, by using (2.23), it is easy to show that $\{u_k\}_{k=1}^{\infty}$ converges to a fixed point u (see, e.g., [13, 34, 74, 67, 150]).

2.4.1 The Picard–Lindelöf method

We consider a Cauchy problem

$$\tfrac{du}{dt} = \varphi(u(t),\ t), \quad u(t_0) = a_0, \quad t \in [t_0 - \varepsilon, t_0 + \varepsilon] \tag{2.24}$$

with (scalar- or vector-valued) solution $u(t)$. Assume that the function $\varphi(u(t),\ t)$ is uniformly Lipschitz continuous with respect to u (i.e., Lipschitz constant can be selected independent of t) and continuous in t. The existence and uniqueness of continuously differentiable $u(t)$ on $[t_0 - \varepsilon, t_0 + \varepsilon]$, $\forall \varepsilon > 0$, follows from the Picard–Lindelöf theorem and the Picard's existence theorem (or Cauchy–Lipschitz the-

orem) (see [33, 88]). Unlike the Picard–Lindelöf theorem, the Peano existence theorem [107] shows only existence, not uniqueness, but imposes weaker requirements on φ (only the continuity with respect to t).

The Picard–Lindelöf method represents (2.24) in the integral form

$$u(t) = \int_{t_0}^{t} \varphi(u(s),\ s)\,\mathrm{d}s + a_0. \tag{2.25}$$

The exact solution of (2.25) is a fixed point that is approximated by the iterative method

$$u_j = \mathcal{T}u_{j-1} + a_0, \quad \mathcal{T}u := \int_{t_0}^{t} \varphi(u(s), s)\,\mathrm{d}s \tag{2.26}$$

provided that $\mathcal{T} : V \to V$ satisfies (2.23) on $[t_0 - \varepsilon, t_0 + \varepsilon]$.

3 MAIN RESULTS

This chapter is devoted to the main theoretical and numerical findings obtained during the PhD studies. Along the exposition of the results, we refer to the published works [PI, PII, PIII, PIV] and preprint [PV], where corresponding matters are thoroughly discussed.

3.1 Fully reliable Adaptive Picard–Lindelöf method

In this section, we make an overview of the work dedicated to the fully reliable APL method which was suggested in [PI] in order to reliably solve the Cauchy problem. The details of the study can also be found in [93, Section 6.7.6].

Let $Q := \{(u,\ t) \mid u \in \mathrm{U},\ t \in I\}$, where U is the set of possible values of u determined during an a priori analysis of the problem and $I := [t_0, t_K]$. We consider the problem (2.24) from Section 2.4 and assume that function $\varphi(u(t),\ t)$ is continuous with respect to both variables and satisfies the Lipschitz condition for any $(u_1,\ t_1), (u_2,\ t_2) \in Q$ in the form

$$\|\varphi(u_2,\ t_2) - \varphi(u_1,\ t_1)\|_{C([t_1,\ t_2])} \le \mathrm{L}_1 \|u_2 - u_1\|_{C([t_1,\ t_2])} + \mathrm{L}_2 |t_2 - t_1|,$$

where L_1 and L_2 are Lipschitz constants.

Assume that I is discretized in the following way:

$$I = \cup_{I^{(k)} \subset \mathcal{F}_K} \overline{I^{(k)}}, \quad \mathcal{F}_K := \{I^{(k)}\}_{k=0}^{K-1}, \quad I^{(k)} := (t_k,\ t_{k+1}), \quad K \in \mathbb{N}. \tag{3.1}$$

If we consider (2.26), it becomes clear that condition $q := \mathrm{L}_1(t_{k+1} - t_k) < 1$ provides the convergence of the algorithm by adapting the length of interval $I^{(k)} \subset \mathcal{F}_K$ to constant L_1. Thus, if $I^{(k)}$ is sufficiently small, the solution can be reconstructed by an iteration scheme, which we call an Adaptive Picard–Lindelöf (APL) method. The corresponding errors of the iterative approximations can be

controlled by the Ostrowski estimates

$$\underline{\mathrm{M}}_j := \tfrac{1}{1+q}\|u_j - u_{j+1}\|_{C(I^{(k)})} \le \|u - u_j\|_{C(I^{(k)})} \le \tfrac{q}{1-q}\|u_j - u_{j-1}\|_{C(I^{(k)})} =: \overline{\mathrm{M}}_j.$$

The latter one is applicable to any iterative process with a contraction operator that possesses the computable contractivity parameter, for example to the iteration algorithm provided in work [55].

However, some technical difficulties arising in iterative integration must be dealt with. Consider $I^{(k)} \in \mathcal{F}_K$ introduced in (3.1) and assume that the initial guess u_0 is defined as a piecewise affine function on a sub-mesh Ω_{S_k} of $I^{(k)}$, i.e., $\Omega_{S_k} = \cup_{s=0}^{S_k-1}[z_s, z_{s+1}]$, where $\Delta_s = z_{s+1} - z_s$, $z_0 = t_k$, and $z_{S_k} = t_{k+1}$. As the first sub-interval, we have

$$u_1(t) = \int_{t_0}^{t} \varphi(u_0(s),\ s)\,\mathrm{d}s + a_0, \quad t \in I^{(0)} := [t_0, t_1]. \tag{3.2}$$

If $q < 1$, the distance between the computed u_1 and u can be found by means of

$$\|u_1(t) - u(t)\|_{C(I^{(0)})} \le \tfrac{q}{1-q}\|u_1(t) - u_0(t)\|_{C(I^{(0)})}. \tag{3.3}$$

However, in (3.2) we obtain piecewise polynomials as a result of the integration of piecewise affine functions. In order to perform iterations on a finite dimensional

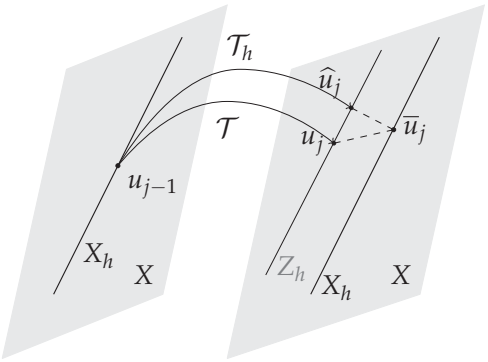


FIGURE 2 Integration and interpolation errors generated by $\mathcal{T}$.

space X_h, the additional errors caused by integration and mapping of a function to this finite dimensional space must be taken into account (see Figure 2). Due to the numerical representation of $\mathcal{T}$, i.e., $\mathcal{T}_h :\ X_h \to Z_h$, where $Z_h \subset X$, the function $\widehat{x}_j = \mathcal{T}_h x_{j-1}$ contains an integration error. Since $Z_h \subset X$ does not coincide with X_h, we must apply a certain projection (interpolation) operator π and evaluate the corresponding error. Henceforth, the errors generated by numerical integration appear in (3.3) as follows:

$$\|u_1 - u_0\|_{C(I^{(0)})} \le\ \|u_0 - \widehat{u}_1\|_{C(I^{(0)})} + \|\widehat{u}_1 - u_1\|_{C(I^{(0)})}. \tag{3.4}$$

Here, $\|\widehat{u}_1 - u_1\|_{C(I^{(0)})} := \|\widehat{e}_1\|_{C(I^{(0)})}$ is the *integration error*. Then we must project the result of numerical integration $\widehat{u}_1 \in Z_h$ to X_h, i.e., $\overline{u}_1(t) = \pi\widehat{u}_1 \in CP^1(I^{(0)})$, where $\pi : Z_h \to CP^1(I^{(0)})$ is the projection operator such that $\pi\widehat{u}(z_s) = \overline{u}(z_s)$, $s = 0, \ldots, S_k$-1. Thus, the RHS of (3.4) is modified as follows

$$\|u_1 - u_0\|_{C(I^{(0)})} \le \|\overline{u}_1 - u_0\|_{C(I^{(0)})} + \|\widehat{u}_1 - u_1\|_{C(I^{(0)})} + \|\widehat{u}_1 - \overline{u}_1\|_{C(I^{(0)})},$$

Here, $\|\widehat{u}_1(t) - \overline{u}_1(t)\|_{C(I^{(0)})} =: \|\bar{e}_1\|_{C(I^{(0)})}$ is the *interpolation error*.

The obtained result states that for any piecewise linear approximation $v(t) := v^k(t)$, $t \in I^{(k)}$, and exact solution $u(t)$ the following estimate

$$\|u(t) - v(t)\|_{C(I^{(k)})} \le \overline{\mathrm{M}}^k, \quad t \in I^{(k)}, \quad I^{(k)} \subset \mathcal{F}_K,$$

holds. Here, the piecewise constant error bound reads as

$$\overline{\mathrm{M}}^k := \tfrac{q}{1-q}(\|\overline{v}_{j+1} - \overline{v}_j\|_{C(I^{(k)})} + e_j^{\mathrm{int}} + e_j^{\mathrm{interp}}).$$

where

$$e_j^{\mathrm{int}} := \sum_{s=0,\ldots,S_k-1} \left(\tfrac{\mathrm{L_s}}{2}\Delta_s^2 - \tfrac{1}{2\mathrm{L_s}}\Big[\varphi(\overline{v}_{j,\,s+1}, z_{s+1}) - \varphi(\overline{v}_{j,\,s}, z_s)\Big]^2 \right), \tag{3.5}$$

and

$$e_j^{\mathrm{interp}} := \sum_{s=0,\ldots,S_k-1} \Delta_s \Bigg(\tfrac{1}{8}\big(\varphi(\overline{v}_{j,\,s+1}, z_{s+1}) - \varphi(\overline{v}_{j,\,s}, z_s)\big) + \tfrac{2}{3}\Big[\mathrm{L}_{1,\,s}|\overline{v}_{j,\,s+1} - \overline{v}_{j,\,s}| + \mathrm{L}_{2,\,s}\Delta_s\Big] \Bigg) \tag{3.6}$$

are integration and interpolation estimates of $\|\widehat{e}_j\|_{C(I^{(k)})}$ and $\|\bar{e}_j\|_{C(I^{(k)})}$ on the j^{th} iteration. Constants in (3.5) and (3.6) are $\mathrm{L_s} = \mathrm{L}_{1,\,s}\, l_s + \mathrm{L}_{2,\,s}$, where l_s is the slope of a piecewise function on every interval $[z_s,\ z_{s+1}]$, $s = 0, ..., S_k$-1, and local Lipschitz constant $\mathrm{L}_{1,\,s}$ analogous to the one in (3.1). In [PI] and [93, Section 6.7.6], we present detailed derivation of estimates for both errors, and confirm the theoretical findings by set of numerical examples.

3.2 Guaranteed error estimates for the solution of parabolic I-BVPs

This section presents two forms of the functional error estimates, which provide a guaranteed upper bound of the deviation $e = u - v$ for the generalized solution u of I-BVP (2.20) with $a \equiv 0$ and any function $v \in H_0^{1,1}(Q_T)$ (generated for instance

by some numerical method) measured in terms of the norm

$$[\,e\,]^2_{(\nu,\theta,\zeta,\chi)} := \int_0^T \left(\nu \,\|\, \nabla e \,\|^2_A \,+\, \theta \,\|\, \lambda\, e\|^2_\Omega \,+\, \chi\|\, \sigma\, e \,\|^2_{\Gamma_R}\right) \mathrm{d}t + \zeta \,\|\, e(\cdot,T) \,\|^2_\Omega, \quad (3.7)$$

where ν, θ, ζ, χ are positive weights and function λ satisfies (2.19). By selecting the weights to balance the components in (3.7) with a desired proportion, we generate a collection of error measures, which can be used for judging the distance between u and v.

The first form of the majorant is presented and numerically tested for evolutionary reaction-diffusion I-BVPs of parabolic type in [PII]. The second (advanced) form of the majorant is studied in [PIII] and [PIV]. Latter one was introduced originally in publication of Repin [124] in order to improve the recovery of the error in balance equation (2.12) by using a special correction function w (see Theorem 3.2). In [PIII], we extend both majorants for problems formulated on domains of complicated geometry by suggesting a method of decomposition of Ω and application of the Poincaré inequalities locally to each element from the collection of subsets. This method does not only help to overcome the complications caused by estimating the global Friedrichs constant but also improves the efficiency of the resulting estimates which exploit the constant on the smaller sub-domains. In [PIV], we encounter difficulties caused by the mixed BC with a non-trivial input function and overcome them by exploiting the Poincaré-type inequalities. The obtained estimates become fully guaranteed, due to results of [99] and [PV], where reliable and easy computable bounds for the constants in the Poincaré-type inequalities are presented for simpleces in $\mathbb{R}^2$ and $\mathbb{R}^3$ (commonly used in FE analysis).

The initial step in the derivation of both upper estimates is the transformation of (2.20) into the integral identity

$$\int_0^T \Big(\|\, \nabla e \,\|^2_A \,+\, \|\, \lambda\, e\|^2_\Omega \,\mathrm{d}t \,+\, \|\, \sigma\, e \,\|^2_{\Gamma_R} \Big)\, \mathrm{d}t \,+\tfrac{1}{2} \,\|\, e(\cdot,T) \,\|^2_\Omega$$
$$= \int_{Q_T} \Big(\big(f - v_t - \lambda^2\, v\big)\, e - A\nabla v \cdot \nabla e \Big)\, \mathrm{d}x\mathrm{d}t + \int_{S_R} -\sigma^2 v\, e\, \mathrm{d}s\mathrm{d}t, \quad (3.8)$$

which follows from the main energy-balance relation of problem (2.12)–(2.16). It is worth mentioning that evolutionary I-BVPs, unlike elliptic BVPs, do not possess variational formulation, therefore the functional error estimates can only be obtained from generalized identity (2.20). Next, we rearrange the RHS of (3.8) by introducing a 'free' vector-valued function

$$y \in Y_{\mathrm{div}}(Q_T) := \Big\{ y \in L^2\big(0,T;L^2\big(\Omega,\mathbb{R}^d\big)\big) \;\Big|\; \mathrm{div} y \in L^2\big(0,T;L^2(\Omega)\big), \; y\cdot n \in L^2\big(0,T;L^2(\Gamma_R)\big) \Big\}.$$

The *residuals* of (2.12), (2.13), and (2.16) are denoted by

$$\mathbf{r}_f(v,y) := f - v_t - \lambda^2 v + \operatorname{div} y, \tag{3.9}$$
$$\mathbf{r}_A(v,y) := y - A\nabla v, \tag{3.10}$$
$$\mathbf{r}_\sigma(v,y) := -\sigma^2 v - y\cdot n, \tag{3.11}$$

respectively. Moreover, we define the weighted residuals

$$\mathbf{r}_f^{\mu}(v,y) := \mu\,\mathbf{r}_f \quad \text{and} \quad \mathbf{r}_f^{1-\mu}(v,y) := (1-\mu)\,\mathbf{r}_f, \tag{3.12}$$

where $\mu(x,t)$ is a real-valued function taking values in $[0,1]$ used in order to split the residual with λ into two parts. This way, the resulting estimate becomes robust to cases in which λ attains drastically different values and may be close to zero in different parts of Ω. A detailed numerical analysis of the majorant with the balancing parameter μ can be found in [PII, Sections 2, 5]. The forthcoming summary demonstrates that a certain weighted combination of norms of (3.9)–(3.12) controls the distance between u and v.

Theorem 3.1. *For any $v \in V_0^{1,1}(Q_T)$, $y \in Y_{\operatorname{div}}(Q_T)$, $\delta \in (0,2]$, and real-valued function $\gamma(t) \in \left[\frac{1}{2}, +\infty\right[$, we have the estimate*

$$\begin{aligned}[e]^2_{(\nu,\theta,1,2)} \le \overline{\mathrm{M}}_{\mathrm{I}}^2(v,y;\delta,\gamma,\mu) := \|e(\cdot,0)\|^2_\Omega + \int_0^T \Big(\gamma(t) \left\| \tfrac{1}{\lambda}\,\mathbf{r}_f^{\mu} \right\|^2_\Omega \\ + \alpha_1 \|\,\mathbf{r}_A\,\|^2_{A^{-1}} + \alpha_2 \tfrac{C_{\mathrm{F}\Omega}^2}{\underline{\nu}_A} \left\| \mathbf{r}_f^{1-\mu} \right\|^2_\Omega + \alpha_3 \tfrac{\widetilde{C}^2_{\mathrm{T}\Gamma_R}}{\underline{\nu}_A} \|\,\mathbf{r}_\sigma\,\|^2_{\Gamma_R} \Big)\,\mathrm{d}t, \end{aligned} \tag{3.13}$$

where

$$\widetilde{C}_{\mathrm{T}\Gamma_R} = C_{\mathrm{T}\Gamma_R}\,(1 + C_{\mathrm{F}\Omega}) \tag{3.14}$$

with Friedrichs' and trace constants in (2.1) and (2.10), respectively, positive parameters $\nu = 2\text{-}\delta$, $\theta(x,t) = \lambda(x)\Big(2 - \frac{1}{\gamma(t)}\Big)^{1/2}$, $\mu(x,t) \in [0,1]$, and $\alpha_1(t)$, $\alpha_2(t)$, $\alpha_3(t)$ are arbitrary positive real-valued functions satisfying the relation

$$\tfrac{1}{\alpha_1(t)} + \tfrac{1}{\alpha_2(t)} + \tfrac{1}{\alpha_3(t)} = \delta. \tag{3.15}$$

Proof. We present below only sketch of the proof. First, the RHS of (3.8) is transformed by means of function $y \in Y_{\operatorname{div}}(Q_T)$, divergence theorem, and integration by parts, resulting into

$$\int_0^T \Big(\|\nabla e\|^2_A + \|\,\lambda\, e\|^2_\Omega + \|\,\sigma\, e\,\|^2_{\Gamma_R} \Big)\,\mathrm{d}t + \tfrac{1}{2}\,\|\,e(\cdot,T)\,\|^2_\Omega = \mathscr{I}_f + \mathscr{I}_A + \mathscr{I}_\sigma + \tfrac{1}{2}\|\,e(\cdot,0)\|^2_\Omega,$$

where

$$\mathscr{I}_f := \int\limits_{Q_T} \mathbf{r}_f\, e\, \mathrm{d}x\mathrm{d}t, \quad \mathscr{I}_A := \int\limits_{Q_T} \mathbf{r}_A \cdot \nabla e\, \mathrm{d}x\mathrm{d}t, \quad \mathscr{I}_\sigma := \int\limits_{S_R} \mathbf{r}_\sigma\, e\, \mathrm{d}s\mathrm{d}t.$$

Next, we estimate $\mathscr{I}_f$, $\mathscr{I}_A$, and $\mathscr{I}_\sigma$, by using Hölder inequality, as follows

$$\mathscr{I}_A \overset{(2.6)}{\leq} \int\limits_0^T \| \mathbf{r}_A \|_{A^{-1}} \|\nabla e\|_A \, \mathrm{d}t \tag{3.16}$$

$$\mathscr{I}_\sigma \overset{(2.6),\,(2.10)}{\leq} \int\limits_0^T \| \mathbf{r}_\sigma \|_{\Gamma_R} \frac{C_{\Gamma_R}}{\sqrt{\underline{\lambda}_A}} \| \nabla e \|_A \, \mathrm{d}t, \quad \text{and} \tag{3.17}$$

$$\mathscr{I}_f \overset{(2.6),\,(2.1)}{\leq} \int\limits_0^T \left(\left\| \tfrac{1}{\varrho}\, \mathbf{r}_f^{\mu} \right\|_\Omega \| \varrho\, e \|_\Omega + \frac{C_{\mathrm{F}\Omega}}{\sqrt{\underline{\lambda}_A}} \left\| \mathbf{r}_f^{1-\mu} \right\|_\Omega \|\nabla e\|_A \right) \mathrm{d}t, \tag{3.18}$$

respectively. Finally, we bound estimates (3.16), (3.18), and (3.17) using the Young–Fenchel inequality

$$\int\limits_0^T \left\| \tfrac{1}{\varrho}\, \mathbf{r}_f^{\mu} \right\|_\Omega \| \varrho\, e \|_\Omega \, \mathrm{d}t \overset{(2.5)}{\leq} \tfrac{1}{2} \int\limits_0^T \left(\gamma \left\| \tfrac{1}{\varrho}\, \mathbf{r}_f^{\mu} \right\|_\Omega^2 + \tfrac{1}{\gamma} \| \varrho\, e \|_\Omega^2 \right) \mathrm{d}t, \tag{3.19}$$

$$\int\limits_0^T \frac{C_{\mathrm{F}\Omega}}{\sqrt{\underline{\lambda}_A}} \left\| \mathbf{r}_f^{1-\mu} \right\|_\Omega \|\nabla e\|_A \, \mathrm{d}t \overset{(2.5)}{\leq} \tfrac{1}{2} \int\limits_0^T \left(\alpha_1 \frac{C_{\mathrm{F}\Omega}^2}{\underline{\lambda}_A} \left\| \mathbf{r}_f^{1-\mu} \right\|_\Omega^2 + \tfrac{1}{\alpha_1} \| \nabla e \|_A^2 \right) \mathrm{d}t, \tag{3.20}$$

$$\int\limits_0^T \| \mathbf{r}_A \|_{A^{-1}} \|\nabla e\|_A \, \mathrm{d}t \overset{(2.5)}{\leq} \tfrac{1}{2} \int\limits_0^T \left(\alpha_2 \, \| \mathbf{r}_A \|_{A^{-1}}^2 + \tfrac{1}{\alpha_2} \| \nabla e \|_A^2 \right) \mathrm{d}t, \tag{3.21}$$

$$\int\limits_0^T \| \mathbf{r}_\sigma \|_{\Gamma_R} \frac{\widetilde{C}_{\mathrm{T}\Gamma_R}}{\sqrt{\underline{\lambda}_A}} \|\nabla e\|_A \, \mathrm{d}t \overset{(2.5)}{\leq} \tfrac{1}{2} \int\limits_0^T \left(\alpha_3 \frac{\widetilde{C}_{\mathrm{T}\Gamma_R}}{\underline{\lambda}_A} \| \mathbf{r}_\sigma \|_{\Gamma_R}^2 + \tfrac{1}{\alpha_3} \| \nabla e \|_A^2 \right) \mathrm{d}t, \tag{3.22}$$

respectively. Here, $\gamma(t)$, $\alpha_1(t)$, $\alpha_2(t)$, and $\alpha_3(t)$ are defined in the theorem's formulation. Then, the estimate (3.13) follows from a combination of (3.19)–(3.22). □

Next, we consider an advanced form of the error majorant of more complicated structure caused by the introduction of the correction function $w \in V_0^{1,1}(Q_T)$, which yields sharper error bounds. Moreover, in [PIII, PIV] we show that the advanced majorant is equivalent to the error measured in terms of the primal energy norm. In this case, the residuals of (2.13), (2.12), and (2.16) read as

$$\begin{aligned}
\mathbf{r}_f(v, y, w) &:= f - (v + w)_t - \lambda^2 (v - w) + \mathrm{div} y, \\
\mathbf{r}_A(v, y, w) &:= y - A\nabla(v - w), \\
\mathbf{r}_\sigma(v, y, w) &:= -\sigma^2 (v - w) - y \cdot n,
\end{aligned}$$

and $\mathbf{r}_f^{\mu}$ and $\mathbf{r}_f^{1-\mu}$ are defined analogously to (3.12).

Theorem 3.2. *For any $v, w \in V_0^{1,1}(Q_T)$, $y \in Y_{\mathrm{div}}(Q_T)$, $\delta \in (0,2]$, real-valued functions $\epsilon(t) \in [1,+\infty[$, and $\gamma(t) \in \left[\frac{1}{2},+\infty\right[$, the following estimate holds:*

$$[e]^2_{(\nu,\theta,\zeta,2)} \le \overline{\mathrm{M}}^2_{\mathrm{II}}(v,y,w;\delta,\epsilon,\gamma,\mu) := \epsilon\|w(\cdot,T)\|^2_{\Omega} + 2L + l$$
$$\int_0^T \left(\gamma \left\| \tfrac{1}{\lambda}\mathbf{r}_f^{\mu} \right\|^2_{\Omega} + \alpha_1 \tfrac{C^2_{\mathrm{F}\Omega}}{\underline{\nu}_A} \| \mathbf{r}_f^{1-\mu} \|^2_{\Omega} + \alpha_2 \|\mathbf{r}_A\|^2_{A^{-1}} + \alpha_3 \tfrac{\widetilde{C}^2_{\mathrm{T}\Gamma_R}}{\underline{\nu}_A} \|\mathbf{r}_\sigma\|^2_{\Gamma_R} \right) \mathrm{d}t, \quad (3.23)$$

where

$$L(v,w) := \int_{Q_T} \Big(f w + v_t\, w - A\nabla v \cdot \nabla w - \lambda^2 v\, w \Big)\, \mathrm{d}x\mathrm{d}t - \int_{S_R} \sigma^2\, w\, \mathrm{d}s\mathrm{d}t,$$

$$l(v,w) := \int_{\Omega} |v(x,0) - \varphi(x)|^2 - 2w(x,0)\big(\varphi(x) - v(0,x)\big)\, \mathrm{d}x,$$

$\widetilde{C}_{\mathrm{T}\Gamma_R}$ is defined in (3.14)*, and parameters $\nu = 2-\delta$, $\theta(x,t) = \lambda(x)\left(2 - \frac{1}{\gamma(t)}\right)^{1/2}$, $\zeta = 1 - \frac{1}{\epsilon}$. Here, $\mu(x,t)$ is a real-valued function taking values in $[0,1]$, and $\alpha_1(t)$, $\alpha_2(t)$, and $\alpha_3(t)$ are positive real-valued functions satisfying the relation (3.15).*

Proof. The proof is analogous to the steps of the proof of Theorem 3.1 and can be found in [124, 54, 121] and [PIII, Theorem 3.1 (i)]. □

Theorem 3.3. *For any $\delta \in (0,2]$, real-valued functions $\gamma(t) \in \left[\frac{1}{2},+\infty\right[$, $\epsilon(t) \in [1,+\infty[$, and $\mu(x,t) \in [0,1]$, the lower bound of the variation problems*

$$\inf_{\substack{v \in V_0^{1,1}(Q_T) \\ y \in Y_{\mathrm{div}}(Q_T)}} \overline{\mathrm{M}}_{\mathrm{I}}^2(v,y) \quad \text{and} \quad \inf_{\substack{v,w \in V_0^{1,1}(Q_T) \\ y \in Y_{\mathrm{div}}(Q_T)}} \overline{\mathrm{M}}^2_{\mathrm{II}}(v,y,w)$$

is zero, and it is attained if and only if $v = u$, $y = A\nabla u$, and $w = 0$.

Proof. See, e.g., [PIII, Theorem 2.1 (ii), Theorem 3.1 (ii)]. □

Computable lower bounds of the error in the exact solutions of PDEs provide useful information, which allows us to judge the quality of error majorants. In [PII], the lower bounds of the error in the solution of a I-BVP are presented and numerically studied for the first time.

Theorem 3.4. *Let v, $\eta \in V_0^{1,1}(Q_T)$, then the following estimate holds:*

$$\underline{\mathrm{M}}^2(\eta,v;\mathbf{k}) := \sup_{\eta \in V_0^{1,1}(Q_T)} \left\{ \sum_{i=1}^{5} G_{v,i}(\eta) + G_{fu_0}(\eta) \right\} \le [\,e\,]^2_{(\underline{\nu},\underline{\theta},\underline{\zeta},\underline{\chi})},$$

where

$$
\begin{aligned}
G_{v,1} &= \int\limits_{Q_T} \Big(- \nabla\eta \cdot A\nabla v - \tfrac{1}{2\kappa_1}|\nabla\eta|^2 \Big)\, \mathrm{d}x\mathrm{d}t,\\
G_{v,2} &= \int\limits_{Q_T} \Big(\eta_t v - \tfrac{1}{2\kappa_2}|\eta_t|^2 \Big)\, \mathrm{d}x\mathrm{d}t,\\
G_{v,3} &= \int\limits_{Q_T} \varrho^2 \Big(- v\eta - \tfrac{1}{2\kappa_3}|\eta|^2 \Big)\, \mathrm{d}x\mathrm{d}t,\\
G_{v,4} &= \int\limits_{\Omega} \Big(- v(x,T)\eta(x,T) - \tfrac{1}{2\kappa_4}\,|\eta(x,T)|^2 \Big)\, \mathrm{d}x,\\
G_{v,5} &= \int\limits_{S_R} \sigma^2 \Big(- v\eta - \tfrac{1}{2\kappa_5}|\eta|^2 \Big)\, \mathrm{d}s\mathrm{d}t,
\end{aligned}
$$

and

$$
G_{fu_0} = \int\limits_{Q_T} f\eta\, \mathrm{d}x\mathrm{d}t + \int\limits_{\Omega} u_0\eta(\cdot,0)\, \mathrm{d}x,
$$

with constant parameters $\underline{\nu} = \frac{\kappa_1}{2}$, $\underline{\theta} = \Big(\frac{1}{2}\big(\kappa_2 + \kappa_3\lambda^2\big)\Big)^{1/2}$, $\underline{\zeta} = \frac{\kappa_4}{2}$, $\underline{\chi} = \frac{\kappa_5}{2}$, *and* $\mathbf{k} = (\kappa_1, \kappa_2, \kappa_3, \kappa_4, \kappa_5)$ *is a vector with positive coordinates.*

Proof. See, e.g., [PII, Section 3] and [PIII, Section 4]. □

3.3 Global minimization of the majorant

In this section, we discuss the algorithm of global majorant minimization, which implies a tool for a posteriori control of the error in the approximate solution of a parabolic I-BVPs. In [138], a priori error estimates are presented for both the semi-discrete problem resulting in a spatial one and for the most commonly used fully discrete schemes obtained by space-time discretization. First, we present a majorant adapted to the time-stepping class of methods and confirm its efficiency (both as an error estimate and an indicator) in Examples 1–3. Since the majorant is defined as the integral over total time-interval $[0, T]$, it is also applied to approximations obtained by space-time discretization techniques on the whole cylinder Q_T (see Examples 4 and 5).

For the reader's convenience, we assume that $\lambda = 0$, which implies $\mu = 0$, matrix $A = A(x)$ is symmetric, $v(\cdot, 0) = u_0$, and $S_T = S_D$ in (2.20). Thus, the error is simplified down to a sum

$$
[\,e\,]^2 := (2-\delta)\, e_{\mathrm{d}}^2 + \|e(\cdot,T)\|_{\Omega} \quad \text{with} \quad e_{\mathrm{d}}^2 = \int\limits_0^T \|\nabla e\|_{\Omega}^2\, \mathrm{d}t, \quad \forall \delta \in (0,2],
$$

where the first term is equivalent to the energy norm and the second one illustrates the error at $t = T$. The majorant, respectively, reads as

$$\overline{\mathrm{M}}_{\mathrm{I}}^2(v, y; \alpha_1, \alpha_2) := \alpha_1 \int_0^T \| y - A\nabla v \|^2_{A^{-1}} \, \mathrm{d}t + \alpha_2 \tfrac{C^2_{\mathrm{F}\Omega}}{\underline{\nu}_A} \int_0^T \| f + \mathrm{div} y - v_t \|^2_{\Omega} \, \mathrm{d}t. \quad (3.24)$$

Here, $\overline{\mathrm{m}}_{\mathrm{f}}^2 := \int_0^T \| f + \mathrm{div} y - v_t \|^2_{\Omega} \, \mathrm{d}t$ assures the reliability of the majorant and measures the violation of the equilibrium equation (2.12), whereas the first term mimics the residual in (2.13) and has confirmed to work as a robust and efficient indicator. Further, the latter one is denoted by $\overline{\mathrm{m}}_{\mathrm{d}}^2 := \int_0^T \| y - A\nabla v \|^2_{A^{-1}} \, \mathrm{d}t$. To measure the reliability and presentation accuracy of $\overline{\mathrm{M}}_{\mathrm{I}}^2$, we use the efficiency index $I_{\mathrm{eff}} := \frac{\overline{\mathrm{M}}_{\mathrm{I}}}{[e]}$.

In order to adapt the majorant (15) to the methods based on time-stepping reconstruction of the approximate solution, we define the following discretization of the time-interval $[0,\ T]$:

$$\mathcal{T}_K = \cup_{k=0}^{K-1} \overline{I^{(k)}}, \quad \text{where} \quad I^{(k)} = (t^k, t^{k+1}). \quad (3.25)$$

Then the time-cylinder can be represented in the form

$$\overline{Q}_T = \cup_{k=0}^{K-1} \overline{Q^{(k)}}, \quad \text{where} \quad Q^{(k)} := I^{(k)} \times \Omega. \quad (3.26)$$

Let $\mathcal{T}_{N_1 \times .. \times N_d}$ be a mesh selected on Ω. Then, $\Theta_{K \times N_1 \times ... \times N_d} = \mathcal{T}_K \times \mathcal{T}_{N_1 \times ... \times N_d}$ denotes the mesh on Q_T. Generally, domain Ω_t of variables x can change its shape in time, i.e., $Q_T := \{(x,t) : x \in \Omega_t, t \in (0,T)\}$, which is more natural to hangle with space-time FEM schemes. For time-incremental methods, we consider only problems on 'right cylinder'.

From now on, we limit our discussion to the time-slice $Q^{(k)}$, such that $y \in Y_{\mathrm{div}}(Q^{(k)})$, $v \in V_0^{1,1}(Q^{(k)})$. We set $\alpha_1 = \frac{1}{\delta}(1 + \frac{1}{\beta})$ and $\alpha_2 = \frac{1}{\delta}(1 + \beta)$, where $\beta(t)$ is a positive bounded function for a.e. $t \in I^{(k)}$. For simplicity, we assume $\beta = \mathrm{const}$. On each $Q^{(k)}$, *the increment of the majorant* (15) is denoted by $\overline{\mathrm{M}}_{\mathrm{I}}^{2,(k)}$, i.e.,

$$\overline{\mathrm{M}}_{\mathrm{I}}^{2,(k)}(v, y; \beta) := \tfrac{1}{\delta}\Big((1+\beta)\, \overline{\mathrm{m}}_{\mathrm{f}}^{2,(k)} + \big(1 + \tfrac{1}{\beta}\big) \tfrac{C^2_{\mathrm{F}\Omega}}{\underline{\nu}_A} \overline{\mathrm{m}}_{\mathrm{d}}^{2,(k)} \Big). \quad (3.27)$$

We intend to define optimal y by minimization of the increment of the majorant, i.e.,

$$\min_{\beta > 0} \min_{y \in Y_{\mathrm{div}}(Q^{(k)})} \overline{\mathrm{M}}_{\mathrm{I}}^{2,(k)}(v, y; \beta).$$

The corresponding *increment of the error* is denoted by $[\, e \,]^{2,(k)}$. The minimum of $\overline{\mathrm{M}}_{\mathrm{I}}^{2,(k)}(y; \beta)$ with respect to β is attained at $\beta_{\min} := \left(\frac{C^2_{\mathrm{F}\Omega} \overline{\mathrm{m}}_{\mathrm{f}}^{2,(k)}}{\underline{\nu}_A \overline{\mathrm{m}}_{\mathrm{d}}^{2,(k)}} \right)^{1/2}$. After β is fixed,

the necessary condition for the minimizer y reads as

$$\frac{\mathrm{d}\overline{\mathrm{M}}_{\mathrm{I}}^{2,(k)}(v, y+\zeta w; \beta)}{\mathrm{d}\zeta}\Big|_{\zeta=0} = 0, \tag{3.28}$$

where $w \in Y_{\mathrm{div}}(Q^{(k)})$. Condition (3.28) yields

$$\int\limits_{Q^{(k)}} \Big(\frac{C_{\mathrm{F}\Omega}^2}{\beta \underline{\nu}_A} \operatorname{div} y \operatorname{div} w + A^{-1} y \cdot w\Big)\, \mathrm{d}x\mathrm{d}t = \int\limits_{Q^{(k)}} \Big(-\frac{C_{\mathrm{F}\Omega}^2}{\beta \underline{\nu}_A}(f - v_t) \operatorname{div} w + \nabla v \cdot w\Big)\, \mathrm{d}x\mathrm{d}t.$$

We reduce the integration with respect to the time by the following linear extension of v and y on increment $Q^{(k)}$

$$v = v^k \tfrac{t^{k+1}-t}{\tau^k} + v^{k+1}\tfrac{t-t^k}{\tau^k}, \quad y = y^k \tfrac{t^{k+1}-t}{\tau^k} + y^{k+1}\tfrac{t-t^k}{\tau^k}, \quad \tau^k = t^{k+1} - t^k, \tag{3.29}$$

such that v^k, $v^{k+1} \in H_0^1(\Omega)$, y^k, $y^{k+1} \in H(\operatorname{div}, \Omega)$, and $w(x,t) = \eta(x) \cdot T(t)$ with $T = \frac{t-t^k}{\tau^k}$ and $\eta \in H(\operatorname{div}, \Omega)$.

As the results, one obtains

$$\begin{aligned}
&\tfrac{C_{\mathrm{F}\Omega}^2}{\beta \underline{\nu}_A} \int\limits_{\Omega} \big(\tfrac{1}{2}\operatorname{div}\mathbf{y^{k+1}} + \operatorname{div} y^k\big) \operatorname{div}\eta \,\mathrm{d}x + \int\limits_{\Omega} A^{-1}\big(\tfrac{1}{2}\mathbf{y^{k+1}} + y^k\big) \cdot \eta \,\mathrm{d}x \\
&= -\tfrac{C_{\mathrm{F}\Omega}^2}{\beta \underline{\nu}_A} \int\limits_{\Omega} \Big(\tfrac{3}{(\tau^k)^2} F_{(t-t^k)}(x) - \tfrac{3(v^{k+1}-v^k)}{2\tau^k}\Big) \operatorname{div}\eta \,\mathrm{d}x + \int\limits_{\Omega} \big(\tfrac{1}{2}\nabla v^{k+1} + \nabla v^k\big) \cdot \eta \,\mathrm{d}x,
\end{aligned}$$

where y^{k+1} is the unknown function we are interested to reconstruct and $F_{(t-t^k)}(x) = \int\limits_{t^k}^{t^{k+1}} f(t - t^k)\,\mathrm{d}t$ is approximated by Gauss quadratures of high order [135, 134].

Assume now that y^k, y^{k+1}, and $\eta \in \operatorname{span}\{\phi_1, \ldots, \phi_N\} =: Y^N \subset H(\operatorname{div}, \Omega)$, i.e., $y^k = \sum\limits_{i=1}^{N} Y_i^k \phi_i$ and $\eta = \phi_j$, $j = 1, \ldots, N$. The condition (3.28) leads to a SLE

$$\Big(\tfrac{C_{\mathrm{F}\Omega}^2}{\beta \underline{\nu}_A} S + K\Big)\, \mathbf{Y^{k+1}} = -\tfrac{1}{2}\Big(\tfrac{C_{\mathrm{F}\Omega}^2}{\underline{\nu}_A} S + K\Big)\, Y^k - \tfrac{C_{\mathrm{F}\Omega}^2}{\beta \underline{\nu}_A} \tfrac{3}{(\tau^k)^2} z + g, \tag{3.30}$$

where $\mathbf{Y^{k+1}} \in \mathbb{R}^N$ is the vector of unknowns, and components of matrices S, K

and vectors z, g are defined as follows:

$$\{S_{ij}\}_{i,j=1}^{N} = \int_{\Omega} \operatorname{div}\phi^i \operatorname{div}\phi^j \,\mathrm{d}x, \tag{3.31}$$

$$\{z_j\}_{j=1}^{N} = \int_{\Omega} \left(F_{(t-t^k)} + \tfrac{(v^k - v^{k+1})\tau^k}{2}\right) \operatorname{div}\phi^j \,\mathrm{d}x, \tag{3.32}$$

$$\{K_{ij}\}_{i,j=1}^{N} = \int_{\Omega} A^{-1}\phi^i \cdot \phi^j \,\mathrm{d}x, \tag{3.33}$$

$$\{g_j\}_{j=1}^{N} = \int_{\Omega} \left(\tfrac{1}{2}\nabla v^{k+1} + \nabla v^k\right) \cdot \phi^j \,\mathrm{d}x. \tag{3.34}$$

The observations above motivate Algorithm 1 (p. 39), which summarizes the steps the optimization $\overline{\mathrm{M}}_{\mathrm{I}}^{2,(k)}(v, y; \beta)$, such that on each time-step the increment of the majorant is reconstructed by means of the iteration procedure. Sequence of fluxes, obtained in the iteration procedure, helps to generate sequence of upper bounds as close to the value of the error as desired. On each $Q^{(k)}$, we obtain optimal y^{k+1}, which is used as initial data on $Q^{(k+1)}$. The second form of the majorant and minorant can be presented in an analogous way. Generally, Algorithm 1 can be extended to work with approximations that have jumps in time (see, e.g., [125]). Moreover, the upper bound can be used as a refinement criteria for schemes adaptive in time. In space-time FE implementation, where time is considered as an extra dimension, we disctretize the majorant (15) by following the steps of Algorithm 3.2 in book of Mali, Neittaanmäki, and Repin [93, Section 3.3.1].

3.4 Numerical experiments

The detailed study of numerical application of $\overline{\mathrm{M}}_{\mathrm{I}}^{2}(v, y; \delta, \gamma, \mu)$ for $\Omega \in \mathbb{R}^d$, $d = \{1, 2\}$ is presented in [PII]. In the same paper, we test $\underline{\mathrm{M}}^{2}(\eta, v; \mathbf{k})$ and compare it to the majorant in (3.13). Besides that, we test the behavior of $\overline{\mathrm{M}}_{\mathrm{I}}^{2}(v, y; \delta, \gamma, \hat{\mu})$ with optimal auxiliary function $\hat{\mu}$ with respect to different λ, and show that the majorant stays robust even with a drastic change in the reaction over Ω. In [PII, Examples 3–5], we consider the numerical behavior of the indicators $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(k)}$, verify its efficiency by several criteria, i.e., different marking procedures denoted by $\mathbb{M}$, quantitative histograms, and other means.

To overcome the drawback in the initial implementation of the functional error estimates in `MATLAB` related to a large number of loops evaluating the local distribution of the majorant, a more advanced set of numerical tests has been performed with the help of `The FEniCS Project` library [137, 91]. We start with test-examples, where the I-BVP is discretized by the incremental method and the majorant is reconstructed and optimized, using the global optimization technique presented in Section 3.3. From here on, parameter δ in (3.27) is set to 1.

Algorithm 1 Global minimization of $\overline{\mathrm{M}}_{\mathrm{I}}^{2,(k)}$

Input: $Q^{(k)}$: $v^k, v^{k+1}, y^k(Y^k)$ {approximate solutions at fixed cuts of time and flux coefficients on $t^k \times \Omega$}
$\phi_i, i = 1, \dots, N$ {basis functions}
$M_{\max}^{\text{iter}}$ {number of iterations}

Assemble matrices S, K and vectors z, g by using

$$\{S_{ij}\}_{i,j=1}^{N} = \int_{\Omega} \operatorname{div}\phi^i \operatorname{div}\phi^j \,\mathrm{d}x, \ \{z_j\}_{j=1}^{N} = \int_{\Omega} \left(F_{(t-t^k)} + \tfrac{(v^k - v^{k+1})\tau}{2}\right) \operatorname{div}\phi^j \,\mathrm{d}x,$$
$$\{K_{ij}\}_{i,j=1}^{N} = \int_{\Omega} A^{-1}\phi^i \cdot \phi^j \,\mathrm{d}x, \quad \{g_j\}_{j=1}^{N} = \int_{\Omega} \left(\tfrac{1}{2}\nabla v^{k+1} + \nabla v^k\right) \cdot \phi^j \,\mathrm{d}x.$$

Approximate flux $y^k = \sum\limits_{i=1}^{N} Y_i^k \phi_i$.
Let $\beta = 1$.
for $m = 1$ **to** $M_{\max}^{\text{iter}}$ **do**
Solve the SLE $\left(\frac{C_{\mathrm{F}\Omega}^2}{\beta \underline{\nu}_A} S + K\right) \mathbf{Y_i^{k+1}} = -\frac{1}{2}\left(\frac{C_{\mathrm{F}\Omega}^2}{\underline{\nu}_A} S + K\right) Y^k - \frac{C_{\mathrm{F}\Omega}^2}{\beta \underline{\nu}_A} \frac{3}{(\tau^k)^2} z + g$.
Approximate flux $y^{k+1} = \sum\limits_{i=1}^{N} Y_i^{k+1} \phi_i$.
Reconstruct v and y on $Q^{(k)}$ by

$$v = v^k \tfrac{t^{k+1} - t}{\tau^k} + v^{k+1} \tfrac{t - t^k}{\tau^k}, \quad y = y^k \tfrac{t^{k+1} - t}{\tau^k} + y^{k+1} \tfrac{t - t^k}{\tau^k}, \quad \tau^k = t^{k+1} - t^k.$$

Compute the components of the majorant by

$$\overline{\mathrm{m}}_{\mathrm{f}}^{2,(k)} := \int_{t_k}^{t_{k+1}} \| f + \operatorname{div} y - v_t \|_{\Omega}^2 \,\mathrm{d}t \quad \text{and} \quad \overline{\mathrm{m}}_{\mathrm{d}}^{2,(k)} := \int_{t_k}^{t_{k+1}} \| y - A\nabla v \|_{A^{-1}}^2 \,\mathrm{d}t.$$

Compute optimal β by $\beta := \left(\frac{C_{\mathrm{F}\Omega}^2 \overline{\mathrm{m}}_{\mathrm{f}}^{2,(k)}}{\underline{\nu}_A \overline{\mathrm{m}}_{\mathrm{d}}^{2,(k)}}\right)^{1/2}$.
end for
Compute result increment of the majorant by

$$\overline{\mathrm{M}}_{\mathrm{I}}^{2}(v, y; \alpha_1, \alpha_2) := \alpha_1 \int_0^T \| y - A\nabla v \|_{A^{-1}}^2 \,\mathrm{d}t + \alpha_2 \frac{C_{\mathrm{F}\Omega}^2}{\underline{\nu}_A} \int_0^T \| f + \operatorname{div} y - v_t \|_{\Omega}^2 \,\mathrm{d}t.$$

Output: $\overline{\mathrm{M}}_{\mathrm{I}}^{2,(k)}(v, y; \beta)$ {incremental majorant on $Q^{(k)}$}
$y^{k+1}(Y^{k+1})$ {reconstruction of the flux on $t^{k+1} \times \Omega$}

Example 1. First, we consider a benchmark problem on unit square $\Omega = (0,1)^2 \subset \mathbb{R}^2$ and $T = 1$ with homogeneous Dirichlet BC, initial function $u_0 = x(1-x)\,y(1-y)$, and $u = x(1-x)\,y(1-y)\,(t^2+t+1)$ as the exact solution (the source function f is calculated respectively). The approximation v is reconstructed by the P_1 Lagrangian FE space (see Figure 3), and the flux y by the linear (RT_1) Raviart-Thomas FE space.

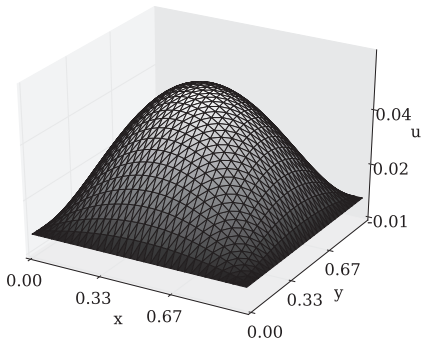
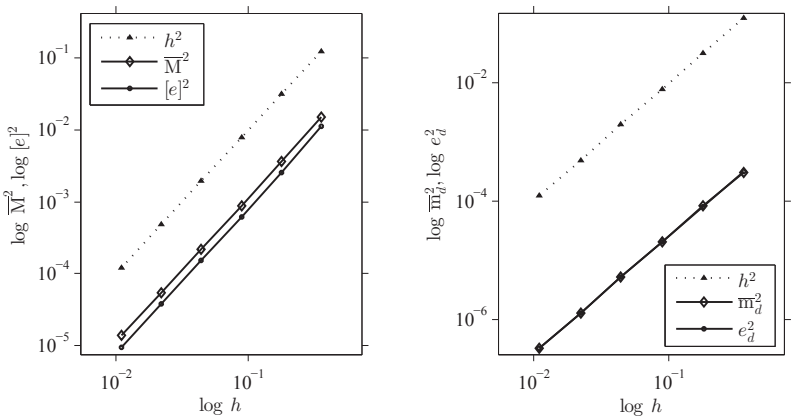


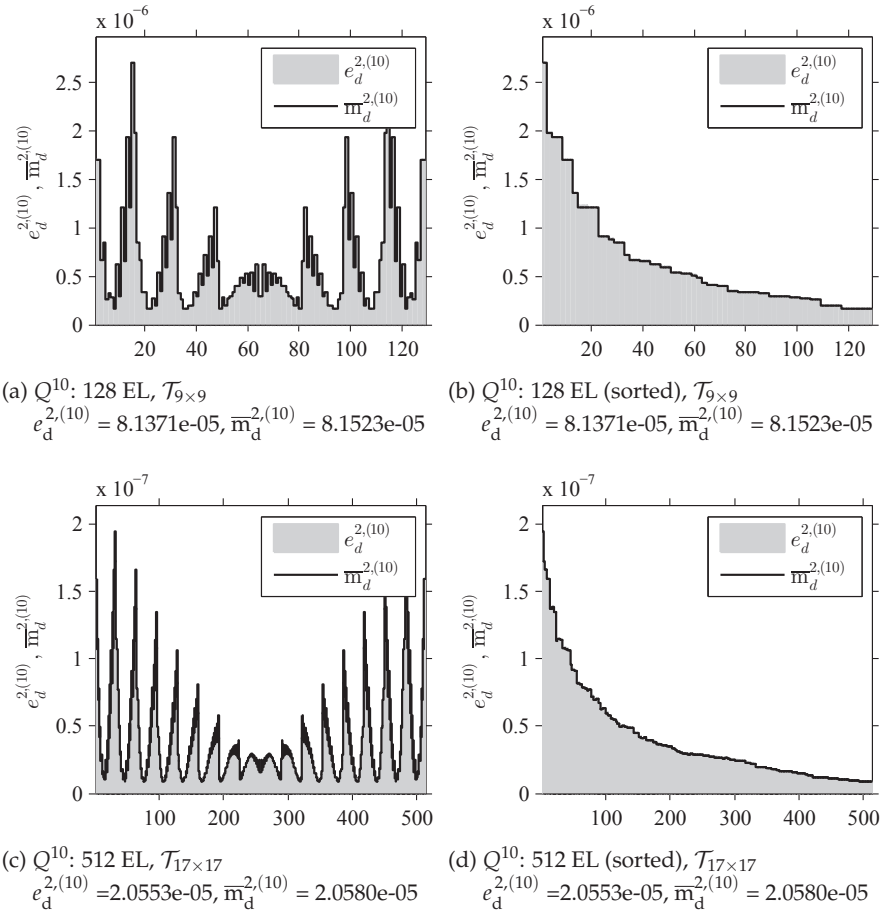
FIGURE 3 Example 1. The approximate solution on the mesh (1089 ND) at $t = 0.1$.

The optimal convergence test for fixed time-step and decreasing mesh size h is illustrated in Figure 4. Here, Figure 4a depicts the total error $[\,e\,]^2$ and majorant $\overline{M}_I^2$, whereas Figure 4b illustrates the part of the true error e_d^2 (the energy norm) and the indicator $\overline{m}_d^2$ decrease with respect to h. We see that both $\overline{m}_d^2$ and $\overline{M}_I^2$ have the expected quadratic convergence, and the values of the indicator (see Figure 4b) practically coincide with e_d^2.



(a) $[\,e\,]^2$ and $\overline{M}_I^2$ (b) e_d^2 and $\overline{m}_d^2$

FIGURE 4 Example 1. Optimal convergence tests.



(a) Q^{10}: 128 EL, $\mathcal{T}_{9\times 9}$
$e_{\mathrm{d}}^{2,(10)}$ = 8.1371e-05, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(10)}$ = 8.1523e-05

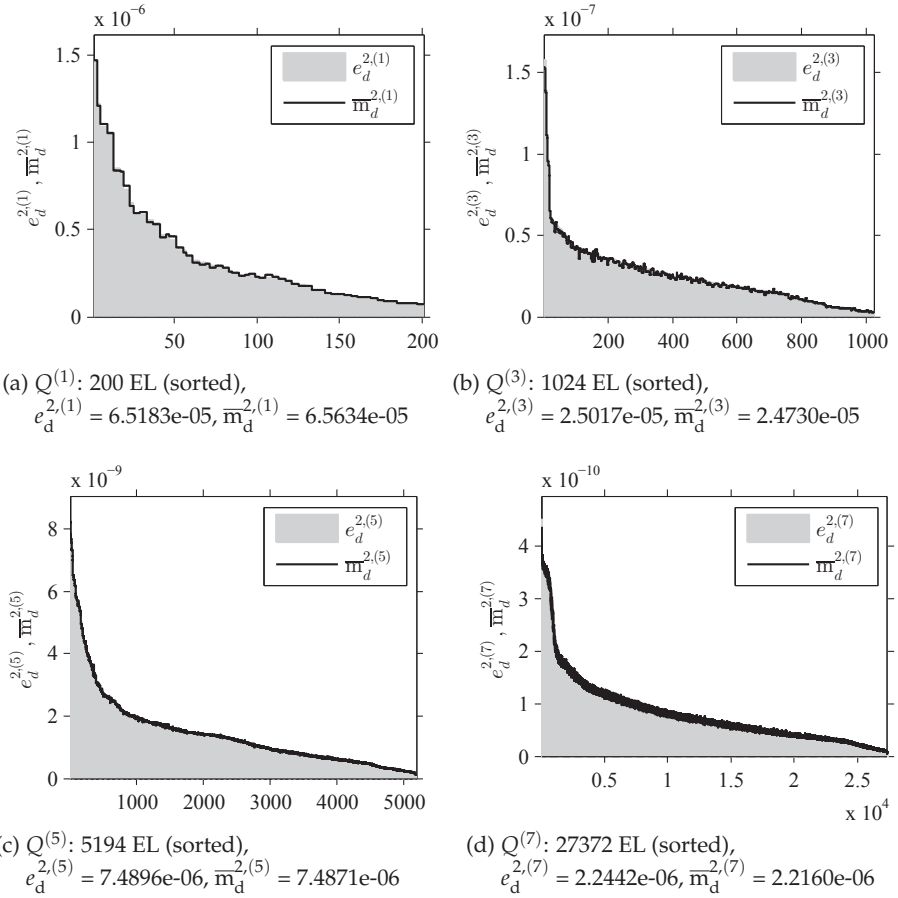
(b) Q^{10}: 128 EL (sorted), $\mathcal{T}_{9\times 9}$
$e_{\mathrm{d}}^{2,(10)}$ = 8.1371e-05, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(10)}$ = 8.1523e-05

(c) Q^{10}: 512 EL, $\mathcal{T}_{17\times 17}$
$e_{\mathrm{d}}^{2,(10)}$ =2.0553e-05, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(10)}$ = 2.0580e-05

(d) Q^{10}: 512 EL (sorted), $\mathcal{T}_{17\times 17}$
$e_{\mathrm{d}}^{2,(10)}$ =2.0553e-05, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(10)}$ = 2.0580e-05

FIGURE 5 Example 1. The distribution of the energy part of the error and the indicator over $Q^{(10)}$.

Next, we consider solving the problem with a fixed mesh on each time-step and confirm that $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(k)}$ does represent the local distribution of $e_{\mathrm{d}}^{2,(k)}$ efficiently on each $Q^{(k)}$. We fix meshes $\Theta_{10\times 9\times 9}$ (Figures 5a–5b) and $\Theta_{10\times 17\times 17}$ (Figures 5c–5d) such that the time discretization parameter $K = 10$ and compare the distributions of $e_{\mathrm{d}}^{2,(10)}$ with $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(10)}$ on $Q^{(10)}$ for both meshes. Here, Figures 5a and 5c present the distributions of $e_{\mathrm{d}}^{2,(10)}$ and $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(10)}$ element-wise, where elements (EL) are enumerated according to the algorithm used in the FE implementation. Whereas, Figs. 5b and 5d illustrate the same distributions, but here the cells are sorted with respect to the decreasing values of the local true errors $e_{\mathrm{d}}^{2,(10)}$. The array containing $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(10)}$ is depicted in the order defined by the indeces obtained after error sorting. Both ways of presenting the results must convince the reader on the quantitative efficiency of the tested error indicator. The evaluation of the error and indicator distributions by such histograms was introduced in [93, Section 3.4, Example 3.4].



(a) $Q^{(1)}$: 200 EL (sorted),
$e_{\mathrm{d}}^{2,(1)}$ = 6.5183e-05, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(1)}$ = 6.5634e-05

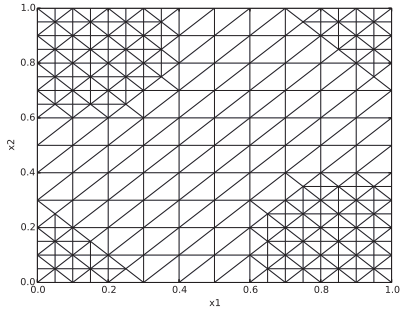
(b) $Q^{(3)}$: 1024 EL (sorted),
$e_{\mathrm{d}}^{2,(3)}$ = 2.5017e-05, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(3)}$ = 2.4730e-05

(c) $Q^{(5)}$: 5194 EL (sorted),
$e_{\mathrm{d}}^{2,(5)}$ = 7.4896e-06, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(5)}$ = 7.4871e-06

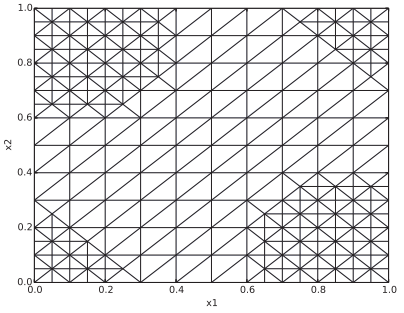
(d) $Q^{(7)}$: 27372 EL (sorted),
$e_{\mathrm{d}}^{2,(7)}$ = 2.2442e-06, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(7)}$ = 2.2160e-06

FIGURE 6 Example 1. The true error and indicator distribution on different time slices $Q^{(k)}$, $k = 1, 3, 5, 7$, with refinement using bulk marking $\mathbb{M}_{0.3}$.

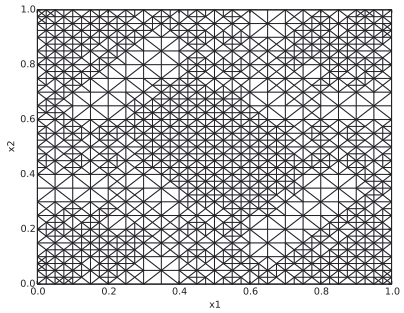
Next, we consider the adaptive refinement strategy with bulk marking criteria $\mathbb{M}_\theta$, where parameter $\theta = 0.3$ (see [42]). The initial mesh is $\mathcal{T}_{11\times 11}$ (200 EL, 121 ND). Figure 6 illustrates the distributions of $e_{\mathrm{d}}^{2,(k)}$ and $RHS\overline{\mathrm{m}}_{\mathrm{d}}^{2,(k)}$ on different slices $Q^{(k)}$, $k = 1, 3, 5, 7$, which demonstrates the quantitative efficiency of the indicator provided by the majorant. Under each sub-plot of Figure 6, we also place information on the total values of $e_{\mathrm{d}}^{2,(k)}$ and $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(k)}$. Moreover, we can analyze meshes obtained during the refinement based either on $e_{\mathrm{d}}^{2,(k)}$ or $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(k)}$ (see Figure 7). In Figures 7a, 7c, and 7e, we present meshes obtained after the refinement process based on the local error distribution, and Figures 7b, 7d, and 7f contain meshes constructed when the refinement is based on the local indicator. It is easy to observe that the meshes on the RHS of Figure 7 resemble the meshes on the LHS. Moreover, the number of EL in the meshes from both sides is close (see Table 1, which illustrates the difference in numbers of EL on slices $Q^{(k)}$). The efficiency of the total majorant is $I_{\mathrm{eff}} = 1.23$.



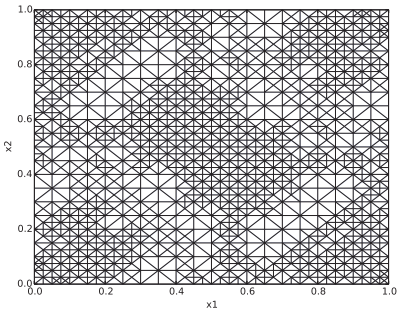
(a) $Q^{(1)}$: 200 EL, 121 ND
$e_{\mathrm{d}}^{2,(1)}$ = 6.5183e-05, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(1)}$ = 6.5634e-05



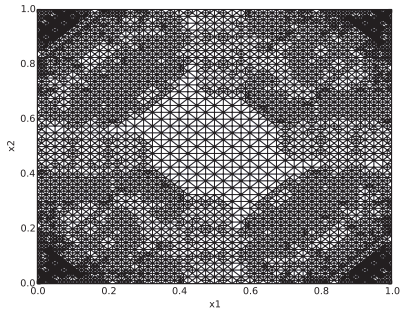
(b) $Q^{(1)}$: 200 EL, 121 ND
$e_{\mathrm{d}}^{2,(1)}$ = 6.5183e-05, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(1)}$ = 6.5634e-05



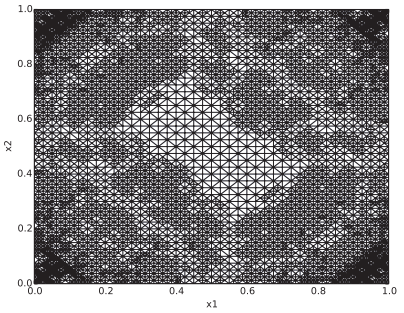
(c) $Q^{(3)}$: 1036 EL, 563 ND
$e_{\mathrm{d}}^{2,(3)}$ = 2.4678e-05, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(3)}$ = 2.4389e-05



(d) $Q^{(3)}$: 1024 EL, 557 ND
$e_{\mathrm{d}}^{2,(3)}$ = 2.5017e-05, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(3)}$ = 2.473e-05

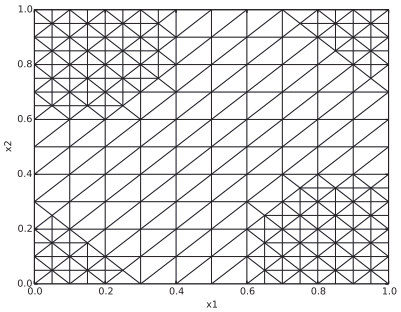


(e) $Q^{(5)}$: 5198 EL, 2692 ND
$e_{\mathrm{d}}^{2,(5)}$ = 7.481e-06, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(5)}$ = 7.4786e-06

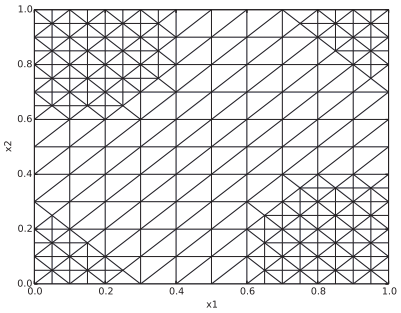


(f) $Q^{(5)}$: 5194 EL, 2692 ND
$e_{\mathrm{d}}^{2,(5)}$ = 7.4896e-06, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(5)}$ = 7.4871e-06

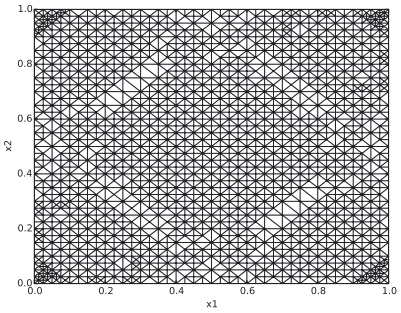
FIGURE 7 Example 1. Evolution of meshes on time-slices $Q^{(k)}$, $k = 1, 3, 5$. The refinement is based on the error (a), (c), (e) and indicator (b), (d), (f) using bulk marker $\mathbb{M}_{0.3}$.



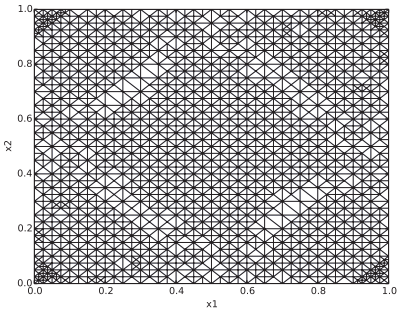
(a) $Q^{(1)}$: 200 EL,121 ND
$e_{\mathrm{d}}^{2,(1)}$ = 6.5183e-05, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(1)}$ = 6.5634e-05



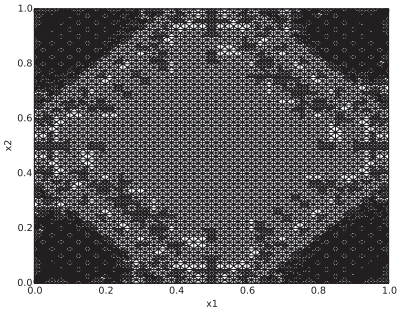
(b) $Q^{(1)}$: 200 EL, 121 ND
$e_{\mathrm{d}}^{2,(1)}$ = 6.5183e-05, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(1)}$ = 6.5634e-05



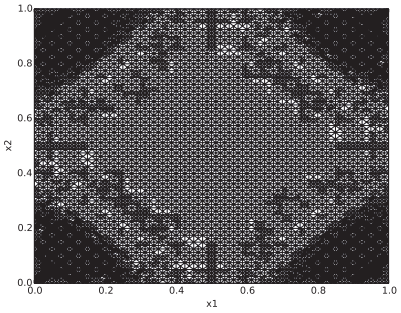
(c) $Q^{(3)}$: 1144 EL, 623 ND
$e_{\mathrm{d}}^{2,(3)}$ = 2.2982e-05, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(3)}$ = 2.2693e-05



(d) $Q^{(3)}$: 1144 EL, 623 ND
$e_{\mathrm{d}}^{2,(3)}$ = 2.2982e-05, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(3)}$ = 2.2693e-05



(e) $Q^{(5)}$: 7748 EL, 3985 ND
$e_{\mathrm{d}}^{2,(5)}$ = 5.1237e-06, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(5)}$ = 5.1238e-06



(f) $Q^{(5)}$: 7660 EL, 3933 ND
$e_{\mathrm{d}}^{2,(5)}$ = 5.1660e-06, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(5)}$ = 5.1662e-06

FIGURE 8 Example 1. Evolution of meshes on time-slices $Q^{(k)}$, $k = 1, 3, 5$. The refinement is based on the error (a), (c), (e) and indicator (b), (d), (f) using marker $\mathbb{M}_{\mathrm{AVR}}$.

TABLE 1 Example 1. The difference in numbers of EL in meshes generated during refinement using $\mathbb{M}_{0.3}$ based on the true error and the indicator.

k	# EL in $\mathcal{T}_{N_1\times N_2}$ (ref. $e_{\mathrm{d}}^{2,(k)}$)	# EL in $\mathcal{T}_{N_1\times N_2}$ (ref. $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(k)}$)	difference in # EL, %
1	200	200	0%
2	420	200	0%
3	1036	1024	1.16%
4	2310	2266	1.9%
5	5198	5194	0.07%
6	11888	11932	0.37%
7	27372	27388	0.06%
8	64334	64264	0.11%
9	154300	152716	1.03%
10	375150	366964	2.18%

TABLE 2 Example 1. The difference in numbers of EL in meshes generated during refinement using $\mathbb{M}_{\mathrm{AVR}}$ based on the true error and the indicator.

k	# EL in $\mathcal{T}_{N_1\times N_2}$ (ref. $e_{\mathrm{d}}^{2,(k)}$)	# EL in $\mathcal{T}_{N_1\times N_2}$ (ref. $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(k)}$)	difference in # EL, %
1	200	200	0%
2	420	420	0%
3	1144	1144	0%
4	3116	3068	1.54%
5	7748	7660	1.14%
6	21112	21180	0.32%
7	55592	55744	0.27%
8	155284	155724	0.28%
9	422300	418304	0.95%

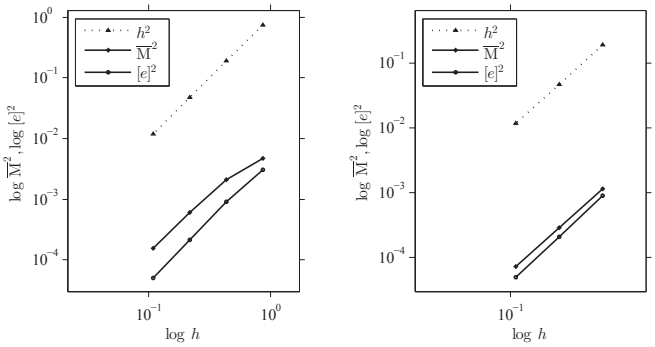
The bulk marking strategy can be compared to the marking determined by the level of the average error $\mathbb{M}_{\mathrm{AVR}}$ [93, Algorithm 2.1]. Figure 8 demonstrates the sequence of the meshes obtained as a result of the refinement based on $e_{\mathrm{d}}^{2,(k)}$ (LHS) and $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(k)}$ (RHS) on $Q^{(k)}$, $k = 1, 3, 5$ (to compare the difference in numbers of EL on different time-slices see Table 2). In this case, we obtain $I_{\mathrm{eff}} = 1.4$. The efficiency index is not as accurate as it is expected, due to the fact that, unlike in the elliptic BVP, there is always a gap between $\overline{\mathrm{M}}_{\mathrm{I}}^2$ and $[\,e\,]^2$ related to the time-derivative v_t in $\overline{\mathrm{m}}_{\mathrm{f}}^2$.

It is important to note that the majorant can be used as a tool to predict 'blow-ups' in time-dependent explicit schemes, which are much less time-consuming than the implicit one but are unstable. Furthermore, for one-dimensional (in space) schemes, the stability condition is written explicitly, whereas there are no such criteria (CFL number) for two- and three-dimensional problems (see [36]). As an example, we consider the mesh $\Theta_{1280\times 121\times 121}$ (28800 EL, 14641 ND) and illustrate the majorant reaction on instability of the scheme (see Table 3). Here, the column DOF(v) reflects the degrees of freedom of v. The LHS of the table contains the total error and the majorant, obtained by using the stable implicit scheme, whereas the RHS illustrates how the majorant drastically increases even when the 'blow-up' is not yet obvious.

TABLE 3 Example 1. Total error, majorant, and efficiency index for approximations generated by implicit and explicit schemes.

	Implicit scheme				Explicit scheme			
k	DOF(v)	$[\![e]\!]^2$	$\overline{\mathrm{M}}$	I_{eff}	DOF(v)	$[\![e]\!]^2$	$\overline{\mathrm{M}}$	I_{eff}
1	14641	2.29e-09	3.78e-09	1.29	14641	1.26e-06	7.89e-05	7.93
2	23627	4.01e-09	6.84e-09	1.31	27175	2.06e-03	4.56e-03	1.49
3	39795	5.05e-09	9.14e-09	1.35	45489	9.19e+03	1.46e+04	1.26
4	67719	5.66e-09	1.06e-08	1.37	82344	1.15e+12	1.63e+12	1.19

Example 2. The same properties can be tested in the example with unit cube $\Omega = (0,1)^3 \subset \mathbb{R}^3$, $T = 1$, initial condition $u_0 = x\,(1-x)\,y\,(1-y)\,z\,(1-z)$, homogeneous Dirichlet BC, and $u = x\,(1-x)\,y\,(1-y)\,z\,(1-z)(t^2+t+1)$ (again, f is defined accordingly). Analogously, we consider $v \in \mathrm{P}_1$. However, in the current example, we compare the performance of the majorant for two different approximations of flux, i.e., $y \in \mathrm{RT}_0$ and $y \in \mathrm{RT}_1$. Figure 9a demonstrates the uniform convergence of $[\![e]\!]^2$ and $\overline{\mathrm{M}}_{\mathrm{I}}^2(y)$ with $y \in \mathrm{RT}_0$, whereas Figure 9b illustrates the same characteristics for $y \in \mathrm{RT}_1$. They both confirm the quadratic order of convergence of the majorant constructed with $y \in \mathrm{RT}_0$ and $y \in \mathrm{RT}_1$.

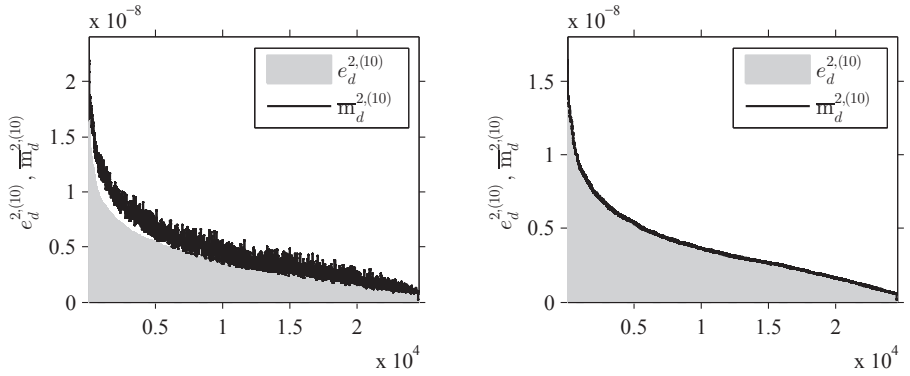


(a) $[\![e]\!]^2, \overline{\mathrm{M}}_{\mathrm{I}}^2(y), y \in RT_0$ (b) $[\![e]\!]^2, \overline{\mathrm{M}}_{\mathrm{I}}^2(y), y \in RT_1$

FIGURE 9 Example 2. Optimal convergence of $[e]^2$ and $\overline{\mathrm{M}}$, (a) $y \in RT_0$ and (b) $y \in RT_1$.

Next, we compare indicators the reconstruction of which is based on fluxes of different regularity. It is easy to see that $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(10)}$ on $Q^{(10)}$ (see Figure 10a) is less efficient than the one reconstructed from $y \in \mathrm{RT}_1$ in Figure 10b . Latter illustrations reaffirm the fact that one must use a flux of higher regularity in order to efficiently predict the local error distribution.

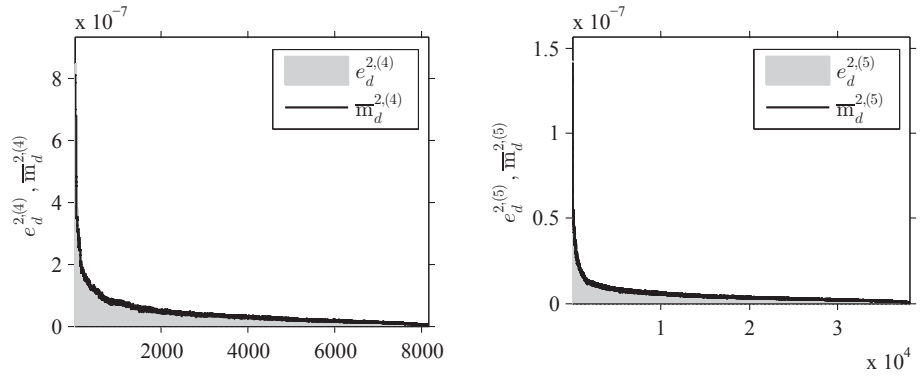
Finally, we consider a refinement strategy with the bulk marking $\mathbb{M}_{0.2}$. We take a coarse initial mesh $\mathcal{T}_{3\times3}$, $K = 5$, and illustrate the obtained error and majorant distribution after the refinement on time-slices $Q^{(4)}$ (Figure 11a) and $Q^{(5)}$ (Figure 11b). The number of obtained EL and the total values of $e_{\mathrm{d}}^{2,(k)}$ and $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(k)}$, $k = 4, 5$, are shown below the figures.



(a) $Q^{(10)}$: 24576 EL (sorted)
$e_{\mathrm{d}}^{2,(10)}$ = 9.1316e-05, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(10)}$ = 1.1578e-04

(b) $Q^{(10)}$: 24576 EL (sorted)
$e_{\mathrm{d}}^{2,(10)}$ = 9.1316e-05, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(10)}$ = 9.1473e-05

FIGURE 10 Example 2. Energy parts of true error and indicator distributions based on (a) $y \in RT_0$ and (b) $y \in RT_1$.



(a) Q^4: 8145 EL (sorted)
$e_{\mathrm{d}}^{2,(4)}$ = 3.7752e-04, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(4)}$ = 3.5351e-04

(b) Q^5: 38058 EL (sorted)
$e_{\mathrm{d}}^{2,(5)}$ = 1.9996e-04, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(5)}$ = 1.9462e-04

FIGURE 11 Example 2. The error and indicator distributions for time-slices $Q^{(4)}$ and $Q^{(5)}$ ($y \in RT_1$).

Example 3. Next, we consider an example with a singularity in the solution. The classical benchmark problem is defined on L-shaped domain $\Omega := (\text{-}1,1) \times (\text{-}1,1) \backslash [0,1) \times [0,\text{-}1)$ with $T = 1$, Dirichlet BC with load $u_D = r^{1/3} \sin\theta$, $r = (x^2 + y^2)$, $\theta = \frac{2}{3}\,\mathrm{atan2}(y,x)$ on S_D, input source function $f = r^{1/3} \sin\theta\ (2\,t + 1)$, and the initial condition $u_0 = r^{1/3} \sin\theta$. The corresponding exact solution is $u = r^{1/3} \sin\theta\ \left(t^2 + t + 1\right)$ with singularity in the point $(0,0)$ (see Figure 12).

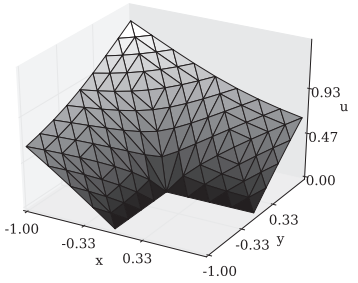
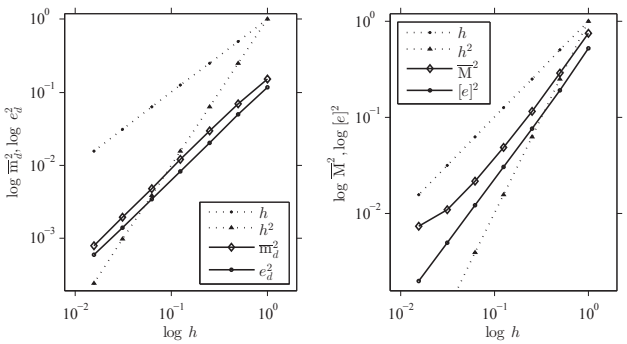
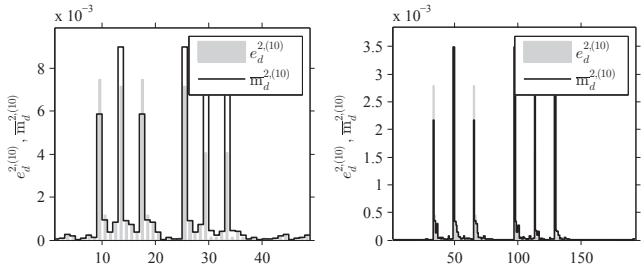


FIGURE 12 Example 3. Approximate solution on the mesh (113 ND) at $t = 0.1$.



(a) e_{d}^2 and $\overline{\mathrm{m}}_{\mathrm{d}}^2$ (b) $[\,e\,]^2$ and $\overline{\mathrm{M}}_{\mathrm{I}}^2$

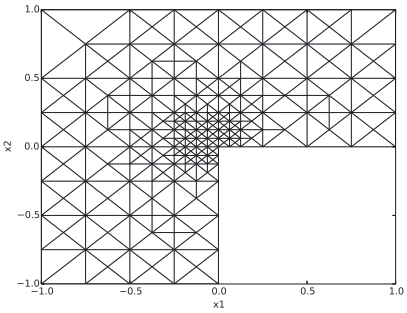
FIGURE 13 Example 3. Optimal convergence rate of (a) e_{d}^2 and $\overline{\mathrm{m}}_{\mathrm{d}}^2$ and (b) $[\,e\,]^2$ and $\overline{\mathrm{M}}_{\mathrm{I}}^2$.



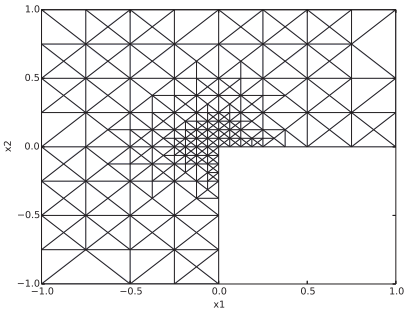
(a) $Q^{(10)}$: 48 EL, 33 ND
$e_{\mathrm{d}}^{2,(10)}$ = 4.9373e-02,
$\overline{\mathrm{m}}_{\mathrm{d}}^{2,(10)}$ = 7.0294e-02

(b) $Q^{(10)}$: 192 EL, 133 ND
$e_{\mathrm{d}}^{2,(10)}$ = 2.9371e-02,
$\overline{\mathrm{m}}_{\mathrm{d}}^{2,(10)}$ = 2.9371e-02

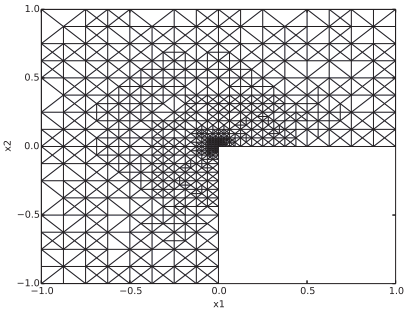
FIGURE 14 Example 3. The error and indicator distributions on $Q^{(10)}$, computed on a mesh with (a) 192 EL and (b) 768 EL.



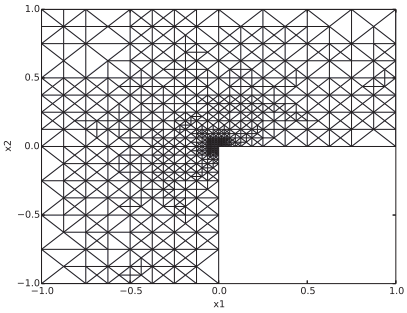
(a) $Q^{(3)}$: 188 EL, 105 ND
$e_{\mathrm{d}}^{2,(3)}$ = 3.00738e-03, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(3)}$ = 3.2841e-03



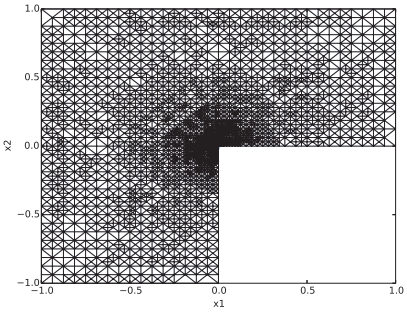
(b) $Q^{(3)}$: 188 EL, 105 ND
$e_{\mathrm{d}}^{2,(3)}$ = 3.0391e-03, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(3)}$ = 3.3192e-03



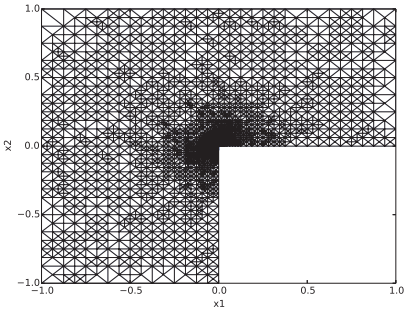
(c) $Q^{(5)}$: 710 EL, 379 DOF
$e_{\mathrm{d}}^{2,(5)}$ = 1.1376e-03, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(5)}$ = 1.2736e-03



(d) $Q^{(5)}$: 686 EL, 366 DOF
$e_{\mathrm{d}}^{2,(5)}$ = 1.1653e-03, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(5)}$ = 1.2981e-03

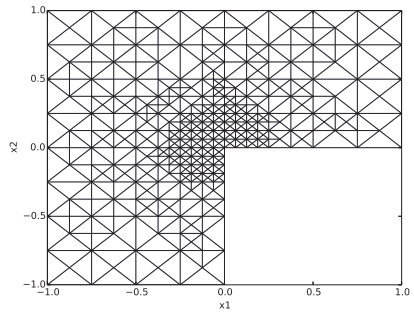


(e) $Q^{(7)}$: 2790 EL, 1445 DOF
$e_{\mathrm{d}}^{2,(7)}$ = 4.6663e-04, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(7)}$ = 4.7780e-04

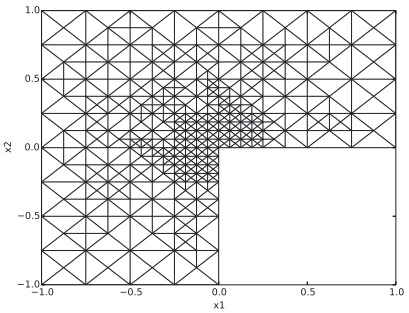


(f) $Q^{(7)}$: 2500 EL, 1296 DOF
$e_{\mathrm{d}}^{2,(7)}$ = 5.0477e-04, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(7)}$ = 5.1600e-04

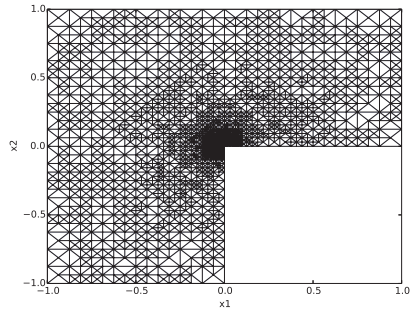
FIGURE 15 Example 3. Evolution of meshes on $Q^{(k)}$, $k = 3, 5, 7$, after refinements based on error (a), (c), (e) and on indicator (b), (d), (f), using marking $\mathbb{M}_{\mathrm{AVR}}$.



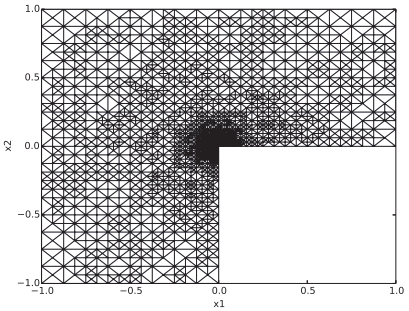
(a) $Q^{(3)}$: 281 EL, 156 ND
$e_{\mathrm{d}}^{2,(3)}$ = 2.6805e-03, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(3)}$ = 2.9591e-03



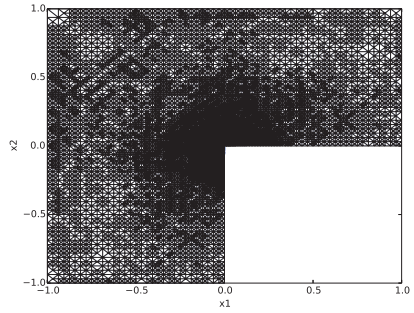
(b) $Q^{(3)}$: 293 EL, 162 ND
$e_{\mathrm{d}}^{2,(3)}$ = 2.6876e-03, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(3)}$ = 2.9662e-03



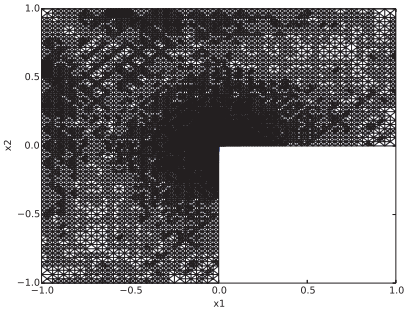
(c) $Q^{(5)}$: 1569 EL, 823 ND
$e_{\mathrm{d}}^{2,(5)}$ = 8.3188e-04, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(5)}$ = 9.6882e-04



(d) $Q^{(5)}$: 1504 EL, 788 ND
$e_{\mathrm{d}}^{2,(5)}$ = 8.4416e-04, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(5)}$ = 9.8012e-04



(e) $Q^{(7)}$: 8762 EL, 4465 ND
$e_{\mathrm{d}}^{2,(7)}$ = 2.7680e-04, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(7)}$ = 2.8537e-04



(f) $Q^{(7)}$: 9529 EL, 4855 ND
$e_{\mathrm{d}}^{2,(7)}$ = 2.8075e-04, $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(7)}$ = 2.8923e-04

FIGURE 16 Example 3. Evolution of meshes on $Q^{(k)}$, $k = 3, 5, 7$, after refinement based on error (a), (c), (e) and on indicator (b), (d), (f), using bulk marking $\mathbb{M}_{0.3}$.

The optimal convergence test of indicator $\overline{\mathrm{m}}_{\mathrm{d}}^2$ is provided in Figure 13a (taking into account that $v \in \mathrm{P}_1$ and $y \in \mathrm{RT}_1$). Analogously, we fix the time-step ($K = 10$) and refine the mesh discretizing Ω. As expected, the speed of convergence of both the error and the majorant lies between linear and quadratic. In Figure 13b, the total error and majorant convergence are provided. The difference in decay of $\overline{\mathrm{m}}_{\mathrm{d}}^2$ and $\overline{\mathrm{M}}_{\mathrm{I}}^2$ can be explained by the presence of the term $-v_t$ in the equilibrium part of the majorant $\overline{\mathrm{m}}_{\mathrm{f}}^2$ and possible accumulation of the error in the flux y (in addition to the accumulation of the error in v).

The distribution of the local errors on $Q^{(10)}$ is indicated quite efficiently by $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(10)}$ (see Figure 14). Figure 14a provides information about the error and majorant distribution on $Q^{(10)}$, where Ω is discretized by the mesh with 192 EL. Figure 14b illustrates analogous characteristics for the mesh with 786 EL. Both figures confirm that $\overline{\mathrm{m}}_{\mathrm{d}}^{2,(k)}$ manages to locate errors associated with corner singularities. Moreover, it is easy to note that the error indicator performs better on the refined mesh with 192 EL.

Finally, we consider the adaptive refinement with two marking procedures and analyze the obtained meshes. Figure 15 shows meshes, obtained by using marker $\mathbb{M}_{\mathrm{AVR}}$, and Figure 16 compares derived meshes after bulk marking $\mathbb{M}_{0.3}$. Analogously to Example 1, we compare the meshes generated during the refinement based on the local true error distribution (LHS) and the local indicator (RHS).

Due to the main drawback of the incremental method (i.e., being time consuming), the space-time FEM, which can be easily parallelized, has been developed. Monograph [60] introduces a scheme that executes a multigrid method for the elliptic problem on each time-step, such that the time is treated as an axis in the space-time grid. Later, more space-time discretization methods were suggested, i.e., the so-called parallel time-stepping method [149], the multigrid waveform relaxation method (space parallelism) [140], and the full space-time multigrid method [63].

Due to the fact that the majorant is formulated on the whole given time-interval, we can apply it to the solution obtained on a discretized space-time cylinder. Therefore, for minimization of $\overline{\mathrm{M}}_{\mathrm{I}}^2$, we follow Algorithm 3.2 in [93, Section 3.3.1]. The examples below discuss the obtained numerical results for '1d + t'- and '2d + t'-dimensional problems.

Example 4. First, we study the numerical properties of $\overline{\mathrm{M}}_{\mathrm{I}}^2$ and the indicator on the unit interval $\Omega = (0, 1) \subset \mathbb{R}$ and $T = 1$ with homogeneous Dirichlet BC. The exact solution is $u = x\,(1 - x)\,(t^2 + t + 1)$ with the IC is $u_0 = x\,(1 - x)$. The approximation v is reconstructed by P_1-FEs (see Figure 17), and the flux y by P_2-FEs. In the current implementation, the time is treated as an extra dimension. Therefore, after several refinement iterations, we can study the optimal convergence of the majorant (see Figure 18a). The plot confirms the quadratic speed of convergence. In Figure 18b, we compare the decay of $\overline{\mathrm{M}}_{\mathrm{I}}^2$ on the uniformly refined

mesh and on the adaptively refined one (using bulk marker $\mathbb{M}_{0.2}$). It becomes clear that for such a problem of non-complicated domain, it is more efficient to do geometric refinement and apply methods based on tensor representations of the data, e.g.., in the way how it is applied to the Fokker–Planck or chemical master equations using tensor train or quantized tensor train formats (see [40, 41]).

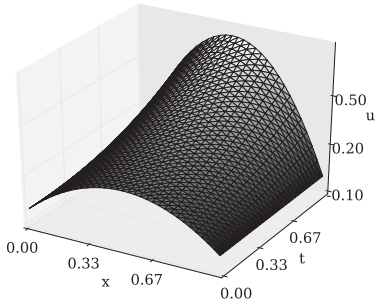
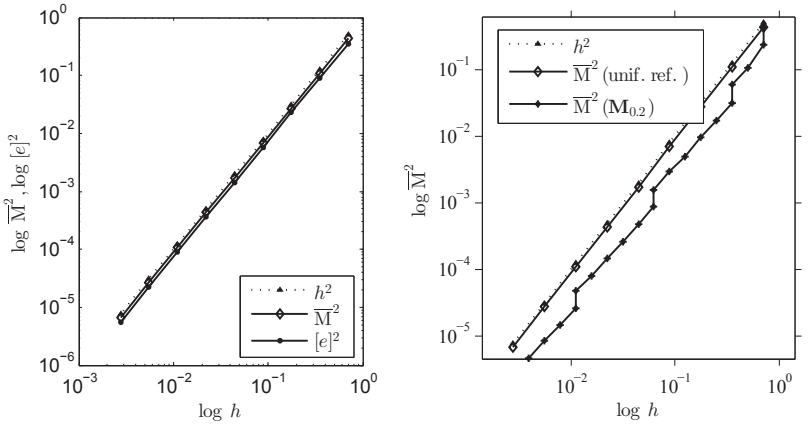


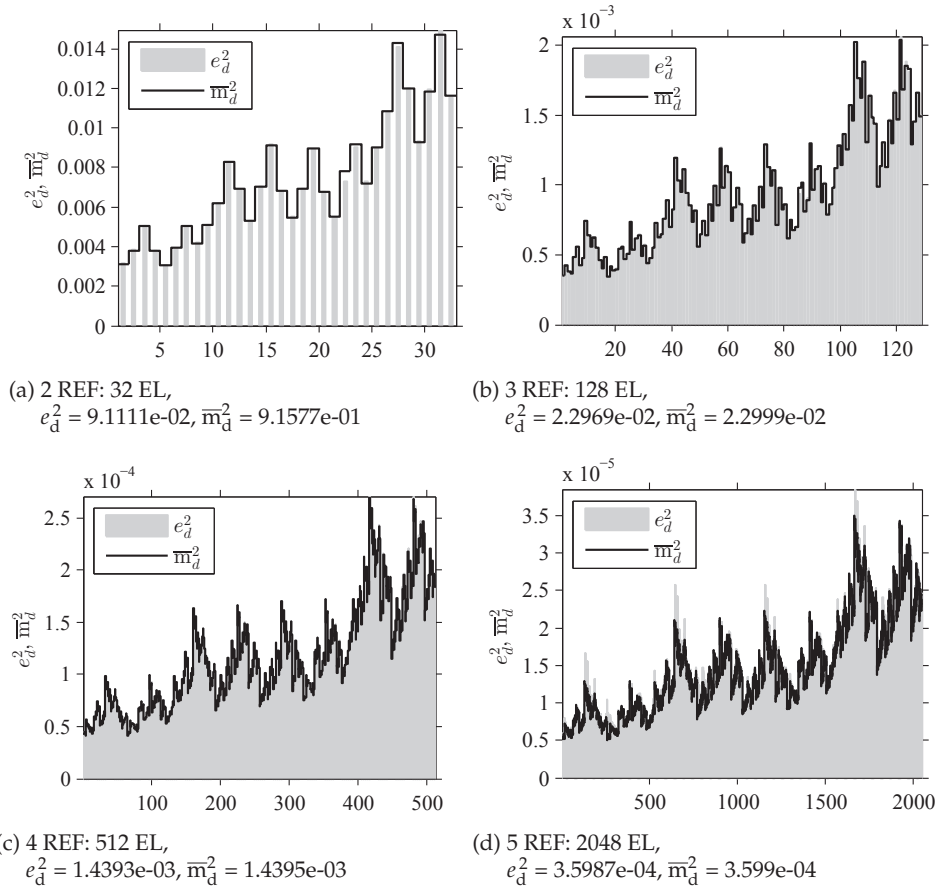
FIGURE 17 Example 4. Approximate solution on the mesh (417 ND), refinement step (REF) 4.



(a) (b)

FIGURE 18 Example 4. (a) Optimal convergence of the total error and majorant. (b) Decay of the majorant on the uniformly refined mesh and on the adaptively refined one.

Next, we consider true error and majorant distributions obtained in each refinement step (see Figure 19). We illustrate e_{d}^2 and $\overline{\mathrm{m}}_{\mathrm{d}}^2$ with respect to EL (numbered by the FE code implementation) after REF 2, 3, 4, 5 in Figures 19a, 19b, 19c, and 19d, respectively (the initial mesh is $\mathcal{T}_{3\times 3}$). The graphic confirms that



(a) 2 REF: 32 EL,
e_{d}^2 = 9.1111e-02, $\overline{\mathrm{m}}_{\mathrm{d}}^2$ = 9.1577e-01

(b) 3 REF: 128 EL,
e_{d}^2 = 2.2969e-02, $\overline{\mathrm{m}}_{\mathrm{d}}^2$ = 2.2999e-02

(c) 4 REF: 512 EL,
e_{d}^2 = 1.4393e-03, $\overline{\mathrm{m}}_{\mathrm{d}}^2$ = 1.4395e-03

(d) 5 REF: 2048 EL,
e_{d}^2 = 3.5987e-04, $\overline{\mathrm{m}}_{\mathrm{d}}^2$ = 3.599e-04

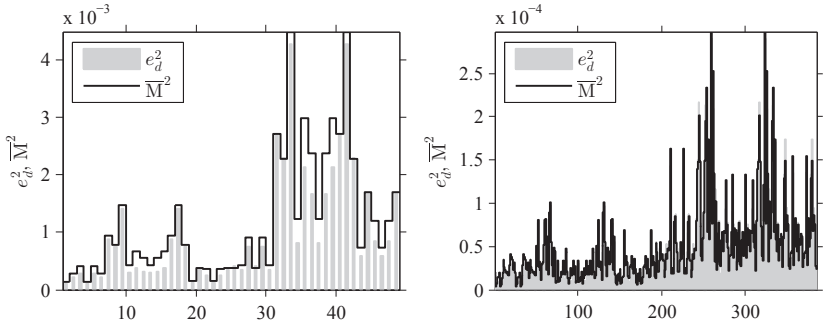
FIGURE 19 Example 4. e_{d}^2 and $\overline{\mathrm{m}}_{\mathrm{d}}^2$ distributions after REF # $= 2, 3, 4, 5$.

the indicator indeed manages to mimic the error distribution and to catch the local jumps very efficiently. Table 4 provides a comprehensive confirmation of the efficiency of the total majorant.

TABLE 4 Example 4. Total error, the majorant, and the efficiency index with respect to the refinement steps.

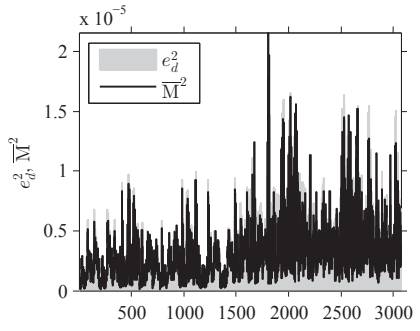
# REF	# EL	$[e]^2$	$\overline{\mathrm{M}}_{\mathrm{I}}^2$	I_{eff}
1	8	3.5229e-01	4.0889e-01	1.08
2	32	9.1112e-02	1.0682e-01	1.08
3	128	2.2969e-02	2.7215e-02	1.09
4	512	5.7541e-03	6.8586e-03	1.09
5	2048	1.4393e-03	1.7209e-03	1.09
6	8192	3.5987e-04	4.3096e-04	1.09
7	131072	2.2493e-05	2.6969e-05	1.09
8	524288	5.6231e-06	6.7435e-06	1.10
9	2097152	1.4058e-06	1.6861e-06	1.10

Example 5. Finally, we study the same problem discussed in Example 1 from the point of view of the space-time discretization. The FE spaces used are as follows: $v \in \mathrm{P}_1$ and $y \in \mathrm{P}_2$. We consider two meshes obtained after the uniform refinement steps REF 1, 2, 3, 4 (see Figures 20a, 20b, 20c, and 20d). Again, the local error and indicator distributions are shown element-wise, where EL are numbered by the algorithm implemented in the code. Table 5 provides the information about the efficiency index of the majorant on every REF.

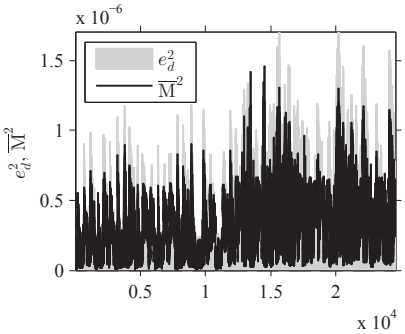


(a) 1 REF: 48 EL
e_{d}^2 = 4.7719e-02, $\overline{\mathrm{M}}_{\mathrm{I}}^2$ = 5.6030e-02

(b) 2 REF: 384 EL
e_{d}^2 = 1.7268e-02, $\overline{\mathrm{M}}_{\mathrm{I}}^2$ = 1.9175e-02



(c) 3 REF: 3072 EL
e_{d}^2 = 8.4558e-03, $\overline{\mathrm{M}}_{\mathrm{I}}^2$ = 9.4992e-03



(d) 4 REF: 24576 EL
e_{d}^2 = 6.1427e-03, $\overline{\mathrm{M}}_{\mathrm{I}}^2$ = 6.9789e-03

FIGURE 20 Example 5. Distribution of error and majorant after refinement steps # = 1, 2, 3, 4.

TABLE 5 Example 5. Total error, the majorant, and the efficiency index with respect to the refinement steps.

# REF	# EL	$[e]^2$	$\overline{\mathrm{M}}_{\mathrm{I}}^2$	I_{eff}
1	48	4.7719e-02	5.6030e-02	1.08
2	384	1.7268e-02	1.9175e-02	1.05
3	3072	8.4558e-03	9.4992e-03	1.06
4	24576	6.1427e-03	6.9789e-03	1.07
5	196608	5.5479e-03	6.2617e-03	1.06

3.5 Guaranteed error estimates for problems on a decomposed domain

In the current section, we present another important result of our work, i.e., error estimates that can be applied for more realistic problems defined on polygonal (polyhedral) domains of a complicated structure and having a mixed BC with non-trivial input functions. The main drawback of the majorant defined in Theorem 3.1 is that it contains global Friedrichs' $C_{\mathrm{F}\Omega}$ and trace $C_{\mathrm{T}\Gamma_{\mathrm{R}}}$ constants, which are hard to calculate or reliably estimate on complex domains. From a numerical point of view, the task is equivalent to the reconstruction of a guaranteed lower bound of the least eigenvalue for the respective differential operator. Therefore, in [PIII, PIV] we suggest the idea of decomposing domain Ω (DD) into a collection of simple non-overlapping sub-domains, such that their characteristics quantities, i.e., constants in (2.7), (2.8) and (2.9), are known or can be estimated (see, e.g., [99, 113, 64, 71, PV]). The method of DD and its numerical efficiency in relation to FEM and boundary element method (BEM) have been thoroughly studied in [82, 28, 59].

Below, we consider the main idea of the suggested approach. Let Ω be decomposed into a collection $\mathcal{O}_\Omega$ of non-overlapping sub-domains $\overline{\Omega} := \cup_{\Omega_i \in \mathcal{O}_\Omega} \overline{\Omega}_i$, $i = 1, \dots, N$. Next, we sort the elements of the obtained collection into two different sets $\mathcal{O}_{\mathrm{P}}$ and $\mathcal{O}_O$ according to the values of the parameter λ, i.e., $\lambda|_{\mathcal{O}_{\mathrm{P}}} \geq \mathrm{P}$ and $\lambda|_{\mathcal{O}_O} \leq \mathrm{P}$, respectively. The sorting is motivated by the strategy of the derivation of error estimates, which depends on the behavior of reaction function λ. For the elements of $\mathcal{O}_O$, we impose the condition

$$\left\{ \mathbf{r}_f^{1-\mu}(v, y) \right\}_{\Omega_i \in \mathcal{O}_O} = 0, \quad \text{for a. a.} \quad t \in]0, T[, \tag{3.35}$$

in order to apply (2.7). Since v and y are in our disposal, condition (3.35) is not difficult to satisfy technically. Subdomains from $\mathcal{O}_{\mathrm{P}}$ are treated by standard arguments due to the presence of a weak term with a relatively large λ. Due to the decomposition of Ω, we obtain $\Gamma_R := \cup_{\Gamma_{Rj} \in \mathcal{S}_R} \Gamma_{Rj}$ such that $\Gamma_{Rj} = \partial\Omega_j \cap \Gamma_R$, $j = 1, \dots, M$, $M \leq N$. Analogously, by imposing local conditions

$$\left\{ \mathbf{r}_\sigma(v, y) \right\}_{\Gamma_{Rj} \in \mathcal{S}_R} = 0, \quad \text{for a. a.} \quad t \in]0, T[, \tag{3.36}$$

constant $C_{\mathrm{T}\Gamma_{\mathrm{R}}}$ can be excluded from (3.13), which is especially advantages if the Robin boundary condition (2.16) is inhomogeneous with the non-trivial input function. Provided that (3.36) is satisfied, we can use (2.7) in order to estimate the error in the residual (3.11).

Based on the above-presented DD into collections $\mathcal{O}_{\mathrm{P}}$, $\mathcal{O}_0$, $\mathcal{S}_{R_j}$ (see Figure 21), we construct complexes

$$R_{\mathcal{O}_{\mathrm{P}},\{\cdot\}}(t) := \sum_{\Omega_l \in \mathcal{O}_{\mathrm{P}}} \frac{|\Omega_l|}{\mathrm{P}^2} \left\{ \mathbf{r}_f^{1-\mu} \right\}_{\Omega_l}^2, \qquad R_{\mathcal{O}_{\mathrm{P}},\|\cdot\|}(t) := \sum_{\Omega_l \in \mathcal{O}_{\mathrm{P}}} \frac{C_{\mathrm{P}\Omega_l}^2}{\lambda_A} \left\| \mathbf{r}_f^{1-\mu} \right\|_{\Omega_l}^2,$$

$$R_{\mathcal{O}_0}(t) := \sum_{\Omega_k \in \mathcal{O}_O} \frac{C^2_{\mathrm{P}\Omega_k}}{\lambda_A} \left\| \mathbf{r}_f^{1-\mu} \right\|^2_{\Omega_k}, \qquad R_{\mathcal{S}_R}(t) := \sum_{\Gamma_{Rj} \in \mathcal{S}_R} \frac{\left(C^{\mathrm{t}}_{\Gamma_{Rj}}\right)^2}{\lambda_A} \|\mathbf{r}_\sigma\|^2_{\Gamma_{Rj}}, \quad (3.37)$$

where the residual functionals $\mathbf{r}_f^{1-\mu}(v,y)$ and $\mathbf{r}_\sigma(v,y)$ are defined in (3.12) and (3.11), respectively. Then generalized estimates follow from Theorem 3.5.

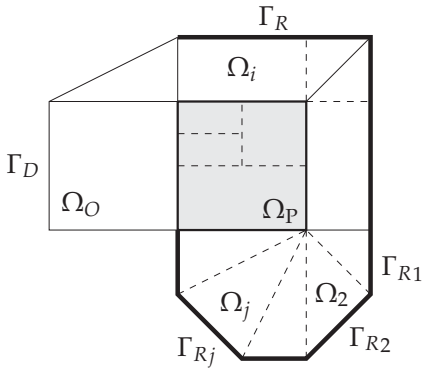


FIGURE 21 Example of the domain decomposition.

Theorem 3.5. *Assume that (3.35) and (3.36) hold. Then for any* $v \in V_0^{1,1}(Q_T)$ *and* $y \in Y_{\mathrm{div}}(Q_T)$ *and* $\delta \in (0,2]$, $\rho_1(t) \in \left[(2-\frac{1}{\rho_2})^{-1}, +\infty\right[$, $\rho_2(t) \in [1, +\infty[$, *the following estimate holds:*

$$[\,e\,]^2_{(v,\theta,1,2)} \le \overline{\mathrm{M}}^2_{\mathrm{I,N}}(v,y;\delta,\rho_1,\rho_2,\mu) := \int_0^T \Big(\rho_1 \left\| \tfrac{1}{\lambda}\, \mathbf{r}_f^{\mu} \right\|^2_{\Omega} + \rho_2 R_{\mathcal{O}_{\mathrm{P}},\{\cdot\}}(t) + \alpha_1 \|\mathbf{r}_A\|^2_{A^{-1}} + \alpha_2 \Big(R_{\mathcal{O}_{\mathrm{P}},\|\cdot\|}(t) + R_{\mathcal{O}_O}(t) \Big) + \alpha_3 R_{\mathcal{S}_R}(t) \Big) \mathrm{d}t. \quad (3.38)$$

Here, $\mathbf{r}_f(v,y)$ *and* $\mathbf{r}_A(v,y)$ *are defined in (3.9) and (3.10), respectively, and* $R_{\mathcal{O}_{\mathrm{P}},\{\cdot\}}$, $R_{\mathcal{O}_{\mathrm{P}},\|\cdot\|}$, $R_{\mathcal{O}_O}$, *and* $R_{\mathcal{S}_R}$ *are determined in* (3.37), $\nu = 2 - \delta$, $\theta(x,t) = \lambda(x)\big(2 - \frac{1}{\rho_1} - \frac{1}{\rho_2}\big)^{1/2}$ *are positive weights,* $\mu(x,t) \in [0,1]$, *and* $\alpha_1(t)$, $\alpha_2(t)$, $\alpha_3(t)$ *are positive scalar-valued functions satisfying the relation* (3.15).

Proof. See, e.g., Theorem 3 (i) in [PIV]. □

In [PIV, Section 3.3], we show that the obtained estimate (3.38) is equivalent to the error measured in the primal-dual norm. By analogous methods used in Theorem 3.5, we obtain the form of the advanced majorant (3.23) applied for the decomposed Ω, which is equivalent to the error measured by (3.7) (see [PIV, Section 3.5]). Due to the fact that minorant in Theorem 3.4 does not contain any global constants, it can be straightforwardly rewritten for the case of a decomposed domain.

3.6 Sharp bounds of constants in Poincaré-type inequalities

In order to apply the estimates discussed in Section 3.5, we must have in our disposal exact values or reliable estimates of the constants in (2.7), (2.8), and (2.9). Due to the use of FE approximations here, the constants in the above-mentioned inequalities on arbitrary nondegenerated triangles and tetrahedrons are of main interest. Therefore, the work presented in [PV] is dedicated to sharp upper bounds of the constants in classic Poincaré and Poincaré-type inequalities for functions with zero mean on the boundary of arbitrary nondegenerated simplexes in $\mathbb{R}^2$ and $\mathbb{R}^3$.

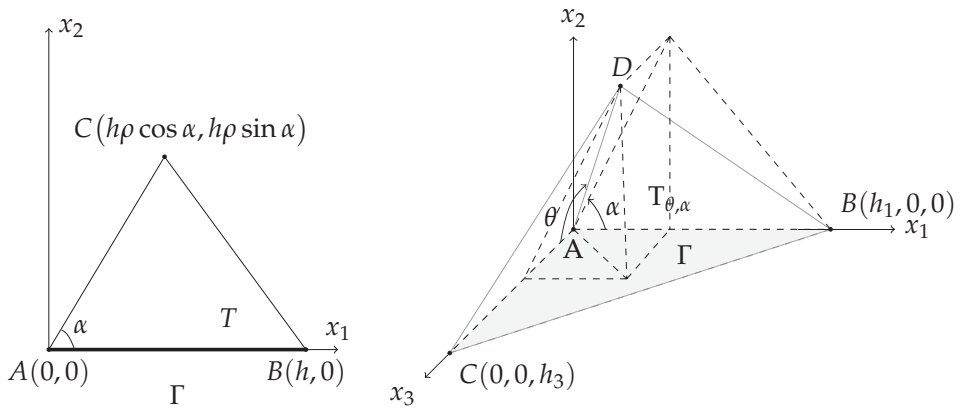


FIGURE 22 Simplices in (a) $\mathbb{R}^2$ and (b) $\mathbb{R}^3$.

For $\mathbb{R}^2$, we consider two basic triangles $\widehat{\mathrm{T}}_{\pi/2} := \mathrm{conv}\big\{(0,0),(1,0),(0,1)\big\}$ and $\widehat{\mathrm{T}}_{\pi/4} := \mathrm{conv}\big\{(0,0),(1,0),\big(\frac{1}{2},\frac{1}{2}\big)\big\}$, with $\widehat{\Gamma} := \big\{x_1 \in [0,1],\ x_2 = 0\big\}$, in order to obtain bounds of constants in (2.8) and (2.9). The corresponding exact values of the constants were recalled in Section 2.1, i.e., $C^{\mathrm{P}}_{\widehat{\Gamma},\pi/2} \approx 0.49291$, $C^{\mathrm{Tr}}_{\widehat{\Gamma},\pi/2} \approx 0.65602$ and $C^{\mathrm{P}}_{\widehat{\Gamma},\pi/4} \approx 0.24646$, $C^{\mathrm{Tr}}_{\widehat{\Gamma},\pi/4} \approx 0.70711$. The estimates of C^{P}_{Γ} and C^{Tr}_{Γ} follow from the Lemma 3.1 [PV, Section 2].

Lemma 3.1. *For any $w \in \widetilde{H}^1(\mathrm{T},\Gamma)$ defined on simplex*

$$\mathrm{T} := \mathrm{conv}\big\{(0,0),(h,0),\big(h\rho\cos\alpha,\ h\rho\ \sin\alpha\big)\big\}$$

(see Figure 22a) with face $\Gamma := \big\{x_1 \in [0,h],\ \ x_2 = 0\big\}$, the Poincaré-type inequalities

$$\|w\|_{\mathrm{T}} \le C^{\mathrm{P}}_{\Gamma}\, h\, \|\nabla w\|_{\mathrm{T}} \quad \text{and} \quad \|w\|_{\Gamma} \le C^{\mathrm{Tr}}_{\Gamma}\, h^{1/2}\, \|\nabla w\|_{\mathrm{T}}$$

hold with

$$C_\Gamma^{\mathrm{P}} \le \overline{C}_\Gamma^{\mathrm{P}} := \min\Big\{\bar{c}_{\mathrm{P},\pi/2}\, C^{\mathrm{P}}_{\widehat{\Gamma},\pi/2},\, \bar{c}_{\mathrm{P},\pi/4}\, C^{\mathrm{P}}_{\widehat{\Gamma},\pi/4}\Big\},$$
$$C_\Gamma^{\mathrm{Tr}} \le \overline{C}_\Gamma^{\mathrm{Tr}} := \min\Big\{\bar{c}_{\mathrm{Tr},\pi/2}\, C^{\mathrm{Tr}}_{\widehat{\Gamma},\pi/2},\, \bar{c}_{\mathrm{Tr},\pi/4}\, C^{\mathrm{Tr}}_{\widehat{\Gamma},\pi/4}\Big\},$$

respectively. Here, the weighting parameters $\bar{c}_{\mathrm{P},\pi/2} = \mu_{\pi/2}^{1/2}$, $\bar{c}_{\mathrm{P},\pi/4} = \mu_{\pi/4}^{1/2}$, $\bar{c}_{\mathrm{Tr},\pi/2} = \big(\rho\,\sin\alpha\big)^{-1/2}\,\bar{c}_{\mathrm{P},\pi/2}$, $\bar{c}_{\mathrm{Tr},\pi/4} = \big(2\rho\,\sin\alpha\big)^{-1/2}\,\bar{c}_{\mathrm{P},\pi/4}$,

$$\mu_{\pi/2}(\rho,\alpha) = \tfrac{1}{2}\Big(1+\rho^2+\big(1+\rho^4+2\,\cos 2\alpha\big)\,\rho^2\big)^{1/2}\Big), \tag{3.39}$$
$$\mu_{\pi/4}(\rho,\alpha) = 2\rho^2 - 2\rho\,\cos\alpha + 1 + \tag{3.40}$$
$$\big((2\rho^2+1)(2\rho^2+1-4\rho\,\cos\alpha+4\rho^2\,\cos 2\alpha)\big)^{1/2}, \tag{3.41}$$

and $C^{\mathrm{P}}_{\widehat{\Gamma},\pi/2}$, $C^{\mathrm{Tr}}_{\widehat{\Gamma},\pi/2}$ *and* $C^{\mathrm{P}}_{\widehat{\Gamma},\pi/4}$, $C^{\mathrm{Tr}}_{\widehat{\Gamma},\pi/4}$ *are the constants in (2.8) and (2.9) for reference triangles* $\widehat{\mathrm{T}}_{\pi/2}$ *and* $\widehat{\mathrm{T}}_{\pi/4}$, *respectively.*

Proof. See [PV, Lemma 1, Section 2]. □

Analogously to Lemma 3.1, the upper bound of the constant in (2.7) can be obtained. In addition to earlier defined $\mathrm{T}_{\pi/4}$ and $\mathrm{T}_{\pi/2}$, we consider third reference triangle $\mathrm{T}_{\pi/3} := \mathrm{conv}\big\{(0,0),(1,0),(\frac{1}{2},\frac{\sqrt{3}}{2})\big\}$. Consequently, the upper bound of C^{P}_{Ω} in (2.7) follows from the Lemma 3.2 [PV, Section 2].

Lemma 3.2. *For any* $w \in \widetilde{H}^1(\mathrm{T})$, *the estimate of the constant in*

$$\|w\|_{\mathrm{T}} \;\le\; C^{\mathrm{P}}_{\Omega} h\, \|\nabla w\|_{\mathrm{T}} \tag{3.42}$$

has the form

$$C^{\mathrm{P}}_{\Omega} \le \overline{C}^{\mathrm{P,MR}}_{\Omega} = \min\Big\{\bar{c}_{\pi/4}\, C^{\mathrm{P}}_{\widehat{\Gamma},\pi/4},\, \bar{c}_{\pi/3}\, C^{\mathrm{P}}_{\widehat{\Gamma},\pi/3},\, \bar{c}_{\pi/2}\, C^{\mathrm{P}}_{\widehat{\Gamma},\pi/2}\Big\} \tag{3.43}$$

Here,

$$\bar{c}_{\pi/4} = \mu_{\pi/4}^{1/2}, \quad \bar{c}_{\pi/3} = \mu_{\pi/3}^{1/2}, \quad \textit{and} \quad \bar{c}_{\pi/2} = \mu_{\pi/2}^{1/2},$$

where $\mu_{\pi/2}$ *and* $\mu_{\pi/4}$ *are defined in* (3.39) *and* (3.41) *and*

$$\mu_{\pi/3}(\rho,\alpha) = \tfrac{2}{3}(1+\rho^2-\rho\,\cos\alpha) + 2\big(\tfrac{1}{9}(1+\rho^2-\rho\,\cos\alpha)^2 - \tfrac{1}{3}\rho^2\sin^2\alpha\big)^{1/2},$$

and $C^{\mathrm{P}}_{\widehat{\mathrm{T}},\pi/4} = \frac{1}{\sqrt{2}\pi}$, $C^{\mathrm{P}}_{\widehat{\mathrm{T}},\pi/3} = \frac{3}{4\pi}$, *and* $C^{\mathrm{P}}_{\widehat{\mathrm{T}},\pi/2} = \frac{1}{\pi}$.

Proof. See [PV, Lemma 2, , Section 2]. □

In [PV, Section 3], the bounds of $\overline{C}^{\mathrm{P}}_{\Gamma}$, $\overline{C}^{\mathrm{Tr}}_{\Gamma}$, and $\overline{C}^{\mathrm{P,MR}}_{\Omega}$ (provided by Lemmas 3.1 and 3.2) are compared with the corresponding minorants, which can be found

by minimization of the Rayleigh quotients

$$\mathcal{R}^{\mathrm{P}}_{\Gamma}[w] = \frac{\|\nabla w\|_{\mathrm{T}}}{\|w-\{w\}_{\Gamma}\|_{\mathrm{T}}}, \quad \mathcal{R}^{\mathrm{Tr}}_{\Gamma}[w] = \frac{\|\nabla w\|_{\mathrm{T}}}{\|w-\{w\}_{\Gamma}\|_{\Gamma}}, \quad \text{and} \quad \mathcal{R}^{\mathrm{P}}_{\Omega}[w] = \frac{\|\nabla w\|_{\mathrm{T}}}{\|w-\{w\}_{\Gamma}\|_{\Gamma}}, \tag{3.44}$$

for all $w \in \widetilde{H}^1(\mathrm{T})$. Moreover, [PV] provides the comparison of $\overline{C}^{\mathrm{P,MR}}_{\Omega}$ with existing estimates from [86, 85] and [31]. Finally, we also discuss the structure of minimizers for Rayleigh quotients (3.44) and their behavior with respect to changing parameters.

The same method can be applied for non-degenerate tetrahedrons $T \in \mathbb{R}^3$ presented in the form

$$\mathrm{T} = \mathrm{conv}\big\{(0,0,0),(h_1,0,0),(0,0,h_3),(D_{x_1},D_{x_2},D_{x_3})\big\}, \tag{3.45}$$

where $(D_{x_1},D_{x_2},D_{x_3}) = (h_1\rho\cos\alpha\,\sin\theta,\, h_1\rho\sin\alpha\,\sin\theta,\, h_1\rho\cos(\theta))$, h_1 and h_3 are the scaling parameters along axises O_{x_1} and O_{x_3}, respectively, α is a polar angle, and θ is an azimuthal angle (see Fig. 22b). Let zero mean condition be imposed on $\Gamma = \mathrm{conv}\big\{(0,0,0),(h_1,0,0),(0,0,h_3)\big\}$, and $\widehat{\mathrm{T}}_{\hat\theta,\hat\alpha}$ denote the reference tetrahedron, where $\hat\theta$ and $\hat\alpha$ are fixed angles. Then, by $\mathcal{F}_{\hat\theta,\hat\alpha}$ we denote the respective mapping $\mathcal{F}_{\hat\theta,\hat\alpha} : \widehat{\mathrm{T}}_{\hat\theta,\hat\alpha} \to \mathrm{T}$.

To the best of our knowledge, exact values of constants in Poincaré-type inequalities for simplexes in $\mathbb{R}^3$ are unknown. Therefore, in [PV] some reference cases are calculated numerically with high accuracy and listed in Table 6. Based on these data, we present the (approximate) bounds for an arbitrary tetrahedron T:

$$\begin{aligned}
&\|v\|_{\mathrm{T}} \le \widetilde{C}^{\mathrm{P}}_{\Gamma}\, h_1\, h_3\, \|\nabla v\|_{\mathrm{T}}, \qquad \widetilde{C}^{\mathrm{P}}_{\Gamma} = \min_{\hat\alpha=\{\pi/4,\pi/3,\pi/2,2\pi/3\}} \Big\{ c^{\mathrm{P}}_{\pi/2,\hat\alpha}\, C^{\mathrm{P}}_{\widehat\Gamma,\pi/2,\hat\alpha} \Big\},\\
&\|v\|_{\Gamma} \le \widetilde{C}^{\mathrm{Tr}}_{\Gamma}\, (h_1\, h_3)^{\frac{1}{2}}\, \|\nabla v\|_{\mathrm{T}}, \quad \widetilde{C}^{\mathrm{Tr}}_{\Gamma} = \min_{\hat\alpha=\{\pi/4,\pi/3,\pi/2,2\pi/3\}} \Big\{ c^{\mathrm{Tr}}_{\pi/2,\hat\alpha}\, C^{\mathrm{Tr}}_{\widehat\Gamma,\pi/2,\hat\alpha} \Big\},
\end{aligned} \tag{3.46}$$

where $C^{\mathrm{P}}_{\widehat\Gamma,\pi/2,\hat\alpha}$ and $C^{\mathrm{Tr}}_{\widehat\Gamma,\pi/2,\hat\alpha}$ are the constants related to four reference thetrahedron from Table 6 and

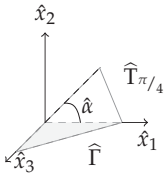
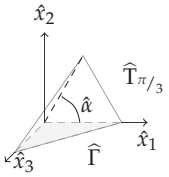
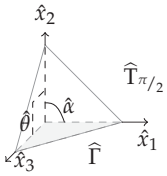
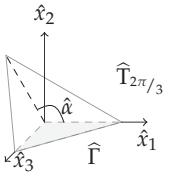
$$c^{\mathrm{P}}_{\pi/2,\hat\alpha} = \frac{\mu^{1/2}_{\pi/2,\hat\alpha}}{h_1 h_3}, \quad c^{\mathrm{Tr}}_{\pi/2,\hat\alpha} = \Big(\frac{h_3\sin\hat\alpha}{\rho\sin\alpha\sin\theta}\Big)^{1/2} c^{\mathrm{P}}_{\pi/2,\hat\alpha}.$$

are the ratios of the mapping $\mathcal{F}_{\pi/2,\hat\alpha}$: $\widehat{\mathrm{T}}_{\pi/2,\hat\alpha} \to \mathrm{T}$. Here, $\widehat{\mathrm{T}}_{\pi/2,\hat\alpha} := \mathrm{conv}\big\{(0,0,0),(1,0,0),(0,0,1),(\cos\hat\alpha,\sin\hat\alpha,0)\big\}$ with $\hat\alpha = \{\frac{\pi}{4},\frac{\pi}{3},\frac{\pi}{2},\frac{2\pi}{3}\}$, T is defined in (3.45), and $\mathcal{F}_{\pi/2,\hat\alpha}(\hat x)$ is presented by the relation

$$x = \mathcal{F}_{\pi/2,\hat\alpha}(\hat x) = B_{\pi/2,\hat\alpha}\hat x, \quad B_{\pi/2,\hat\alpha} = \begin{pmatrix} h_1 & \frac{h_1}{\sin\hat\alpha}(\rho\cos\alpha\sin\theta - \cos\hat\alpha) & 0\\ 0 & h_1\rho\frac{\sin\alpha\sin(\theta)}{\sin\hat\alpha} & 0\\ 0 & h_1\rho\frac{\cos\theta}{\sin\hat\alpha} & h_3 \end{pmatrix}.$$

The numerical tests and detailed discussions of the practical aspects of this

TABLE 6 $C^{\mathrm{P},M}_{\widehat{\Gamma},\pi/2,\hat{\alpha}}$ and $C^{\mathrm{Tr},M}_{\widehat{\Gamma},\pi/2,\hat{\alpha}}$ with respect to $M(N)$ for $\widehat{\mathrm{T}}_{\hat{\theta},\hat{\alpha}}$ with $\rho = 1$, $\hat{\theta} = \frac{\pi}{2}$, and several $\hat{\alpha}$.

	$\hat{\theta} = \frac{\pi}{2}, \hat{\alpha} = \frac{\pi}{4}$		$\hat{\theta} = \frac{\pi}{2}, \hat{\alpha} = \frac{\pi}{3}$	
				
$M(N)$	$C^{\mathrm{P},M}_{\widehat{\Gamma},\pi/2,\hat{\alpha}}$	$C^{\mathrm{Tr},M}_{\widehat{\Gamma},\pi/2,\hat{\alpha}}$	$C^{\mathrm{P},M}_{\widehat{\Gamma},\pi/2,\hat{\alpha}}$	$C^{\mathrm{Tr},M}_{\widehat{\Gamma},\pi/2,\hat{\alpha}}$
7	0.32431	0.760099	0.325985	0.654654
26	0.338539	0.829445	0.340267	0.761278
63	0.341122	0.831325	0.342556	0.762901
124	0.341147	0.831335	0.342589	0.762905
215	**0.341147**	**0.831335**	**0.342589**	**0.762905**
	$\hat{\theta} = \frac{\pi}{2}, \hat{\alpha} = \frac{\pi}{2}$		$\hat{\theta} = \frac{\pi}{2}, \hat{\alpha} = \frac{2\pi}{3}$	
				
$M(N)$	$C^{\mathrm{P},M}_{\widehat{\Gamma},\pi/2,\hat{\alpha}}$	$C^{\mathrm{Tr},M}_{\widehat{\Gamma},\pi/2,\hat{\alpha}}$	$C^{\mathrm{P},M}_{\widehat{\Gamma},\pi/2,\hat{\alpha}}$	$C^{\mathrm{Tr},M}_{\widehat{\Gamma},\pi/2,\hat{\alpha}}$
7	0.360532	0.654654	0.4152099	0.686161
26	0.373669	0.751615	0.4274757	0.863324
63	0.375590	0.751994	0.4286444	0.864595
124	0.375603	0.751999	0.4286652	0.864630
215	**0.375603**	**0.751999**	**0.4286652**	**0.864630**

study can be found in [PV]. In [PV, Section 5], we provide an example that shows possible applications of the results and derive a computable majorant of the difference between the exact solution of a boundary value problem and an arbitrary finite dimensional approximation computed on a simplicial mesh. Using the above presented constants, one can weaken the pointwise continuity condition of normal components of the auxiliary flux on inner faces of the mesh in the functional estimates (3.38). Instead, it is enough that the mean values of the normal components are continuous, which allows us to use a wider space for the approximation of the flux.

4 CONCLUSIONS AND OUTLOOK

In this chapter, we summarize the results presented in the study and give an outlook on some future research. The first part of the thesis presents a new version of the Picard–Lindelöf method for nonlinear ODEs supplied with guaranteed and explicitly computable upper bounds of approximation errors. The estimates derived take into account interpolation and integration errors and, therefore, provide objective on the accuracy of computed approximations (see [PI]).

In the second (major) part of the work, guaranteed bounds of distance to the exact solution of the evolutionary reaction-diffusion problem with mixed BC are discussed. We show that two-sided error estimates are directly computable and equivalent to the error. Numerical experiments have confirmed that the estimates provide accurate two-sided bounds of the overall error and generate efficient indicators of local error distribution (see [PII] and Section 3.4 of the current work).

Earlier, we have generalized two-sided bounds to evolutionary reaction-diffusion problems and adapted them to domains of complicated structure with mixed Dirichlet–Robin BC. The estimates are also valid for problems with complicated nonlinear source functions. To overcome computational difficulties, the domain decomposition method was used. To obtain the error estimate, we have exploited the classical Poincaré and Poincaré-type inequalities for functions with zero mean boundary traces. Therefore, the new corresponding bounds of the distance to the exact solution contain only constants in local Poincaré-type inequalities associated with subdomains, which quantitatively improves the majorant value. Moreover, it has been proved that the bounds are equivalent to the primal and primal-dual energy norms of the error (see [PIII, PIV]).

The above-introduced estimates require exact values of guaranteed and realistic bounds of constants in respective functional inequalities. Therefore, in the last part of the thesis, we present sharp estimates of constants in Poincaré and Poincaré-type inequalities for functions having zero mean value on the boundary of a Lipschitz domain or on a measurable part of it. These estimates are particularly used in a posteriori error estimation methods for I-BVPs introduced in [PIII] and [PIV]. Our focus was on computable relations that provide sharp bounds of the constants in the above-mentioned inequalities on simplexes in 2D

and 3D, which, based on numerical simulations, have confirmed to provide efficient numerical results. Also, we have numerically studied the behavior of the constants in the classical Poincaré inequalities and compared these results with known analytical estimates.

In the context of partial differential equations, we have studied only linear models. Thus, extension of these methods to nonlinear I-BVPs is a matter of future work. Moreover, it would be interesting to extend the application of majorant for nonconforming approximations. The estimates based on the domain decomposition technique and local classical Poincaré and Poincaré-type inequalities for functions with zero mean trace obtained in [PIV] can be further developed. Finally, one of the most important directions for the future work is to improve the speed of majorant reconstruction, e.g., to implement a highly parallel algorithm for its minimization.

YHTEENVETO (FINNISH SUMMARY)

Tämä väitöskirja käsittelee funktionaalisia a posteriori virhe-estimaatteja ajasta riippuville ongelmille. Väitöskirja koostuu viidestä julkaisusta, joista ensimmäiset käsittelevät epälineaaristen differentiaaliyhtälöiden Cauchy-ongelmaa, sekä aika-riippuvaista parabolista reaktio-diffuusio-konvektio-ongelmaa. Julkaisuissa tehdään teoreettinen a posteriori virhe-analyysi ja kattava numeerinen analyysi. Kaksi viimeistä julkaisua keskittyvät Poincare-tyyppisiin epäyhtälöihin ja niiden sisältämien vakioiden numeeriseen laskentaan. Näille vakioille johdetaan ylärajat kolmioille ja tetraedreille. Osittaisdifferentiaaliyhtälöiden laskenta-alue diskretoidaan tyypillisesti tämän muotoisiin osa-alueisiin. Näiden vakioiden ylärajat ovat erityisen hyödyllisiä sovellettuna funktionaalisiin virhe-ylärajoihin.

Tässä työssä esitellyt virhe-estimaatit ovat täysin laskettavissa, eli ne eivät sisällä tuntemattomia muuttujia. Ne ovat myös täysin luotettavia, eli tarjoavat aina aidon ylärajan virheelle. Työssä suoritetut kattavat numeeriset kokeet osoittavat funktionaalisten virhe-estimaattien ja niistä johdettujen virhe-indikaattorien tehokkuuden ja luotettavuuden. Kaikki tässä työssä tehdyt numeeriset kokeet suoritettiin MATLABilla ja FEniCSin Python-rajapinnalla.

REFERENCES

[1] A. Agouzal. On the saturation assumption and hierarchical a posteriori error estimator. *Comput. Meth. Appl. Math.*, 2(2):125–131, 2002.

[2] M. Ainsworth. A posteriori error estimation for fully discrete hierarchic models of elliptic boundar–value problems on thin domains. *Numer. Math.*, 80(3):325–362, 1998.

[3] M. Ainsworth and J. T. Oden. A procedure for a posteriori error estimation for *h-p* finite element methods. *Comput. Methods Appl. Mech. Engrg.*, 101(1-3):73–96, 1992. Reliability in computational mechanics (Kraków, 1991).

[4] M. Ainsworth and J. T. Oden. A unified approach to a posteriori error estimation using element residual methods. *Numer. Math.*, 65(1):23–50, 1993.

[5] M. Ainsworth and J. T. Oden. *A posteriori error estimation in finite element analysis*. Wiley and Sons, New York, 2000.

[6] M. Ainsworth and R. Rankin. Fully computable error bounds for discontinuous Galerkin finite element approximations on meshes with an arbitrary number of levels of hanging nodes. *SIAM J. Numer. Anal.*, 47(6):4112–4141, 2010.

[7] G. Akrivis, C. Makridakis, and R. H. Nochetto. Galerkin and Runge-Kutta methods: unified formulation, a posteriori error estimates and nodal superconvergence. *Numer. Math.*, 118(3):429–456, 2011.

[8] I. Babuška and W. C. Rheinboldt. Error estimates for adaptive finite element computations. *SIAM J. Numer. Anal.*, 15(4):736–754, 1978.

[9] I. Babuška and R. Rodríguez. The problem of the selection of an a posteriori error indicator based on smoothening techniques. *Internat. J. Numer. Methods Engrg.*, 36(4):539–567, 1993.

[10] I. Babuška and T. Strouboulis. *The finite element method and its reliability*. Numerical mathematics and scientific computation. The Clarendon Press Oxford University Press, New York, 2001.

[11] I. Babuška and W. C. Rheinboldt. A-posteriori error estimates for the finite element method. *Internat. J. Numer. Meth. Engrg.*, 12:1597–1615, 1978.

[12] I. Babuška, J. R. Whiteman, and T. Strouboulis. *Finite elements, an introduction to the method and error estimation*. Oxford University Press, New York, 2011.

[13] S. Banach. Sur les opérations dans les ensembles abstraits et leur application aux équations intégrales. *Fund. Math.*, 3:133–181, 1922.

[14] W. Bangerth and R. Rannacher. *Adaptive finite element methods for differential equations*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2003.

[15] S. Bartels and C. Carstensen. Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids. Part II: Higher order FEM. *Math. Comput.*, 239(71):971–994, 2002.

[16] 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(4):237–264, 1996.

[17] I. Bendixson. Détermination des équations résolubles algébriquement dans lesquelles chaque racine peut s'exprimer en fonction rationnelle de l'une d'entre elles. *Ann. Fac. Sci. Toulouse Sci. Math. Sci. Phys.*, 7(2):C1–C7, 1893.

[18] P. H. Bérard. Spectres et groupes cristallographiques. I. Domaines euclidiens. *Invent. Math.*, 58(2):179–199, 1980.

[19] M. Besier and R. Rannacher. Goal-oriented space-time adaptivity in the finite element Galerkin method for the computation of nonstationary incompressible flow. *Internat. J. Numer. Methods Fluids*, 70(9):1139–1166, 2012.

[20] F. Black and M. Scholes. The pricing of options and corporate liabilities [reprint of J. Polit. Econ. **81** (1973), no. 3, 637–654]. In *Financial risk measurement and management*, volume 267 of *Internat. Lib. Crit. Writ. Econ.*, pages 100–117. Edward Elgar, Cheltenham, 2012.

[21] D. Braess. *Finite elements*. Cambridge University Press, Cambridge, second edition, 2001. Theory, fast solvers, and applications in solid mechanics, Translated from the 1992 German edition by Larry L. Schumaker.

[22] S. Brenner and R. L. Scott. *The mathematical theory of finite element methods*. Springer, New York, 1994.

[23] J. R. Cannon. *The one-dimensional heat equation*, volume 23 of *Encyclopedia of Mathematics and its Applications*. Addison-Wesley Publishing Company, Advanced Book Program, Reading, MA, 1984.

[24] H. S. Carslow and J. C. Jaeger. *Conduction of Heat in Solids*. Oxford Univ. Press (Clarendon), London and New York, 1948.

[25] C. Carstensen. Quasi–interpolation and a posteriori error analysis of finite element methods. *Mathematical Modelling in Numerical Analysis*, 6(33):1187–1202, 1999.

[26] C. Carstensen and S. A. Funken. Fully reliable localized error control in the FEM. *SIAM J. Sci. Comput.*, 21(4):1465–1484, 2000.

[27] C. Carstensen and J. Gedicke. Guaranteed lower bounds for eigenvalues. *Math. Comp.*, 83(290):2605–2629, 2014.

[28] C. Carstensen, M. Kuhn, and U. Langer. Fast parallel solvers for symmetric boundary element domain decomposition equations. *Numer. Math.*, 79(3):321–347, 1998.

[29] C. Carstensen and S. A. Sauter. A posteriori error analysis for elliptic PDEs on domains with complicated structures. *Numer. Math.*, 96(4):691–721, 2004.

[30] C. Carstensen and R. Verfürth. Edge residuals dominate a posteriori error estimates for low order finite element methods. *SIAM J. Numer. Anal.*, 5(36):1571–1587, 1999.

[31] S. Y. Cheng. Eigenvalue comparison theorems and its geometric applications. *Math. Z.*, 143(3):289–297, 1975.

[32] P. G. Ciarlet. *The finite element method for elliptic problems*. North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978. Studies in Mathematics and its Applications, Vol. 4.

[33] E. A. Coddington and N. Levinson. *Theory of ordinary differential equations*. McGraw-Hill Book Company, Inc., New York-Toronto-London, 1955.

[34] L. Collatz. *Funktionan alysis und numerische Mathematik*. Springer-Verlag, Berlin, 1964.

[35] R. Courant. Variational methods for the solution of problems of equilibrium and vibrations. *Bull. Amer. Math. Soc.*, 49:1–23, 1943.

[36] R. Courant, K. Friedrichs, and H. Lewy. On the partial difference equations of mathematical physics. *IBM J. Res. Develop.*, 11:215–234, 1967.

[37] J. Crank. *The mathematics of diffusion*. Clarendon Press, Oxford, second edition, 1975.

[38] P. Deuflhard, P. Leinen, and H. Yserentant. Concept of an adaptive hierarchical finite element code. *Impact Computing Sci. Engrg.*, 1(1):3–35, 1989.

[39] C. R. Dohrmann, A. Klawonn, and O. B. Widlund. Domain decomposition for less regular subdomains: overlapping Schwarz in two dimensions. *SIAM J. Numer. Anal.*, 46(4):2153–2168, 2008.

[40] S. Dolgov and B. Khoromskij. Simultaneous state-time approximation of the chemical master equation using tensor product formats. *Numer. Linear Algebra Appl.*, 22(2):197–219, 2015.

[41] S. V. Dolgov, B. N. Khoromskij, and I. V. Oseledets. Fast solution of parabolic problems in the tensor train/quantized tensor train format with initial application to the Fokker-Planck equation. *SIAM J. Sci. Comput.*, 34(6):A3016–A3038, 2012.

[42] W. Dörfler. A convergent adaptive algorithm for Poisson's equation. *SIAM J. Numer. Anal.*, 33(3):1106–1124, 1996.

[43] W. Dörfler and R. H. Nochetto. Small data oscillation implies the saturation assumption. *Numer. Math.*, 91(1):1–12, 2002.

[44] W. Dörfler and M. Rumpf. An adaptive strategy for elliptic problems including a posteriori controlled boundary approximation. *Math. Comp.*, 67(224):1361–1382, 1998.

[45] R. Duran, M. A. Muschietti, and R. Rodriguez. On the asymptotic exactness of error estimators for linear triangle elements. *Numer. Math.*, 59(2):107–127, 1991.

[46] T. Eirola, A. M. Krasnosel'skii, M. A. Krasnosel'skii, N. A. Kuznersov, and O. Nevanlinna. Incomplete corrections in nonlinear problems. *Nonlinear Analysis*, 25(7):717–728, 1995.

[47] K. Eriksson and C.Johnson. An adaptive finite element method for linear elliptic problems. *Math. Comp.*, 50(182):361–383, 1988.

[48] L. C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.

[49] N. Filonov. On an inequality for the eigenvalues of the Dirichlet and Neumann problems for the Laplace operator. *Algebra i Analiz*, 16(2):172–176, 2004.

[50] J. B. J. Fourier. Analytical theory of heat. In *Great Books of the Western World, no. 45*, Great Books of the Western World, no. 45, pages 163–251. Encyclopaedia Britannica, Inc., Chicago, London, Toronto, 1952.

[51] A. Friedman. *Partial differential equations of parabolic type*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1964.

[52] K. Friedrichs. On certain inequalities and characteristic value problems for analytic functions and for functions of two variables. *Trans. Amer. Math. Soc.*, 41(3):321–364, 1937.

[53] M. Fuchs. Computable upper bounds for the constants in Poincaré-type inequalities for fields of bounded deformation. *Math. Methods Appl. Sci.*, 34(15):1920–1932, 2011.

[54] A. V. Gaevskaya and S. I. Repin. A posteriori error estimates for approximate solutions of linear parabolic problems. *Springer, Differential Equations*, 41(7):970–983, 2005.

[55] V. Girault, K. Kumar, and M. F. Wheeler. Convergence of iterative coupling of geomechanics with flow in a fractured poroelastic medium. Technical Report ICES REPORT 15-05, The Institute for Computational Engineering and Sciences The University of Texas at Austin, Austin, Texas 78712, 2015.

[56] R. Glowinski. *Numerical methods for nonlinear variational problems*. Springer, New York, 1984.

[57] R. Glowinski, J. L. Lions, and R. Trémolierés. *Analyse numérique des inéquations variationnelles*. Dunod, Paris, 1976.

[58] C. Grossmann, H.-G. Roos, and M. Stynes. *Numerical treatment of partial differential equations*. Universitext. Springer, Berlin, 2007. Translated and revised from the 3rd (2005) German edition by Martin Stynes.

[59] G. Haase, B. Heise, M. Kuhn, and U. Langer. Adaptive domain decomposition methods for finite and boundary element equations. In *Boundary element topics (Stuttgart, 1995)*, pages 121–147. Springer, Berlin, 1997.

[60] W. Hackbusch. Parabolic multigrid methods. In *Computing methods in applied sciences and engineering, VI (Versailles, 1983)*, pages 189–197. North-Holland, Amsterdam, 1984.

[61] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving ordinary differential equations. I*, volume 8 of *Springer series in computational mathematics*. Springer-Verlag, Berlin, second edition, 1993. Nonstiff problems.

[62] B.-O. Heimsund, X.-C. Tai, and J. Wang. Superconvergence for the gradient of finite element approximations by L^2 projections. *SIAM J. Numer. Anal.*, 40(4):1263–1280, 2002.

[63] G. Horton and S. Vandewalle. A space-time multigrid method for parabolic partial differential equations. *SIAM J. Sci. Comput.*, 16(4):848–864, 1995.

[64] Y. Hoshikawa and H. Urakawa. Affine Weyl groups and the boundary value eigenvalue problems of the Laplacian. *Interdiscip. Inform. Sci.*, 16(1):93–109, 2010.

[65] P. Houston, R. Rannacher, and E. Süli. A posteriori error analysis for stabilised finite element approximations of transport problems. *Comput. Methods Appl. Mech. Engrg.*, 190(11–12):1483–1508, 2000.

[66] A. Hrennikoff. Solution of problems of elasticity by the framework method. *J. Appl. Mech.*, 8:A–169–A–175, 1941.

[67] V. I. Istratescu. *Fixed Point Theory, An Introduction*. The Netherlands, 1981.

[68] C. Johnson. Error estimates and adaptive time-step control for a class of one-step methods for stiff ordinary differential equations. *SIAM J. Numer. Anal.*, 25(4):908–926, 1988.

[69] C. Johnson. *Numerical solution of partial differential equations by the finite element method*. Dover Publications Inc., Mineola, NY, 2009. Reprint of the 1987 edition.

[70] C. Johnson and P. Hansbo. Adaptive finite elements in computational mechanics. *Comput. Methods Appl. Mech. Engrg.*, 101(1-2):143–181, 1992.

[71] F. Kikuchi and X. Liu. Estimation of interpolation error constants for the P_0 and P_1 triangular finite elements. *Comput. Methods Appl. Mech. Engrg.*, 196(37-40):3750–3758, 2007.

[72] A. Klawonn, O. Rheinbach, and O. B. Widlund. An analysis of a FETI-DP algorithm on irregular subdomains in the plane. *SIAM J. Numer. Anal.*, 46(5):2484–2504, 2008.

[73] M. Kollmann, M. Kolmbauer, U. Langer, M. Wolfmayr, and W. Zulehner. A robust finite element solver for a multiharmonic parabolic optimal control problem. *Comput. Math. Appl.*, 65(3):469–486, 2013.

[74] A. N. Kolmogorov and S. V. Fomin. *Introductory real analysis*. Dover Publications, Inc., New York, 1975.

[75] M. Křížek and P. Neittaanmäki. Superconvergence phenomenon in the finite element method arising from averaging gradients. *Numer. Math.*, 45(1):105–116, 1984.

[76] M. Křížek and P. Neittaanmäki. On superconvergence techniques. *Acta Appl. Math.*, 9(3):175–198, 1987.

[77] M. Křížek, P. Neittaanmäki, and R. Stenberg. Superconvergence, post-processing and a posteriori error estimates. In M. Křížek, P. Neittaanmäki, and R. Stenberg, editors, *Lecture notes in pure and applied mathematics*, volume 196. Marcel Dekker, New York, 1998.

[78] P. Ladevéze and D. Leguillon. Error estimate procedure in the finite element method and applications. *SIAM J. Numer. Anal.*, 20(3):485–509, 1983.

[79] O. A. Ladyzhenskaya. *The boundary value problems of mathematical physics*. Springer, New York, 1985.

[80] O. A. Ladyzhenskaya, V. A. Solonnikov, and N.N. Uraltseva. *Linear and quasilinear equations of parabolic type*. Nauka, Moscow, 1967.

[81] J. Lang. *Adaptive multilevel solution of nonlinear parabolic PDE systems*, volume 16 of *Lecture Notes in Computational Science and Engineering*. Springer-Verlag, Berlin, 2001. Theory, algorithm, and applications.

[82] U. Langer. Parallel iterative solution of symmetric coupled FE/BE-equations via domain decomposition. In *Domain decomposition methods in science and engineering (Como, 1992)*, volume 157 of *Contemp. Math.*, pages 335–344. Amer. Math. Soc., Providence, RI, 1994.

[83] U. Langer, S. Repin, and M. Wolfmayr. Functional a posteriori error estimates for parabolic time-periodic boundary value problems. *CMAM*, 15(3):353–372, 2015.

[84] U. Langer and M. Wolfmayr. Multiharmonic finite element analysis of a time-periodic parabolic optimal control problem. *J. Numer. Math.*, 21(4):265–300, 2013.

[85] R. S. Laugesen and B. A. Siudeja. Maximizing Neumann fundamental tones of triangles. *J. Math. Phys.*, 50(11):112903, 18, 2009.

[86] R. S. Laugesen and B. A. Siudeja. Minimizing Neumann fundamental tones of triangles: an optimal Poincaré inequality. *J. Differential Equations*, 249(1):118–135, 2010.

[87] A. Lax. Decaying shocks. A comparison of an approximate analytic solution with a finite difference method. *Communications on Appl. Math.*, 1:247–257, 1948.

[88] E. Lindelöf. Sur l'application de la méthode des approximations successives aux équations différentielles ordinaires du premier ordre. In *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, volume 114, pages 454–457. Juillet, 1894.

[89] J. Liouville. Sur la théorie de la variation des constantes arbitraires. *Liouville J. de Math.*, 3:342–349, 1838.

[90] X. Liu and S. Oishi. Guaranteed high-precision estimation for P_0 interpolation constants on triangular finite elements. *Jpn. J. Ind. Appl. Math.*, 30(3):635–652, 2013.

[91] A. Logg, K.-A. Mardal, and G. N. Wells, editors. *Automated solution of differential equations by the finite element method*, volume 84 of *Lecture Notes in Computational Science and Engineering*. Springer, Heidelberg, 2012. The FEniCS book.

[92] C. Makridakis and R. H. Nochetto. A posteriori error analysis for higher order dissipative methods for evolution problems. *Numer. Math.*, 104(4):489–514, 2006.

[93] O. Mali, P. Neittaanmäki, and S. Repin. *Accuracy verification methods*, volume 32 of *Computational Methods in Applied Sciences*. Springer, Dordrecht, 2014. Theory and algorithms.

[94] MathWorks MATLAB and Simulink for Technical Computing. Mathworks. products and services, 2015.

[95] D. Meidner, R. Rannacher, and B. Vexler. A priori error estimates for finite element discretizations of parabolic optimization problems with pointwise state constraints in time. *SIAM J. Control Optim.*, 49(5):1961–1997, 2011.

[96] D. Meidner, R. Rannacher, and J. Vihharev. Goal-oriented error control of the iterative solution of finite element equations. *J. Numer. Math.*, 17(2):143–172, 2009.

[97] S. G. Mikhlin. *Constants in some inequalities of analysis*. A Wiley-Interscience Publication. John Wiley and Sons, Ltd., Chichester, 1986. Translated from the Russian by Reinhard Lehmann.

[98] K. W. Morton and D. F. Mayers. *Numerical solution of partial differential equations*. Cambridge University Press, Cambridge, 1994. An introduction.

[99] A. I. Nazarov and S. I. Repin. Exact constants in Poincare type inequalities for functions with zero mean boundary traces. *Mathematical Methods in the Applied Sciences*, 2014. Puplished in arXiv.org in 2012, math/1211.2224.

[100] P. Neittaanmäki and S. Repin. *Reliable methods for computer simulation*, volume 33 of *Studies in Mathematics and its Applications*. Elsevier Science B.V., Amsterdam, 2004. Error control and a posteriori estimates.

[101] O. Nevanlinna. Remarks on Picard–Lindelöf iteration Part I. *Springer, BIT Numerical Mathematics*, 29(2):328–346, 1989.

[102] O. Nevanlinna. Remarks on Picard–Lindelöf iteration Part II. *Springer, BIT Numerical Mathematics*, 29(3):535–562, 1989.

[103] J. T. Oden and S. Prudhomme. Goal–oriented error estimation and adaptivity for the finite element method. *Comput. Methods Appl.*, 41(5–6):735–756, 2001.

[104] L. A. Oganesjan and L. A. Ruhovec. An investigation of the rate of convergence of variation-difference schemes for second order elliptic equations in a two-dimensional region with smooth boundary. *Ž. Vyčisl. Mat. i Mat. Fiz.*, 9:1102–1120, 1969.

[105] D. Pauly. On Maxwell's and Poincaré's constants. *Discrete Contin. Dyn. Syst. Ser. S*, 8(3):607–618, 2015.

[106] L. E. Payne and H. F. Weinberger. An optimal Poincaré inequality for convex domains. *Arch. Rational Mech. Anal.*, 5:286–292 (1960), 1960.

[107] G. Peano. Sull'integrabilitÃ delle equazioni differenziali del primo ordine. *Atti Accad. Sci. Torino*, 21:437–445, 1886.

[108] G. Peano. Intégration par séries des équations différentielles linéaires. *Math. Annalen*, 32:450–456, 1888.

[109] K. Pearson. The problem of the random walk. *Nature*, 72:294, 318, 342, 1905.

[110] J. Peraire and A. T. Patera. Bounds for linear-functional outputs of coercive partial differential equations: Local indicators and adaptive refinement. In P. Ladevéze and J. T. Oden, editors, *Advances in adaptive computational methods in mechanics*, pages 199–228. Elsevier, New York, 1998.

[111] E. Picard. Mémoire sur la théorie des équations aux dérivées partielles et la méthode des approximations successives. *J. de Math. pures et appl., 4e série*, 6:145–210, 1890.

[112] E. Picard. *Traité d'Analyse. III Volums*. Paris, 1891–1896.

[113] M. A. Pinsky. The eigenvalues of an equilateral triangle. *SIAM J. Math. Anal.*, 11(5):819–827, 1980.

[114] H. Poincare. Sur les Equations aux Derivees Partielles de la Physique Mathematique. *Amer. J. Math.*, 12(3):211–294, 1890.

[115] H. Poincare. Sur les Equations de la Physique Mathematique. *Rend. Circ. Mat. Palermo*, 8:57–156, 1894.

[116] R. Rannacher. The dual-weighted-residual method for error control and mesh adaptation in finite element methods. In *The mathematics of finite elements and applications, X, MAFELAP 1999 (Uxbridge)*, pages 97–116. Elsevier, Oxford, 2000.

[117] R. Rannacher and B. Vexler. Adaptive finite element discretization in PDE-based optimization. *GAMM-Mitt.*, 33(2):177–193, 2010.

[118] S. Repin. A posteriori error estimation for variational problems with power growth functionals based on duality theory. *Zapiski Nauchnych Seminarov POMI*, 249:244–255, 1997.

[119] S. Repin. A posteriori estimates for approximate solutions of variational problems with strongly convex functionals. *Problems of Mathematical Analysis*, 17:199–226, 1997.

[120] S. Repin. A posteriori error estimation for variational problems with uniformly convex functionals. *Math. Comput.*, 69(230):481–500, 2000.

[121] S. Repin. *A posteriori estimates for partial differential equations*, volume 4 of *Radon Series on Computational and Applied Mathematics*. Walter de Gruyter GmbH & Co. KG, Berlin, 2008.

[122] S. Repin and S. Sauter. Functional a posteriori estimates for the reaction-diffusion problem. *C. R. Acad. Sci. Paris*, 343(1):349–354, 2006.

[123] S. I. Repin. A unified approach to a posteriori error estimation based on duality error majorants. *Math. Comput. Simulation*, 50(1-4):305–321, 1999. Modelling '98 (Prague).

[124] S. I. Repin. Estimates of deviations from exact solutions of initial-boundary value problem for the heat equation. *Rend. Mat. Acc. Lincei*, 13(9):121–133, 2002.

[125] S. I. Repin and S. K. Tomar. A posteriori error estimates for approximations of evolutionary convection-diffusion problems. *J. Math. Sci. (N. Y.)*, 170(4):554–566, 2010. Problems in mathematical analysis. No. 50.

[126] T. Richter, A. Springer, and B. Vexler. Efficient numerical realization of discontinuous Galerkin methods for temporal discretization of parabolic problems. *Numer. Math.*, 124(1):151–182, 2013.

[127] W. T. Rouleau and J. F. Osterle. The application of finite difference methods to boundary-layer type flows. *J. Aero. Sci.*, 22:249–254, 1955.

[128] M. Schmich and B. Vexler. Adaptivity with dynamic meshes for space-time finite element discretizations of parabolic equations. *SIAM J. Sci. Comput.*, 30(1):369–393, 2007/08.

[129] D. Schötzau and C. Schwab. Time discretization of parabolic problems by the hp-version of the discontinuous Galerkin finite element method. *SIAM J. Numer. Anal.*, 38(3):837–875, 2000.

[130] D. Schötzau and T. P. Wihler. A posteriori error estimation for hp-version time-stepping methods for parabolic partial differential equations. *Numer. Math.*, 115(3):475–509, 2010.

[131] E. Stein and S. Ohnimus. Coupled model- and solution-adaptivity in the finite element method. *Comput. Methods Appl. Mech. Engrg.*, 150(1–4):327–350, 1997.

[132] E. Stein, M. Rüter, and S. Ohnimus. Error-controlled adaptive goal-oriented modeling and finite element approximations in elasticity. *Comput. Methods Appl. Mech. Engrg.*, 196(37–40):3598–3613, 2007.

[133] G. Strang and G. Fix. *An analysis of the finite element method*. Prentice Hall, Englewood Cliffs, 1973.

[134] A. H. Stroud. *Numerical quadrature and solution of ordinary differential equations*. Springer-Verlag, New York-Heidelberg, 1974. A textbook for a beginning course in numerical analysis, Applied Mathematical Sciences, Vol. 10.

[135] A. H. Stroud and Don Secrest. *Gaussian quadrature formulas*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1966.

[136] G. Teschl. *Ordinary differential equations and dynamical systems*, volume 140 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2012.

[137] The FEniCS Project ©Copyright 2015. The fenics project, 2015.

[138] V. Thomée. *Galerkin finite element methods for parabolic problems*, volume 25 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 2006.

[139] A. Toselli and O. Widlund. *Domain decomposition methods—algorithms and theory*, volume 34 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2005.

[140] S. Vandewalle and R. Piessens. Efficient parallel algorithms for solving initial-boundary value and time-periodic parabolic partial differential equations. *SIAM J. Sci. Statist. Comput.*, 13(6):1330–1346, 1992.

[141] J. L. Vázquez. *The porous medium equation*. Oxford Mathematical Monographs. The Clarendon Press, Oxford University Press, Oxford, 2007. Mathematical theory.

[142] R. Verfürth. *A review of a posteriori error estimation and adaptive mesh-refinement techniques*. Wiley and Sons, Teubner, New-York, 1996.

[143] R. Verfürth. A posteriori error estimates for finite element discretizations of the heat equation. *Calcolo*, 40(3):195–212, 2003.

[144] L. B. Wahlbin. *Superconvergence in Galerkin finite element methods*, volume 1605 of *Lecture notes in mathematics*. Springer-Verlag, Berlin, 1995.

[145] J. Wang. Superconvergence analysis of finite element solutions by the least-squares surface fitting on irregular meshes for smooth problems. *J. Math. Study*, 33:229–243, 2000.

[146] J. Wang and X. Ye. Superconvergence analysis for the Navier–Stokes equations. *Applied Numerical Mathematics*, 41:515–527, 2002.

[147] D. V. Widder. *The heat equation*. Academic Press [Harcourt Brace Jovanovich, Publishers], New York-London, 1975. Pure and Applied Mathematics, Vol. 67.

[148] J. Wloka. *Partial Differential Equations*. Cambridge University Press, 1987.

[149] D. E. Womble. A time-stepping algorithm for parallel computers. *SIAM J. Sci. Statist. Comput.*, 11(5):824–837, 1990.

[150] E. Zeidler. *Nonlinear functional analysis and its applications. I: Fixed-point theorems*. Springer-Verlag, New York, 1986.

[151] E. Zeidler. *Nonlinear functional analysis and its applications. II/A*. Springer-Verlag, New York, 1990. Linear monotone operators, Translated from the German by the author and Leo F. Boron.

[152] E. Zeidler. *Nonlinear functional analysis and its applications. II/B*. Springer-Verlag, New York, 1990. Nonlinear monotone operators, Translated from the German by the author and Leo F. Boron.

[153] Zh. Zhang and A. Naga. A new finite element gradient recovery method: superconvergence property. *SIAM J. Sci. Comput.*, 26(4):1192–1213, 2005.

[154] O. C. Zienkiewicz, B. Boroomand, and J. Z. Zhu. Recovery procedures in error estimation and adaptivity: adaptivity in linear problems. In P. Ladeveze and J.T. Oden, editors, *Advances in adaptive computational methods in mechanics (Cachan, 1997)*, volume 47 of *Stud. Appl. Mech.*, pages 3–23. Elsevier, Amsterdam, 1998.

[155] O. C. Zienkiewicz and J. Z. Zhu. A simple error estimator and adaptive procedure for practical engineering analysis. *Internat. J. Numer. Meth. Engrg.*, 24(2):337–357, 1987.

[156] O. C. Zienkiewicz and J. Z. Zhu. Adaptive techniques in the finite element method. *Commun. Appl. Numer. Methods*, 4:197–204, 1988.

[157] M. Zlámal. On the finite element method. *Numer. Math.*, 12:394–409, 1968.

[158] M. Zlámal. Some superconvergence results in the finite element method. In *Mathematical aspects of finite element methods (Proc. Conf., Consiglio Naz. delle Ricerche (C.N.R.), Rome, 1975)*, pages 353–362. Lecture Notes in Math., Vol. 606. Springer, Berlin, 1977.

ORIGINAL PAPERS

PI

GUARANTEED ERROR BOUNDS FOR A CLASS OF PICARD-LINDELÖF ITERATION METHODS

by

S. Matculevich, P. Neittaanmäki, and S. Repin (2013)

Numerical methods for differential equations, optimization, and technological problems, Comput. Methods Appl. Sci., **27**: 175–189

Guaranteed Error Bounds for a Class of Picard-Lindelöf Iteration Methods

Svetlana Matculevich, Pekka Neittaanmäki, and Sergey Repin

Abstract We present a new version of the Picard-Lindelöf method for ordinary differential equations (ODEs) supplied with guaranteed and explicitly computable upper bounds of an approximation error. The upper bounds are based on the Ostrowski estimates and the Banach fixed point theorem for contractive operators. The estimates derived in the paper take into account interpolation and integration errors and, therefore, provide objective information on the accuracy of computed approximations.

1 Introduction

In this paper, we discuss a new version of the Picard-Lindelöf method for solving the Cauchy problem

$$\frac{du}{dt} = \varphi(u(t),t), \quad u(t_0) = u_0, \tag{1}$$

where the solution $u(t)$ (which may be a scalar or vector function) must be found on the interval $[t_0, t_K]$.

Existence and uniqueness of the solutions follow from the Picard-Lindelöf theorem and the Picard existence theorem or from the Cauchy–Lipschitz theorem (see [1, pp. 1–15], [3]).

Svetlana Matculevich, Pekka Neittaanmäki
Department of Mathematical Information Technology, University of Jyväskylä, P.O. Box 35 (Agora), FI-40014 University of Jyväskylä, Finland, e-mail: svmatsul@student.jyu.fi, pekka.neittaanmaki@mit.jyu.fi

Sergey Repin
V. A. Steklov Institute of Mathematics in St. Petersburg, Fontanka 27, RU-191024, St. Petersburg, Russia, e-mail: repin@pdmi.ras.ru, and Department of Mathematical Information Technology, University of Jyväskylä, PO Box 35 (Agora), FI-40014 University of Jyväskylä, Finland

The problem (1) can be numerically solved by various well-known methods (e.g., the methods of Runge–Kutta and Adams). Typically, the methods are furnished by a priori asymptotic estimates which show theoretical properties of the iteration algorithm. However, these estimates may have mainly a qualitative meaning and do not provide all necessary information about error bounds. This is the goal of a posteriori error estimation methods. We deduce such type of estimates and suggest a version of the Picard-Lindelöf method as a tool for constructing a fully reliable approximation of (1).

The Picard-Lindelöf iteration is one of the efficient known numerical methods for ODEs. Furthermore, it can be used not only for ODEs but for t-dependent algebraic and functional equations (see, e.g., [5, 6]). It was shown that the speed of convergence is quite independent of the step sizes. Numerical methods based on Picard-Lindelöf iterations for dynamical processes (the so-called waveform relaxation in the context of electrical networks) are discussed in [2].

The approach discussed in this paper is based on two-sided a posteriori estimates derived by Ostrowski [7] (see also systematic exposition presented in the books [4, 8]). The algorithm includes natural adaptation of the integration step and provides guaranteed bounds for the accuracy on the time interval $[t_0, t_K]$.

In Sect. 2, we present the main idea of the Picard-Lindelöf method and obtain the conditions which not only provide convergence of the method but also allow applying a posteriori error estimates. However, these estimates cannot be directly used. In practice computations based on the Picard-Lindelöf method we must take into account interpolation and integration errors. This analysis is done in Sect. 3. It leads to error bounds, derived in Sect. 4, which include the interpolation and integration errors. The structure of the algorithm is exposed in Sect. 5, where results of numerical tests are presented.

2 Picard-Lindelöf Method

Assume that the function $\varphi(\xi(t), t)$ (which is allowed to be a vector-valued function) in (1) is continuous with respect to both variables in terms of the continuous norm

$$\|u\|_{C([t_k,t_{k+1}])} := \max_{t\in[t_k,t_{k+1}]} |u(t)| \tag{2}$$

and satisfies the Lipschitz condition in the form

$$\|\varphi(u_2,t_2)-\varphi(u_1,t_1)\|_{C([t_1,t_2])} \le \mathrm{L}_1\|u_2-u_1\|_{C([t_1,t_2])} + \mathrm{L}_2|t_2-t_1|, \quad \forall (u_1,t_1),(u_2,t_2)\in Q, \tag{3}$$

where L_1, L_2 are Lipschitz constants, and

$$Q := \{(\xi,t) \mid \xi \in \mathrm{U},\ t_0 \le t \le t_N\}. \tag{4}$$

U is the set of possible values of u which comes from an a priori analysis of the problem. (It is clear that $u_0 \in \mathrm{U}$.)

In the Picard-Lindelöf method, we represent the differential equation in the integral form

$$u(t) = \int_{t_0}^{t} \varphi(u(s), s)\, ds + u_0. \tag{5}$$

Now, the exact solution is a fixed point of (5), which can be found by the iteration method

$$u_j(t) = \int_{t_0}^{t} \varphi(u_{j-1}(s), s)\, ds + u_0. \tag{6}$$

We write in the form $u_j = \mathscr{T} u_{j-1} + u_0$, where $\mathscr{T} : X \to X$ is the integral operator.

It is easy to show that the operator

$$\mathscr{T} u := \int_{t_k}^{t} \varphi(u(\tau), \tau) d\tau + u_{0,k}$$

is q-contractive on $I_k = [t_k, t_{k+1}]$, where I_k is a subinterval of the mesh $\mathscr{T}_K = \cup_{k=0}^{K-1} [t_k, t_{k+1}]$ defined on the interval $[t_0, t_K]$, with respect to the norm $\|u\|_{C(I_k)}$, if the condition

$$q := \mathrm{L}_1 (t_{k+1} - t_k) < 1 \tag{7}$$

is provided.

Therefore, if the interval $[t_{k+1}, t_k]$ is small enough, then the solution can be found by the iteration procedure. In the next sections, we call this method the Adaptive Picard-Lindelöf (APL) method.

3 Application of the Ostrowski Estimates

For the considered problem, the Ostrowski estimate reads as follows:

Theorem 1 ([7]). *Assume that* (7) *is satisfied on* $I_k := [t_k, t_{k+1}]$. *Then, the following estimate holds:*

$$M_j^{\ominus} := \frac{1}{1+q} \|u_j - u_{j+1}\|_{C(I_k)} \le \|u - u_j\|_{C(I_k)} \le \frac{q}{1-q} \|u_j - u_{j-1}\|_{C(I_k)} =: M_j^{\oplus}. \tag{8}$$

Remark 1. It is possible to derive more accurate error bounds for $\|u - u_j\|_{C(I_k)}$ by using additional elements of the sequence $\{u_j\}_{j=1}^{\infty}$ that have indexes greater than j:

$$\|u - u_j\|_{C(I_k)} \le M_j^{\oplus, p} := \frac{1}{1-q^p} \|u_j - u_{j+p}\|_{C(I_k)}. \tag{9}$$

By the mathematical induction method it can be proved that the optimal form of the majorant and minorant based on P correspondent elements of the sequence are as follows:

$$M_j^{\ominus,P} := \sup_{p=1,\dots,P} \left\{ \frac{1}{1+q^p} \|u_j - u_{j+p}\|_{C(I_k)} \right\},$$
$$M_j^{\oplus,P} := \inf_{p=1,\dots,P} \left\{ \frac{1}{1-q^p} \|u_j - u_{j+p}\|_{C(I_k)} \right\}. \tag{10}$$

However, the estimates (8) cannot be directly used because numerical approximations include interpolation and integration errors, which must be taken into account by fully reliable schemes.

Let us discuss this issue within the paradigm of a single (e.g., the first) step of the APL:

$$u_1(t) = \int_{t_0}^{t} \varphi(u_0(\tau), \tau) d\tau, \; t \in [t_0, t_1], \tag{11}$$

where u_0 is the initial approximation defined as a piecewise affine function on the mesh $\Omega_{S_k} = \cup_{s=0}^{S_k-1} [z_s, z_{s+1}]$ on the interval $[t_0, t_1]$.

If $q < 1$ and u_1 is computed exactly, then

$$\|u_1(t) - u(t)\|_{C([t_0,t_1])} \le \frac{q}{1-q} \|u_1(t) - u_0(t)\|_{C([t_0,t_1])}. \tag{12}$$

However, in general, u_1 is approximated by a piecewise affine continuous function

$$\overline{u}_1(t) = \pi u_1 \in CP^1([z_s, z_{s+1}]), \quad s = 0, \dots, S_k - 1, \tag{13}$$

where π is the projection operator $\pi : C \to CP^1([t_0, t_1])$ satisfying the relation $\pi\, u(z_s) = \overline{u}(z_s)$. Thus, on the right-hand side of (12) we can estimate as follows:

$$\|u_1(t) - u_0(t)\|_{C([t_0,t_1])} \le \|\overline{u}_1(t) - u_0(t)\|_{C([t_0,t_1])} + \|\overline{u}_1(t) - u_1(t)\|_{C([t_0,t_1])}. \tag{14}$$

Here $\|\overline{u}_1(t) - u_1(t)\|_{C([t_0,t_1])} = \|\overline{e}_1\|_{C([t_0,t_1])}$ is an interpolation error. In general, this term is unknown, but we can estimate it using an interpolation error estimate.

Numerical integration generates other errors which must be taken into account. Indeed, the values $\overline{u}(z_s)$, $s = 0, \dots, S_k$, cannot be found exactly. Hence, at every node z_s instead of $\overline{u}_1(z_s)$ we have $\widehat{u}_1(z_s)$. Now, (14) implies

$$\|u_1(t) - u_0(t)\|_{C([t_0,t_1])} \le \|\widehat{u}_1(t) - u_0(t)\|_{C([t_0,t_1])} + \\ + \|\widehat{u}_1(t) - \overline{u}_1(t)\|_{C([t_0,t_1])} + \|\overline{u}_1(t) - u_1(t)\|_{C([t_0,t_1])}, \tag{15}$$

where $\|\widehat{u}_1(t) - \overline{u}_1(t)\|_{C([t_0,t_1])} = \|\widehat{e}_1\|_{C([t_0,t_1])}$ is the integration error.

4 Estimates of Interpolation and Integration Errors

4.1 Interpolation Error

We study the difference between u_1 and $\overline{u}_1$, where $\overline{u}_1$ is the linear interpolant of u_1 defined at the points $\{z_s\}_{s=0}^{S_k}$:

$$u_1(z_s) = \overline{u}_1(z_s) = \int_0^{z_s} \varphi(u_0(t),t)\,dt. \tag{16}$$

For all $z \in [z_s, z_{s+1}]$,

$$\overline{u}_1(z) = u_1(z_s) + \frac{u_1(z_{s+1}) - u_1(z_s)}{\Delta_s}(z - z_s). \tag{17}$$

Then,

$$\begin{aligned}\overline{e} = \overline{u}_1(z) - u_1(z) = \\ = \left[\int_0^{z_s} \varphi(u_0(t),t)\,dt + \frac{\int_{z_s}^{z_{s+1}} \varphi(u_0(t),t)dt}{\Delta_s}(z - z_s)\right] - \int_0^{z} \varphi(u_0(t),t)\,dt = \\ = \frac{z - z_s}{\Delta_s}\int_{z_s}^{z_{s+1}} \varphi(u_0(t),t)dt - \int_{z_s}^{z} \varphi(u_0(t),t)\,dt. \end{aligned} \tag{18}$$

Taking into account that u_0 is affinely interpolated, consider the last integral on the right-hand side of (18)

$$\int_{z_s}^{z} \varphi(u_0(t),t)\,dt = \int_{z_s}^{z} \varphi\left(u_{0,s} + \frac{u_{0,s+1} - u_{0,s}}{\Delta_s}(t - z_s), t\right) dt. \tag{19}$$

Define

$$\lambda = \frac{t - z_s}{\Delta_s} = \frac{t - z_s}{z_{s+1} - z_s}, \tag{20}$$

where z_s and z_{s+1} are nodes of the mesh defined in Section 3. Substitute $t = z_s + (z_{s+1} - z_s)\lambda$ to $\varphi(u_0(t),t)$

$$\begin{aligned}\varphi\left(u_{0,s} + \frac{u_{0,s+1} - u_{0,s}}{\Delta_s}(t - z_s), t\right) = \\ = \varphi(u_{0,s} + (u_{0,s+1} - u_{0,s})\lambda, z_s + \lambda(z_{s+1} - z_s)) = \\ = \varphi(\lambda u_{0,s+1} + (1-\lambda)u_{0,s}, \lambda z_{s+1} + (1-\lambda)z_s). \end{aligned} \tag{21}$$

Let

$$\widetilde{\varphi}_{[s,s+1]} := \varphi_s + \frac{\varphi_{s+1} - \varphi_s}{\Delta_s}(t - z_s), \tag{22}$$

where $\varphi_s = \varphi(u_{0,s}, z_s)$ and $\varphi_{s+1} = \varphi(u_{0,s+1}, z_{s+1})$. Using (20), we rewrite (22)

$$\widetilde{\varphi}_{[s,s+1]} = \varphi_s + (\varphi_{s+1} - \varphi_s)\lambda = \lambda\varphi_{s+1} + (1-\lambda)\varphi_s. \tag{23}$$

Thus, we can derive the following estimate with the help of (23) and (3):

$$\begin{aligned}
&\left|\varphi\left(u_{0,s} + \frac{u_{0,s+1}-u_{0,s}}{\Delta_s}(t-z_s), t\right) - \widetilde{\varphi}_{[s,s+1]}\right| \leq \\
&\leq |\varphi(\lambda u_{0,s+1} + (1-\lambda)u_{0,s},\ \lambda z_{s+1} + (1-\lambda)z_s) - \lambda\varphi_{s+1} + (1-\lambda)\varphi_s| \leq \\
&\leq (1-\lambda)\Big[\mathrm{L}_{1,s}|\lambda u_{0,s+1} + (1-\lambda)u_{0,s} - u_{0,s}| + \\
&\qquad\qquad + \mathrm{L}_{2,s}|\lambda z_{s+1} + (1-\lambda)z_s - z_s|\Big] + \\
&\quad + \ \lambda\Big[\mathrm{L}_{1,s}|\lambda u_{0,s+1} + (1-\lambda)u_{0,s} - u_{0,s+1}| + \\
&\qquad\qquad + \mathrm{L}_{2,s}|\lambda z_{s+1} + (1-\lambda)z_s - z_{s+1}|\Big] \leq \\
&\leq 2\lambda(1-\lambda)\left[\mathrm{L}_{1,s}|u_{0,s+1} - u_{0,s}| + \mathrm{L}_{2,s}|z_{s+1} - z_s|\right] \\
&\leq 2\frac{(z_{s+1}-t)(t-z_s)}{\Delta_s^2}\left[\mathrm{L}_{1,s}|u_{0,s+1} - u_{0,s}| + \mathrm{L}_{2,s}\Delta_s\right].
\end{aligned} \tag{24}$$

We decompose (19)

$$\begin{aligned}
&\int_{z_s}^{z} \varphi(u_0(t), t)dt = \\
&\quad = \int_{z_s}^{z} \widetilde{\varphi}_{[s,s+1]}(t)dt + \int_{z_s}^{z}\left[\varphi(u_{0,s} + \frac{u_{0,s+1}-u_{0,s}}{\Delta_s}(t-z_s), t) - \widetilde{\varphi}_{[s,s+1]}\right]dt.
\end{aligned} \tag{25}$$

Let us denote the first integral on the right-hand side of (25) by $\widetilde{i}_s(z)$. Then,

$$\widetilde{i}_s(z) := \int_{z_s}^{z}\left(\varphi_s + \frac{\varphi_{s+1}-\varphi_s}{\Delta_s}(t-z_s)\right)dt = (z-z_s)\left[\varphi_s + \frac{\varphi_{s+1}-\varphi_s}{2\Delta_s}(z-z_s)\right]. \tag{26}$$

The second integral on the right-hand side of (25) is estimated with the help of (24):

$$\begin{aligned}
&\int_{z_s}^{z}\left|\varphi\left(u_{0,s} + \frac{u_{0,s+1}-u_{0,s}}{\Delta_s}(t-z_s), t\right) - \widetilde{\varphi}_{[s,s+1]}\right|dt \leq \\
&\quad \leq \frac{2\left[\mathrm{L}_{1,s}|u_{0,s+1} - u_{0,s}| + \mathrm{L}_{2,s}\Delta_s\right]}{\Delta_s^2}\int_{z_s}^{z}(t-z_s)(z_{s+1}-t)dt = \\
&\quad = \frac{2\left[\mathrm{L}_{1,s}|u_{0,s+1} - u_{0,s}| + \mathrm{L}_{2,s}\Delta_s\right]}{\Delta_s^2}\int_{z_s}^{z}(t-z_s)(z_s+\Delta_s-t)dt = \\
&\quad = \frac{2\left[\mathrm{L}_{1,s}|u_{0,s+1} - u_{0,s}| + \mathrm{L}_{2,s}\Delta_s\right]}{\Delta_s^2}(z-z_s)^2\left[\frac{\Delta_s}{2} - \frac{z-z_s}{3}\right] = \\
&\quad = \frac{\left[\mathrm{L}_{1,s}|u_{0,s+1} - u_{0,s}| + \mathrm{L}_{2,s}\Delta_s\right]}{3\Delta_s^2}(z-z_s)^2(2z_s + 3\Delta_s - 2z).
\end{aligned} \tag{27}$$

Since

$$\max_{z\in[z_s,z_{s+1}]}(z-z_s)^2(2z_s+3\Delta_s-2z)=\Delta_s^3, \tag{28}$$

we find that

$$\int_{z_s}^{z}\left|\varphi(u_{0,s}+\frac{u_{0,s+1}-u_{0,s}}{\Delta_s}(t-z_s),t)-\widetilde{\varphi}_{[s,s+1]}\right|dt\leq$$
$$\leq\frac{[\mathrm{L}_{1,s}|u_{0,s+1}-u_{0,s}|+\mathrm{L}_{2,s}\Delta_s]\,\Delta_s^3}{3\Delta_s^2}=$$
$$=\frac{[\mathrm{L}_{1,s}|u_{0,s+1}-u_{0,s}|+\mathrm{L}_{2,s}\Delta_s]\,\Delta_s}{3}. \tag{29}$$

We represent the interpolation error (18) using (26),

$$\overline{u}_1(z)-u_1(z)=\frac{z-z_s}{\Delta_s}\int_{z_s}^{z_{s+1}}\varphi(u_0(t),t)dt-\int_{z_s}^{z}\varphi(u_0(t),t)dt=$$
$$=\frac{z-z_s}{\Delta_s}\widetilde{i}_s(z_{s+1})-\widetilde{i}_s(z)+\varepsilon_1(z)+\varepsilon_2(z), \tag{30}$$

where

$$\begin{aligned}\varepsilon_1&=\int_{z_s}^{z_{s+1}}\left|\varphi(u_{0,s}+\frac{u_{0,s+1}-u_{0,s}}{\Delta_s}(t-z_s),t)-\widetilde{\varphi}_{[s,s+1]}\right|dt,\\ \varepsilon_2&=\int_{z_s}^{z}\left|\varphi(u_{0,s}+\frac{u_{0,s+1}-u_{0,s}}{\Delta_s}(t-z_s),t)-\widetilde{\varphi}_{[s,s+1]}\right|dt.\end{aligned} \tag{31}$$

Thus, we estimate the interpolation error as follows:

$$\overline{e}=\|\overline{u}_1(z)-u_1(z)\|_{C([z_s,z_{s+1}])}\leq$$
$$\leq\max_{z\in[z_s,z_{s+1}]}\left|\frac{z-z_s}{\Delta_s}\widetilde{i}_s(z_{s+1})-\widetilde{i}_s(z)\right|+\max_{z\in[z_s,z_{s+1}]}|\varepsilon_1(z)+\varepsilon_2(z)|. \tag{32}$$

For the first term on the right hand side of (32) we have (see (26))

$$\max_{z\in[z_s,z_{s+1}]}\left|\frac{z-z_s}{\Delta_s}\widetilde{i}_s(z_{s+1})-\widetilde{i}_s(z)\right|dt\leq\frac{|\varphi_{s+1}-\varphi_s|}{2\Delta_s}\max_{z\in[z_s,z_{s+1}]}|(z-z_s)(z_{s+1}-z)|\leq$$
$$\leq\frac{|\varphi_{s+1}-\varphi_s|}{2\Delta_s}\frac{\Delta_s^2}{4}=\frac{1}{8}|\varphi_{s+1}-\varphi_s|\Delta_s. \tag{33}$$

For the second term, we have (see (29))

$$\max_{z\in[z_s,z_{s+1}]}|\varepsilon_1(z)+\varepsilon_2(z)|\leq 2\frac{\Delta_s\left[\mathrm{L}_{1,s}|u_{0,s+1}-u_{0,s}|+\mathrm{L}_{2,s}\Delta_s\right]}{3}. \tag{34}$$

Hence, the overall estimate of the interpolation error has the form

$$\|\overline{u}_1(z) - u_1(z)\|_{C([z_s,z_{s+1}])} \leq \frac{\varphi_{s+1}-\varphi_s}{8}\Delta_s + \frac{2}{3}\Delta_s\left[\mathrm{L}_{1,s}|u_{0,s+1}-u_{0,s}| + \mathrm{L}_{2,s}\,\Delta_s\right]. \quad (35)$$

4.2 Integration Error

The interpolation error estimate (35) does not account for the fact that computations of the integral are performed approximately. It is not difficult to evaluate the integration errors by noting that for a Lipschitz function $f(t)$ the error encompassed in the simplest trapezoidal quadrature formula

$$\int_{t_0}^{t_1} f(t)dt \simeq \frac{f(t_0)+f(t_1)}{2}(t_1-t_0) \quad (36)$$

can be estimated as follows:

$$e_{int} \leq \frac{L}{4}(t_1-t_0)^2 - \frac{1}{4L}[f(t_1)-f(t_0)]^2. \quad (37)$$

Then, it is not difficult to show that the integration error can be estimated as

$$\|\widehat{u}_1(t) - \overline{u}_1(t)\|_{C([z_s,z_{s+1}])} \leq \frac{\mathrm{L}_s}{4}\Delta_s^2 - \frac{1}{4\mathrm{L}_s}[\varphi_{s+1}-\varphi_s]^2, \quad (38)$$

where $\mathrm{L}_s = \mathrm{L}_{1,s}\,l_s + \mathrm{L}_{2,s}$. (Here, l_s is the slope of the piecewise function on every interval $[z_s, z_{s+1}]$, $s = 0,...,S_k-1$.)

4.3 Guaranteed Error Bounds for Picard-Lindelöf Method

Thus, on every subinterval $[z_s, z_{s+1}]$ the interpolation error can be estimated with the help of (35). Then, for whole interval $[t_0,t_1] := \cup_{s=0}^{S_k-1}[z_s,z_{s+1}]$ the interpolation error estimate is the following:

$$\|\overline{u}_1(t) - u_1(t)\|_{C([t_0,t_1])} \leq$$
$$\leq \sum_{s=0,...,S_k-1} \frac{\varphi_{s+1}-\varphi_s}{8}\Delta_s + \frac{2}{3}\left[\mathrm{L}_{1,s}|u_{0,s+1}-u_{0,s}| + \mathrm{L}_{2,s}\Delta_s\right]\Delta_s. \quad (39)$$

Analogously, for the integration error

$$\|\overline{u}_1(t) - \widehat{u}_1(t)\|_{C([t_0,t_1])} \leq \sum_{s=0,...,S_k-1} \frac{\mathrm{L}_s}{2}\Delta_s^2 - \frac{1}{2\mathrm{L}_s}[\varphi_{s+1}-\varphi_s]^2. \quad (40)$$

Then, the inequality (15) implies the estimate

$$\|u_1(t)-u_0(t)\|_{C([t_0,t_1])} \le \|\widehat{u}_1(t)-u_0(t)\|_{C([t_0,t_1])}+ \\ + \sum_{s=0,\dots,S_k-1}\left(\frac{\varphi_{s+1}-\varphi_s}{8}\Delta_s+\frac{2}{3}\Delta_s\left[\mathrm{L}_{1,s}|u_{0,s+1}-u_{0,s}|+\mathrm{L}_{2,s}\Delta_s\right]\right)+ \\ + \sum_{s=0,\dots,S_k-1}\left(\frac{\mathrm{L}_s}{2}\Delta_s^2-\frac{1}{2\mathrm{L_s}}[\varphi_{s+1}-\varphi_s]^2\right). \quad (41)$$

After j steps of the iterations we obtain

$$\|u_{j+1}(t)-u_j(t)\|_{C([t_0,t_1])} \le \mathrm{M}_{j+1}^{\oplus,1}(\widehat{u}_j) := \\ \|\widehat{u}_{j+1}(t)-\widehat{u}_j(t)\|_{C([t_0,t_1])}+E^1_{interp}+E^1_{integr}, \quad (42)$$

where

$$E^1_{interp} := \sum_{s=0,\dots,S_k-1}\left(\frac{\varphi(\widehat{u}_{j,s+1},z_{s+1})-\varphi(\widehat{u}_{j,s},z_s)}{8}\Delta_s+ \right. \\ \left. +\frac{2}{3}\Delta_s\left[\mathrm{L}_{1,s}|\widehat{u}_{j,s+1}-\widehat{u}_{j,s}|+\mathrm{L}_{2,s}\Delta_s\right]\right) \quad (43)$$

and

$$E^1_{integr} := \sum_{s=0,\dots,S_k-1}\left(\frac{\mathrm{L_s}}{2}\Delta_s^2-\frac{1}{2\mathrm{L_s}}\left[\varphi(\widehat{u}_{j,s+1},z_{s+1})-\varphi(\widehat{u}_{j,s},z_s)\right]^2\right), \quad (44)$$

where for $j=0$ the function $\widehat{u}_j$ is taken as a piecewise affine interpolation of u_0, and for $j\ge 1$ it is taken from the previous iteration step.

The quantity $\mathrm{M}_j^{\oplus,1}$ is fully computable, and it shows the overall error associated with the step number j on the first interval.

Remark 2. Estimate of the overall error related to the interval $[t_0,t_K]$ includes all errors computed on the intervals. In other words the error associated with $[t_0,t_{k-1}]$ is appended to the error on $[t_{k-1},t_k]$ (which formally follows from the fact that the initial condition on $[t_{k-1},t_k]$ includes errors on the previous intervals).

Thus, we have shown that fully guaranteed and computable bounds can indeed be derived for the problem (1) with the Lipschitz function φ, i.e. for every finite time interval $[t_0,t_K]$ and for every a priori required accuracy ε an approximate solution of the problem can be found by the APL method discussed above.

5 APL Algorithm and Numerical Examples

Let ε be a required accuracy of an approximate solution. Then, practical computation can be performed by Algorithm 1.

Algorithm 1 The algorithm of the APL method

Input: ε {required accuracy on the interval}, u_0 {input initial boundary condition}

$\mathscr{T}_K = \bigcup_{k=0}^{K-1} [t_k, t_{k+1}]$ { constructed by *Mesh Generation Procedure*}

$\varepsilon^k = \frac{\varepsilon}{K}$ {obtain accuracy of the approximate solution on interval $[t_k, t_{k+1}]$}

$\Omega_{S_k} = \bigcup_{s=0}^{S_k-1} [z_s, z_{s+1}]$ {initial mesh for each subinterval}

for $k = 1$ to K **do**

  $j = 0$

  do

    if $k = 1$

      $a = u_0$

    else

      $a = v^{k-1}(t_{k-1})$

    endif

    $v_j^k = \mathit{Integration\ Procedure}(\varphi, v_{j-1}^k, S_k) + a$

    calculate E_{interp}^k and E_{integr}^k by using (43) and (44)

    $M_j^{\oplus,k} = \|v_j^k - v_{j-1}^k\|_{C([t_{k-1},t_k])} + E_{interp}^k + E_{integr}^k$

    $e_j^{\oplus} = \frac{q}{1-q} M_j^{\oplus,k}$

    if $E_{interp}^k + E_{integr}^k > \varepsilon_k$

      $S_k = 2\, S_k$ {refine the mesh Ω_{S_k}}

    endif

    $j = j + 1$

  while $e_j^{\oplus} > \varepsilon^k$

  $v^k = v_j^k$ {approximate solution on the interval $[t_{k-1}, t_k]$}

  $e^{\oplus,k} = e_j^{\oplus}$ {error bound achieved for the interval $[t_{k-1}, t_k]$}

end for

Output: $\{v^k\}_{k=1}^K$ {approximate solution}

$\{e^{\oplus,k}\}_{k=1}^K$ {error bounds estimates on sub intervals}

In general, the algorithm should start with the generation of a suitable mesh (i.e., select time intervals). Here, we do not discuss this question in detail, but only note that the *Mesh Guaranteed Procedure* must adapt the mesh to the nature of $\varphi(u(t), t)$, which requires information about U (see (4)). In practise, such information can be obtained by solving the problem (1) numerically with the help of some heuristic (e.g., Runge-Kutta) method on a coarse mesh.

The APL algorithm is a cycle over all the intervals of the mesh $\mathscr{T}_K = \cup_{k=0}^{K-1} [t_k, t_{k+1}]$. On each subinterval, the algorithm is realized as a subcycle (whose index is j). In the subcycle, we apply the PL method and try to find an approximation that meets the accuracy requirements imposed (i.e., the accuracy must be higher than ε^k). Initial data are taken from the previous step (for the first step, the initial condition is defined by u_0).

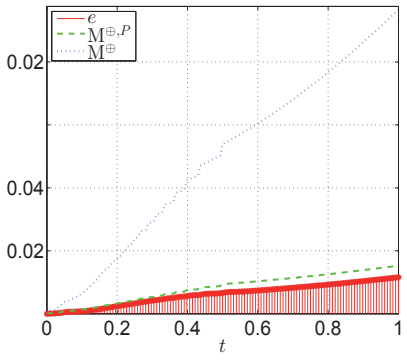


Fig. 1 The error and error majorants

After computing an approximation on $[t_k, t_{k+1}]$ we use our majorant and find a guaranteed upper bound (which includes the interpolation and integration errors). Iterations are continued unless the required accuracy ε^k has been achieved. After that we save the results and proceed to the next interval.

Note that in Algorithm 1, we do not discuss in detail the process of integration on an interval, which is performed on a local mesh with a certain amount of subintervals (whose size is Δ_s). In principle, it may happen that the desired level of accuracy, ε^k, is not achieved with the Δ_s selected. This fact will be easily detected because interpolation and integration errors will dominate and do not allow the overall error to decrease below ε^k. In this case, Δ_s must be reduced, and computations on the corresponding interval must be repeated.

Example 1. Consider the problem

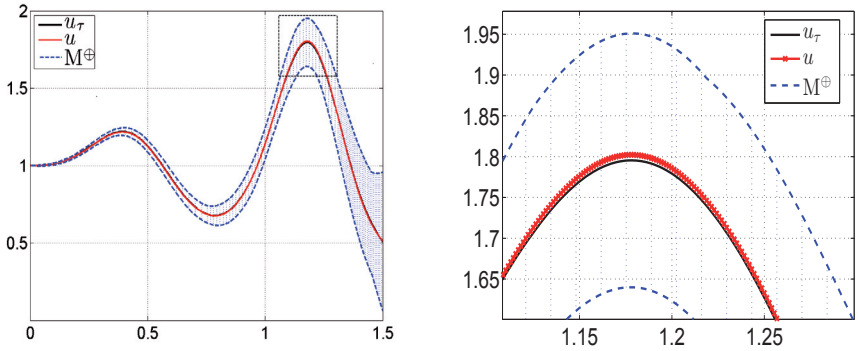
$$\begin{aligned} \frac{du}{dt} &= 4ut\sin(8t), \quad t \in [0, 3/2], \\ u(0) &= u_0 = 1 \end{aligned} \tag{45}$$

with the exact solution

$$u = e^{\frac{1}{16}\sin(8t) - \frac{1}{2}t\cos(8t)}.$$

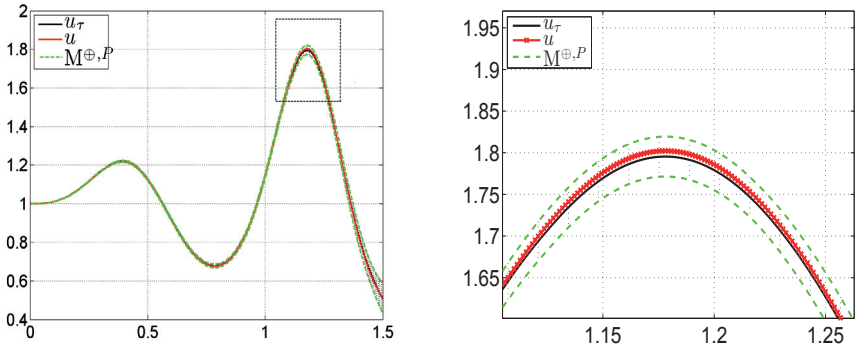
In Fig. 1, we depict the error (bold dots), error bounds computed by the Ostrowski estimates (dotted line) and by the advanced form of the estimate (dashed line). In order to make the results more transparent, we depict the approximate solution together with the zone which contains the exact solution (see Figs. 2(a) and 3(a)). The form of this (shaded) zone is determined by the a posteriori estimates.

Thus, the APL method computes two-sided guaranteed bounds containing the exact solution. It may happen that the desired level of accuracy has been exceeded at some moment $t' < t_K$ and further Picard-Lindelöf iterations are unable to reduce the error. This situation may arise if the amount of internal points used for numerical integration on each interval is too small. In this case, we must enlarge the number of internal nodes (which will reduce integration and interpolation errors) and repeat



(a) (b)

Fig. 2 (a) Exact and approximate solutions with guaranteed bounds of deviation computed by the Ostrowski estimate. (b) A zoomed interval of exact and approximate solutions with bounds of deviation computed by the majorant.



(a) (b)

Fig. 3 (a) Exact and approximate solutions with guaranteed bounds of deviation computed by the advanced form of the estimate. (b) A zoomed interval of exact and approximate solutions with bounds of deviation computed by the majorant.

the computations. Numerical results illustrated in Figs. 2(a) and 3(a) show that the advanced majorant provides much sharper bounds of the deviation.

Values of the components of the estimate (the first term, the *estimate of* $\|\overline{e}\|$ and the *estimate of* $\|\widehat{e}\|$ from (42)) are presented in Table 1. We see that in this example the values of S_k were selected properly, so that interpolation and integration error estimates are insignificant with respect to the first term.

Example 2. The APL method works with stiff problems as well. Consider the classical stiff equation

Table 1 Components of the general estimate

Estimate of $\|e_j\|$	Estimate of $\|\overline{e_j}\|$	Estimate of $\|\hat{e}_j\|$
2.2658e-002	8.6160e-008	9.5725e-008
4.6095e-002	1.8847e-007	5.8148e-007
5.4949e-002	2.5299e-007	5.9301e-007
7.4818e-002	2.5768e-007	2.3618e-006
9.5993e-002	3.0190e-007	2.3699e-006
1.0302e-001	3.4216e-007	2.3807e-006
1.5427e-001	4.8963e-007	2.4320e-006
1.5647e-001	6.1877e-007	2.4999e-006
2.3495e-001	9.4891e-007	2.6183e-006
2.7145e-001	9.8935e-007	2.6328e-006
3.0533e-001	9.9923e-007	2.6373e-006
3.2838e-001	1.0158e-006	2.6404e-006
4.4629e-001	1.0182e-006	2.6517e-006

$$\begin{aligned} \frac{du}{dt} &= 50\cos(t) - 50u, \quad t = [0,1], \\ u(0) &= u_0 = 1 \end{aligned} \tag{46}$$

with the exact solution

$$u = \frac{1}{2501}e^{-50t} + \frac{2500}{2501}\cos(t) + \frac{50}{2501}\sin(t).$$

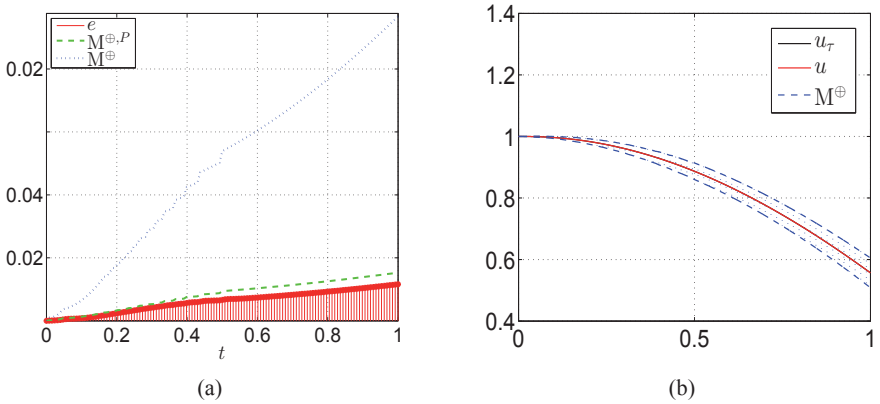
Analogously to the previous example, in Fig. 4(a) the general error (lines with dots on the top) estimated by the Ostrowski estimate (dotted line) and the advanced form of the estimate (dashed line) are illustrated. Another way to depict obtained results is shown in Fig. 4(b).

Example 3. The APL method can also be applied to stiff systems of ODEs. As an example, we consider the system

$$\begin{cases} \dfrac{du_1}{dt} = 998u_1 + 1998u_2, \\ \dfrac{du_2}{dt} = -999u_1 - 1999u_2, \\ u_1(t_0) = 1, u_2(t_0) = 1, \\ t \in [0, 5 \cdot 10^{-3}] \end{cases}$$

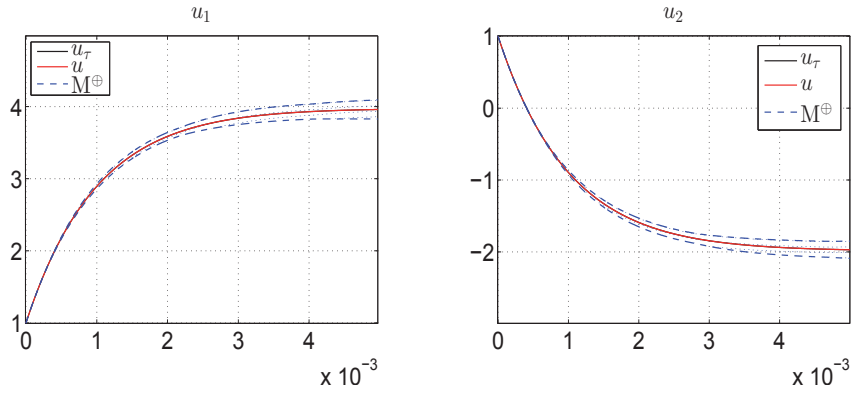
with the exact solutions $u_1 = 4e^{-t} - 3e^{-1000t}$ and $u_2 = -2e^{-t} + 3e^{-1000t}$. In Figs. 5(a), 5(b), 6(a) and 6(b), we present the same type of information (behavior of the solution and guaranteed bounds) as in the previous examples.

We note that for stiff equations getting an approximate solution with guaranteed and sharp error bounds requires much larger expenditures than in relatively simple



(a) (b)

Fig. 4 (a) Error and error majorants. (b) Exact and approximate solutions with a guaranteed deviation bound.



(a) (b)

Fig. 5 Exact solutions and approximate solutions of the system and guaranteed error bounds computed by the Ostrowski method.

Examples 1 and 2. This result is not surprising because (as it is quite natural to expect) for such type of problems fully reliable computations will be much more expensive.

References

1. E. A. Coddington and N. Levinson. *Theory of ordinary differential equations*. Tata McGraw-Hill, New York, 1972.

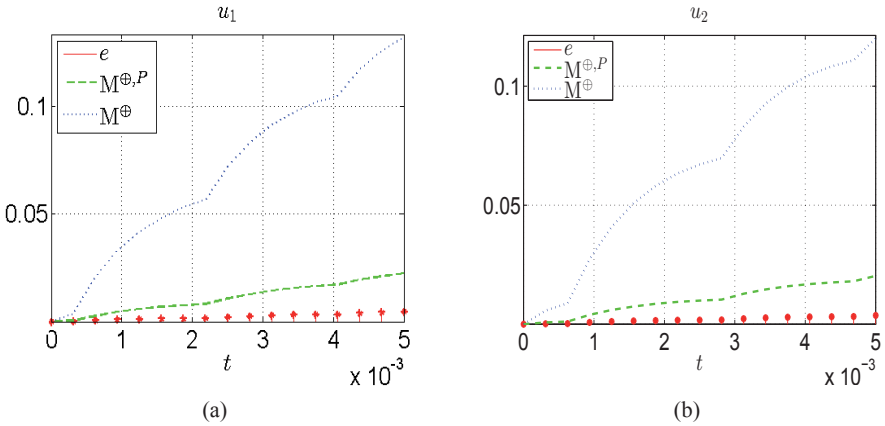


Fig. 6 The error and error majorants for the solutions u_1, u_2 of the system.

2. T. Eirola, A. M. Krasnosel'skii, M. A. Krasnosel'skii, N. A. Kuznetsov, and O. Nevanlinna. Incomplete corrections in nonlinear problems. *Nonlinear Analysis*, 25(7):717–728, 1995.
3. E. Lindelöf. Sur l'application de la méthode des approximations successives aux équations différentielles ordinaires du premier ordre. *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 114:454–457, 1894.
4. P. Neittaanmäki and S. Repin. *Reliable methods for computer simulation. Error control and a posteriori estimates*. Elsevier, Amsterdam, 2004.
5. O. Nevanlinna. Remarks on Picard-Lindelöf iteration. Part I. *BIT. Numerical Mathematics*, 29(2):328–346, 1989.
6. O. Nevanlinna. Remarks on Picard-Lindelöf iteration. Part II. *BIT. Numerical Mathematics*, 29(3):535–562, 1989.
7. A. Ostrowski. Les estimations des erreurs a posteriori dans les procédés itératifs. *Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences. Séries A et B*, 275:A275–A278, 1972.
8. S. Repin. *A posteriori estimates for partial differential equations*. Walter de Gruyter, Berlin, 2008.

PII

COMPUTABLE ESTIMATES OF THE DISTANCE TO THE EXACT SOLUTION OF THE EVOLUTIONARY REACTION-DIFFUSION EQUATION

by

S. Matculevich and S. Repin (2014)

Applied Mathematics and Computation, **247**: 329–347

PIII

A POSTERIORI ERROR ESTIMATES FOR TIME-DEPENDENT REACTION-DIFFUSION PROBLEMS BASED ON THE PAYNE–WEINBERGER INEQUALITY

by

S. Matculevich, P. Neittaanmäki, and S. Repin (2015)

Discrete and Continuous Dynamical Systems - Series A, AIMS, **35**(6): 2659–2677

PIV

ESTIMATES OF THE DISTANCE TO THE EXACT SOLUTION OF EVOLUTIONARY REACTION-DIFFUSION PROBLEMS BASED ON LOCAL POINCARÉ TYPE INEQUALITIES

by

S. Matculevich and S. Repin (2014)

Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov (POMI), **425**(1): 7–34

PV

SHARP BOUNDS OF CONSTANTS IN POINCARÉ TYPE INEQUALITIES FOR POLYGONAL DOMAINS

by

S. Matculevich and S. Repin (2015)

arXiv, math/1504.031662

Sharp bounds of constants in Poincaré-type inequalities for simplicial domains

S. Matculevich and S. Repin

Department of Mathematical Information Technology, University of Jyväskylä
FIN-40100 Jyväskylä, FINLAND
e-mails: svetlana.v.matculevich@jyu.fi, sergey.repin@jyu.fi

St. Petersburg Dept. of V.A. Steklov Institute of Mathematics of RAS
St. Petersburg, Russia

September 25, 2015

Abstract

The paper is concerned with sharp estimates of constants in classical Poincaré inequalities and Poincaré-type inequalities for functions having zero mean value in a simplicial domain or on a part of its boundary. These estimates are important for quantitative analysis of problems generated by differential equations where numerical approximations are typically constructed with the help of simplicial meshes. We suggest easily computable relations that provide sharp bounds of the respective constants and compare these results with analytical estimates (if they are known). In the last section, we present an example that shows possible applications of the results and derive a computable majorant of the difference between the exact solution of a boundary value problem and an arbitrary finite dimensional approximation computed on a simplicial mesh, which uses above mentioned constants.

1 Introduction

Let T be an open bounded connected domain in $\mathbb{R}^d$ ($d \geq 2$) with Lipschitz boundary ∂T. It is well known that the Poincaré inequality ([29, 30])

$$\|w\|_T \leq C_T^{\mathrm{P}} \|\nabla w\|_T \tag{1}$$

holds for any

$$w \in \widetilde{H}^1(T) := \Big\{ w \in H^1(T) \,\Big|\, \{\!|w|\!\}_T = 0 \Big\},$$

where $\|w\|_T$ denotes the norm in $L^2(T)$, $\{\!|w|\!\}_T := \frac{1}{|T|}\int_T w\,\mathrm{d}x$ is the mean value of w, and $|T|$ is the Lebesgue measure of T. The constant C_T^{P} depends only on T and d.

Poincaré-type inequalities also hold for

$$w \in \widetilde{H}^1(T,\Gamma) := \Big\{ w \in H^1(T) \,\Big|\, \{\!|w|\!\}_\Gamma = 0 \Big\},$$

where Γ is a measurable part of ∂T such that $\mathrm{meas}_{d-1}\Gamma > 0$ (in particular, Γ may coincide with the whole boundary). For any $w \in \widetilde{H}^1(T,\Gamma)$, we have two inequalities similar to (1). The first one

$$\|w\|_T \leq C_\Gamma^{\mathrm{p}} \|\nabla w\|_T \tag{2}$$

is another form of the Poincaré inequality (1), which is stated for a different set of functions and contains a different constant, i.e. $C_T^{\mathrm{P}} \leq C_\Gamma^{\mathrm{p}}$. The constant C_Γ^{p} is associated with the minimal positive eigenvalue of the problem

$$-\Delta u = \lambda u \ \text{ in } \ T; \quad \partial_n u = \lambda \{\!|u|\!\} \ \text{ on } \ \Gamma; \quad \partial_n u = 0 \ \text{ on } \ \partial T \backslash \Gamma, \quad \forall u \in H^1(T,\Gamma). \tag{3}$$

The second inequality

$$\|w\|_\Gamma \leq C_\Gamma^{\mathrm{Tr}} \|\nabla w\|_T \tag{4}$$

estimates the trace of $w \in \widetilde{H}^1(T,\Gamma)$ on Γ. It is associated with the minimal nonzero eigenvalue of the problem

$$-\Delta u = 0 \ \text{ in } \ T; \quad \partial_n u = \lambda u \ \text{ on } \ \Gamma; \quad \partial_n u = 0 \ \text{ on } \ \partial T \backslash \Gamma, \quad \forall u \in H^1(T,\Gamma). \tag{5}$$

The problem (5) is a special case of the Steklov problem with spectral parameter appearing in the boundary condition. It is called the sloshing problem, which describes the oscillations of fluid in a container. The extensive study of the properties of sloshing eigenvalues and eigenfunctions can be found in [9, 2, 15, 16, 17] and references therein. The question on the spectrum of the operator corresponding to the latter problem also have gained interest from the viewpoint of spectral geometry (see [11]).

Poincaré-type inequalities are often used in analysis of nonconforming approximations (e.g., discontinuous Galerkin or mortar methods), domain decomposition methods (see, e.g., [14, 8] and [34]), analysis of problems described in terms of vector valued functions (see, e.g., [10, 26]), a posteriori estimates, and other applications related to quantitative analysis of partial differential equations. In [13, 20], the analysis of error constants for piecewise constant and linear interpolations over triangular finite elements can be found. Paper [4] introduces fully computable two-sided bounds on the eigenvalues of the Laplace operator based on the approximation of the corresponding eigenfunction in the nonconforming Crouzeix-Raviart FE space. Therefore, exact values of respective constants (or sharp and guaranteed bounds of them) are interesting from both analytical and computational points of view.

It is known that for convex domains $C_T^{\mathrm{P}} \le \frac{\operatorname{diam}(T)}{\pi}$ (see [27]). However, for triangles this estimate was improved in [19], i.e., $C_T^{\mathrm{P}} \le \frac{\operatorname{diam}(T)}{j_{1,1}}$, where $j_{1,1} \approx 3.8317$ is the smallest positive root of the Bessel function J_1. Moreover, for isosceles triangles it was shown that

$$C_T^{\mathrm{P}} \le \overline{C}_T^{\mathrm{P},\triangle} := \operatorname{diam}(T) \cdot \begin{cases} \frac{1}{j_{1,1}} & \alpha \in (0, \frac{\pi}{3}], \\ \min\left\{ \frac{1}{j_{1,1}}, \frac{1}{j_{0,1}} \big(2(\pi-\alpha)\tan(\alpha/2)\big)^{-1/2} \right\} & \alpha \in (\frac{\pi}{3}, \frac{\pi}{2}], \\ \frac{1}{j_{0,1}} \big(2(\pi-\alpha)\tan(\alpha/2)\big)^{-1/2} & \alpha \in (\frac{\pi}{2}, \pi). \end{cases} \tag{6}$$

Here, $j_{0,1} \approx 2.4048$ and $j_{1,1} \approx 3.8317$ are the smallest positive roots of the Bessel functions J_0 and J_1, respectively. A lower bound of C_T^{P} for convex domains was derived in [6]. It was shown that

$$C_T^{\mathrm{P}} \ \ge \ \tfrac{\operatorname{diam}(T)}{2\, j_{0,1}}. \tag{7}$$

This estimate compliments the upper bound presented by the Payne–Weinberger estimate, and according to [1], is known to be the best lower bound for general domains with diameter scaling among all known so far. However, for triangles work [18] provides lower bound

$$C_T^{\mathrm{P}} \ \ge \ \tfrac{P}{4\,\pi}, \tag{8}$$

which improves (7) for some cases. Here, P is perimeter of T.

Exact values of C_Γ^{p} and C_Γ^{Tr} were derived in [25] for parallelepipeds, rectangles, and right triangles. Below, we present a concise summary of these results related to the case $d = 2$:

1. If $T := \operatorname{conv}\big\{(0,0),(0,h),(h,0)\big\}$, and $\Gamma := \big\{x_1 \in [0,h],\, x_2 = 0\big\}$ (i.e., Γ coincides with one of the legs of the isosceles right triangle), then

$$C_\Gamma^{\mathrm{p}} = \tfrac{h}{\zeta_0}, \quad \text{and} \quad C_\Gamma^{\mathrm{Tr}} = \Big(\tfrac{h}{\zeta_0\, \tanh(\hat{\zeta}_0)}\Big)^{1/2}, \tag{9}$$

   where ζ_0 and $\hat{\zeta}_0$ are unique roots in $(0,\pi)$ of the equations

$$z\cot(z) + 1 = 0 \quad \text{and} \quad \tan(z) + \tanh(z) = 0, \tag{10}$$

   respectively.

2. If $T := \operatorname{conv}\big\{(0,0),(0,h),\big(\frac{h}{2},\frac{h}{2}\big)\big\}$ and Γ coincides with the hypotenuse of the isosceles right triangle, then

$$C_\Gamma^{\mathrm{p}} = \tfrac{h}{2\zeta_0}, \quad \text{and} \quad C_\Gamma^{\mathrm{Tr}} = \big(\tfrac{h}{2}\big)^{1/2}.$$

It is worth noting that values of constant C_Γ^{Tr} on the right isosceles triangle follow from the exact solutions of Stecklov problem on the square. This specific case was mentioned in the work [11].

Exact value of constants in classical Poincaré inequality on equilateral triangle $T_{\pi/3} := \operatorname{conv}\big\{(0,0),(1,0),(\frac{1}{2},\frac{\sqrt{3}}{2})\big\}$ is derived in [28], i.e., $C^{\mathrm{p}}_{\Gamma,\pi/2,\pi/3} = \frac{3}{4\pi}$. Constants for the right isosceles triangles with legs $\frac{\sqrt{2}}{2}$ and 1, which are defined correspondingly as $T_{\pi/4} := \operatorname{conv}\big\{(0,0),(1,0),(\frac{1}{2},\frac{1}{2})\big\}$ and $T_{\pi/2} := \operatorname{conv}\big\{(0,0),(1,0),(0,1)\big\}$, are $C^{\mathrm{P}}_{\widehat{T},\pi/4} = \frac{1}{\sqrt{2}\pi}$

and $C^{\mathrm{P}}_{\widehat{T},\pi/2} = \frac{1}{\pi}$, respectively. The latter one can be found from [12] and [13]. Explicit formulas of the same constants for some three-dimensional domains can be found in papers [3] and [12].

Above mentioned results form a basis for deriving sharp bounds of the constants C^{p}_{Γ}, C^{Tr}_{Γ}, and C^{P}_{T} for arbitrary non-degenerate triangles and tetrahedrons, which are typical objects in various discretization methods. In Section 2, we deduce guaranteed and easily computable bounds of C^{p}_{Γ}, C^{Tr}_{Γ}, and C^{P}_{T} for triangular domains. The efficiency of these bounds is tested in Section 3, where C^{p}_{Γ}, C^{Tr}_{Γ} are compared with lower bounds computed numerically by solving generalized eigenvalue problem generated by Rayleigh quotients discretized over sufficiently representative sets of trial functions. In the same section, we make a similar comparison of numerical lower bounds related to the constant C^{P}_{T} with obtained upper bounds and existing estimates known from works [18, 19] and [6]. Lower bounds of the constants presented in Section 3 have been computed by two independent codes. The first code is based on MATLAB Symbolic Math Toolbox [33], and the second one uses The FEniCS Project [21]. Section 4 is devoted to tetrahedrons. We combine numerical and theoretical estimates in order to derive two sided bounds of the constants. For convinience of the reader, we collect all the figures in Appendix 6. Finally, in Section 5 we present an example that shows one possible application of the estimates considered in previous sections. Here, the constants are used in order to deduce a guaranteed and fully computable upper bound of the distance between the exact solution of an elliptic boundary value problem and an arbitrary function (approximation) in the respective energy space.

2 Majorants of C^{p}_{Γ} and C^{Tr}_{Γ} for triangular domains

We set

$$T = \mathrm{conv}\big\{(0,0),(h,0),\big(h\rho\cos\alpha,\, h\rho\sin\alpha\big)\big\} \quad \text{and} \quad \Gamma := \big\{x_1 \in [0,h];\;\; x_2 = 0\big\}, \tag{11}$$

where $\rho > 0$, $h > 0$, and $\alpha \in (0,\pi)$ are geometrical parameters that fully define a triangle T (see Fig. 1). Lemma below is based on analysis of the mapping from the reference triangles to (11) using well known transformation of the integrals (see, e.g., [7]). Easily computable bounds of C^{p}_{Γ} and C^{Tr}_{Γ} are presented below.

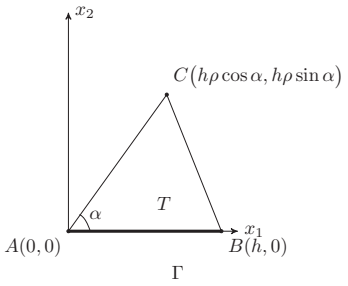


Figure 1: Simplex in $\mathbb{R}^2$.

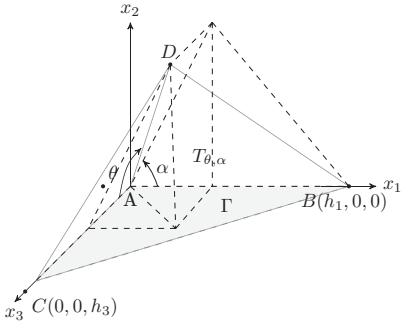


Figure 2: Simplex in $\mathbb{R}^3$.

Lemma 1 *For any $w \in \widetilde{H}^1(T,\Gamma)$, the estimates*

$$\|w\|_T \;\le\; C^{\mathrm{p}}_{\Gamma}\, h\, \|\nabla w\|_T \quad \text{and} \quad \|w\|_{\Gamma} \le\; C^{\mathrm{Tr}}_{\Gamma}\, h^{1/2}\, \|\nabla w\|_T \tag{12}$$

hold with

$$C^{\mathrm{p}}_{\Gamma} \le \overline{C}^{\mathrm{p}}_{\Gamma} = \min\Big\{c^{\mathrm{p}}_{\pi/2}\, C^{\mathrm{p}}_{\widehat{\Gamma},\pi/2},\; c^{\mathrm{p}}_{\pi/4}\, C^{\mathrm{p}}_{\widehat{\Gamma},\pi/4}\Big\} \quad \text{and} \quad C^{\mathrm{Tr}}_{\Gamma} \le \overline{C}^{\mathrm{Tr}}_{\Gamma} = \min\Big\{c^{\mathrm{Tr}}_{\pi/2}\, C^{\mathrm{Tr}}_{\widehat{\Gamma},\pi/2},\; c^{\mathrm{Tr}}_{\pi/4}\, C^{\mathrm{Tr}}_{\widehat{\Gamma},\pi/4}\Big\},$$

respectively. Here,

$$c^{\mathrm{p}}_{\pi/2} = \mu^{1/2}_{\pi/2}, \quad c^{\mathrm{Tr}}_{\pi/2} = \big(\rho\,\sin\alpha\big)^{-1/2}\, c^{\mathrm{p}}_{\pi/2}, \quad c^{\mathrm{p}}_{\pi/4} = \mu^{1/2}_{\pi/4}, \quad c^{\mathrm{Tr}}_{\pi/4} = \big(2\rho\,\sin\alpha\big)^{-1/2}\, c^{\mathrm{p}}_{\pi/4},$$

where

$$\mu_{\pi/2}(\rho,\alpha) = \tfrac{1}{2}\Big(1+\rho^2+\big(1+\rho^4+2\,\rho^2\,\cos 2\alpha\big)^{1/2}\Big), \tag{13}$$

$$\mu_{\pi/4}(\rho,\alpha) = 2\rho^2 - 2\rho\,\cos\alpha + 1 + \big((2\rho^2+1)(2\rho^2+1-4\rho\,\cos\alpha+4\rho^2\,\cos 2\alpha)\big)^{1/2}, \tag{14}$$

and $C^{\rm p}_{\widehat{\Gamma},\pi/2} \approx 0.49291$, $C^{\rm Tr}_{\widehat{\Gamma},\pi/2} \approx 0.65602$ *and* $C^{\rm p}_{\widehat{\Gamma},\pi/4} \approx 0.24646$, $C^{\rm Tr}_{\widehat{\Gamma},\pi/4} \approx 0.70711$.

Proof: Consider a linear mapping $\mathcal{F}_{\pi/2} : \widehat{T}_{\pi/2} \to T$

$$x = \mathcal{F}_{\pi/2}\,(\hat{x}) = B_{\pi/2}\,\hat{x}, \quad \text{where} \quad B_{\pi/2} = \begin{pmatrix} h & \rho h \cos\alpha \\ 0 & \rho h \sin\alpha \end{pmatrix}, \quad \det B_{\pi/2} = \rho h^2\, \sin\alpha.$$

For any $\hat{w} \in \widetilde{H}^1(\widehat{T}_{\pi/2}, \widehat{\Gamma})$, we have the estimate

$$\| \hat{w} \|_{\widehat{T}_{\pi/2}} \leq C^{\rm p}_{\widehat{\Gamma},\pi/2} \| \nabla \hat{w} \|_{\widehat{T}_{\pi/2}}, \tag{15}$$

where $C^{\rm p}_{\widehat{\Gamma},\pi/2}$ is the constant associated with the basic simplex

$$T_{\pi/2} := \mathrm{conv}\big\{(0,0),(1,0),(0,1)\big\}. \tag{16}$$

Note that

$$\| \hat{w} \|^2_{\widehat{T}_{\pi/2}} = \tfrac{1}{\rho h^2 \sin\alpha} \| w \|^2_T, \tag{17}$$

and

$$\| \nabla \hat{w} \|^2_{\widehat{T}_{\pi/2}} \leq \tfrac{1}{\rho h^2 \sin\alpha} \int_T A_{\pi/2}(h,\rho,\alpha) \nabla w \cdot \nabla w \, \mathrm{d}x, \tag{18}$$

where

$$A_{\pi/2}(h,\rho,\alpha) = h^2 \begin{pmatrix} 1+\rho^2\, \cos^2\alpha & \rho^2 \sin\alpha\, \cos\alpha \\ \rho^2 \sin\alpha\, \cos\alpha & \rho^2 \sin^2\alpha \end{pmatrix}.$$

It is not difficult to see that

$$\lambda_{\max}(A_{\pi/2}) = h^2 \mu_{\pi/2}(\rho,\alpha), \quad \mu_{\pi/2}(\rho,\alpha) = \tfrac{1}{2}\Big(1+\rho^2+\big(1+\rho^4+2\,\cos 2\alpha\,\rho^2\big)^{1/2}\Big),$$

where $\mu_{\pi/2}(\rho,\alpha)$ is defined in (13). We use (15), (17), and (18), and obtain

$$\| w \|_T \,\leq\, c^{\rm p}_{\pi/2}\, C^{\rm p}_{\widehat{\Gamma},\pi/2}\, h \,\| \nabla w \|_T, \quad c^{\rm p}_{\pi/2}(\rho,\alpha) = \mu^{1/2}_{\pi/2}(\rho,\alpha). \tag{19}$$

Next, in view of (4), for any $\hat{w} \in \widetilde{H}^1(\widehat{T}_{\pi/2}, \widehat{\Gamma})$ we have

$$\| \hat{w} \|_{\widehat{\Gamma}} \leq C^{\rm Tr}_{\widehat{\Gamma},\pi/2} \| \nabla \hat{w} \|_{\widehat{T}_{\pi/2}},$$

where $C^{\rm Tr}_{\widehat{\Gamma},\pi/2}$ is the constant associated with the reference simplex

$$\widehat{T}_{\pi/4} := \mathrm{conv}\big\{(0,0),(1,0),\big(\tfrac{1}{2},\tfrac{1}{2}\big)\big\}. \tag{20}$$

Since

$$\| \hat{w} \|^2_{\widehat{\Gamma}} = \tfrac{1}{h} \| w \|^2_{\Gamma},$$

we obtain

$$\| w \|_\Gamma \,\leq\, c^{\rm Tr}_{\pi/2}\, C^{\rm Tr}_{\widehat{\Gamma},\pi/2}\, h^{1/2} \| \nabla w \|_T, \quad c^{\rm Tr}_{\pi/2}(\rho,\alpha) = \Big(\tfrac{\mu_{\pi/2}(\rho,\alpha)}{\rho \sin\alpha}\Big)^{1/2}. \tag{21}$$

The mapping

$$x = \mathcal{F}_{\pi/4}(\hat{x}) = B_{\pi/4}\,\hat{x}, \quad \text{where} \quad B_{\pi/4} = \begin{pmatrix} h & 2\rho h \cos\alpha - h \\ 0 & 2\rho h \sin\alpha \end{pmatrix} \quad \text{and} \quad \det B_{\pi/4} = 2\rho h^2\, \sin\alpha > 0,$$

yields another pair of estimates for the functions in $\widetilde{H}^1(T,\Gamma)$:

$$\| w \|_T \,\leq\, c^{\rm p}_{\pi/4}\, C^{\rm p}_{\widehat{\Gamma},\pi/4}\, h \,\| \nabla w \|_T, \quad c^{\rm p}_{\pi/4}(\rho,\alpha) = \mu^{1/2}_{\pi/4}(\rho,\alpha), \tag{22}$$

and

$$\| w \|_\Gamma \,\leq\, c^{\rm Tr}_{\pi/4}\, C^{\rm Tr}_{\widehat{\Gamma},\pi/4}\, h^{1/2} \| \nabla w \|_T, \quad c^{\rm Tr}_{\pi/4}(\rho,\alpha) = \Big(\tfrac{\mu_{\pi/4}(\rho,\alpha)}{2\rho \sin\alpha}\Big)^{1/2}, \tag{23}$$

where $\mu_{\pi/4}(\rho,\alpha)$ is defined in (14). Now, (12) follows from (19), (21), (22), and (23). □

Remark 1 The selection of Γ and α in T depends on the finite element implementation, i.e., we select Γ out three edges of T such that it satisfies the condition $\{\!|w|\!\}_\Gamma = 0$. According to the results of numerical experiments, to provide optimal values of $C_\Gamma^{\rm P}$, Γ must coincide with with longest side of T, and between two adjacent angles we select minimum one to be α. For the constant $C_\Gamma^{\rm Tr}$, minimum values are attained for α lying in the interval $(\frac{\pi}{3}, \frac{2\pi}{3})$ (see Section 3).

Remark 2 Let us show that obtained w belongs to the correct space. Assume that $w \in \widetilde{H}^1(\widehat{T}, \widehat{\Gamma})$ and consider mean value of w after the transformation. We note that

$$\{\!|w|\!\}_\Gamma := \int\limits_\Gamma w(x)\,{\rm d}s = h \int\limits_{\widehat{\Gamma}} w(x(\hat{x}))\,{\rm d}\hat{s} = h \int\limits_{\widehat{\Gamma}} \hat{w}\,{\rm d}\hat{s} = 0.$$

Therefore, the mean value of the function on the boundary remains zero and indeed $w \in \widetilde{H}^1(T, \Gamma)$.

Analogously to Lemma 1, one can obtain the upper bound of the constant in (1). For that we consider three reference triangle $T_{\pi/2}$, $T_{\pi/4}$ (defined in (16) and (20)), and $T_{\pi/3} := \mathrm{conv}\{(0,0), (1,0), (\frac{1}{2}, \frac{\sqrt{3}}{2})\}$.

Lemma 2 *For any $w \in \widetilde{H}^1(T)$, the estimate of the constant in*

$$\|w\|_T \;\le\; C_\Omega^{\rm P} h\, \|\nabla w\|_T \tag{24}$$

has the form

$$C_\Gamma^{\rm p} \le \overline{C}_T^{\rm P} = \min\left\{ \overline{c}_{\pi/4}\, C^{\rm P}_{\widehat{T},\pi/4};\; \overline{c}_{\pi/3}\, C^{\rm P}_{\widehat{T},\pi/3};\; \overline{c}_{\pi/2}\, C^{\rm P}_{\widehat{T},\pi/2} \right\} \tag{25}$$

Here,

$$\overline{c}_{\pi/4} = \mu^{1/2}_{\pi/4}, \quad \overline{c}_{\pi/3} = \mu^{1/2}_{\pi/3}, \quad \textit{and} \quad \overline{c}_{\pi/2} = \mu^{1/2}_{\pi/2},$$

where $\mu_{\pi/2}$ and $\mu_{\pi/4}$ are defined in (13) *and* (14) *and*

$$\mu_{\pi/3}(\rho,\alpha) = \tfrac{2}{3}(1 + \rho^2 - \rho\,\cos\alpha) + 2\big(\tfrac{1}{9}(1+\rho^2-\rho\,\cos\alpha)^2 - \tfrac{1}{3}\rho^2\sin^2\alpha\big)^{1/2}, \tag{26}$$

and $C^{\rm P}_{\widehat{T},\pi/4} = \frac{1}{\sqrt{2}\pi}$, $C^{\rm P}_{\widehat{T},\pi/3} = \frac{3}{4\pi}$, $C^{\rm P}_{\widehat{T},\pi/2} = \frac{1}{\pi}$.

Proof: The mapping $\mathcal{F}_{\pi/2} : \widehat{T}_{\pi/2} \to T$ coincides with (2) from Lemma 1. Therefore the bound

$$\|\,w\,\|_T \;\le\; \overline{c}_{\pi/2}\, C^{\rm P}_{\widehat{T},\pi/2}\, h\, \|\,\nabla w\,\|_T, \quad \overline{c}_{\pi/2}(\rho,\alpha) = \mu^{1/2}_{\pi/2}(\rho,\alpha) \tag{27}$$

is obtained by following the steps of the previous proof. From analysis of mappings

$$x = \mathcal{F}_{\pi/3}(\hat{x}) = B_{\pi/3}\,\hat{x}, \quad \text{where} \quad B_{\pi/3} = \begin{pmatrix} h & \frac{h}{\sqrt{3}}(2\rho\cos\alpha - 1) - h \\ 0 & \frac{2h}{\sqrt{3}}\rho\sin\alpha \end{pmatrix}, \quad \det B_{\pi/3} = \tfrac{2h^2}{\sqrt{3}}\,\sin\alpha > 0,$$

and

$$x = \mathcal{F}_{\pi/4}(\hat{x}) = B_{\pi/4}\,\hat{x}, \quad \text{where} \quad B_{\pi/4} = \begin{pmatrix} h & 2\rho h\cos\alpha - h \\ 0 & 2\rho h\sin\alpha \end{pmatrix}, \quad \det B_{\pi/4} = 2\rho h^2\,\sin\alpha > 0,$$

we obtain alternative estimates for function $w \in \widetilde{H}^1(T)$

$$\|\,w\,\|_T \;\le\; \overline{c}_{\pi/3}\, C^{\rm P}_{\widehat{T},\pi/3}\, h\, \|\,\nabla w\,\|_T, \quad \overline{c}_{\pi/3}(\rho,\alpha) = \mu^{1/2}_{\pi/3}(\rho,\alpha), \tag{28}$$

$$\|\,w\,\|_T \;\le\; \overline{c}_{\pi/4}\, C^{\rm P}_{\widehat{T},\pi/4}\, h\, \|\,\nabla w\,\|_T, \quad \overline{c}_{\pi/4}(\rho,\alpha) = \mu^{1/2}_{\pi/4}(\rho,\alpha), \tag{29}$$

where $\mu_{\pi/3}(\rho,\alpha)$ and $\mu_{\pi/4}(\rho,\alpha)$ are defined in (26) and (14), respectively. Therefore, (25) follows from (27), (28), and (29). Analogously to Remark 2, if $\hat{w} \in \widetilde{H}^1(\widehat{T})$, it follows that $w \in \widetilde{H}^1(T)$. □

3 Minorants of C_Γ^{p} and C_Γ^{Tr} for triangular domains

3.1 Two-sided bounds of C_Γ^{p} and C_Γ^{Tr}

Majorants of C_Γ^{p} and C_Γ^{Tr} provided by Lemma 1 should be compared with the corresponding minorants, which can be found by solving generalized eigenvalue problem generated by the discretized Rayleigh quotients

$$\mathcal{R}_\Gamma^{\mathrm{p}}[w] = \frac{\|\nabla w\|_T}{\|w-\{w\}_\Gamma\|_T} \quad \text{and} \quad \mathcal{R}_\Gamma^{\mathrm{Tr}}[w] = \frac{\|\nabla w\|_T}{\|w-\{w\}_\Gamma\|_\Gamma}. \tag{30}$$

Here, w is approximated by the basis of finite dimensional subspaces $V^N \subset H^1(T)$ formed by sufficiently representative collections of test functions. For this purpose, we use either power or Fourier series and introduce

$$V_1^N := \operatorname{span}\Big\{\, x^i y^j \,\Big\}, \quad \text{and} \quad V_2^N := \operatorname{span}\Big\{\, \cos(\pi i x)\cos(\pi j y) \,\Big\}, \;\; i,j = 0,\dots,N, \;\; (i,j) \neq (0,0),$$

with $\dim V_1^N = \dim V_2^N = M(N) := (N+1)^2 - 1$. The corresponding constants are denoted by $\underline{C}_\Gamma^{M,\mathrm{p}}$ and $\underline{C}_\Gamma^{M,\mathrm{Tr}}$, where $M(N)$ indicates the amount of used basis functions in the finite dimensional subspace. Since above defined finite dimensional spaces are limit dense in $H^1(T)$, the minorants tend to exact constants as $M(N)$ tends to infinity. The quotients (30) follow from the definition of the constants C_Γ^{p} and C_Γ^{Tr} for $w \in H^1(T)$, i.e.,

$$\|w - \{w\}_\Gamma\|_T \le C_\Gamma^{\mathrm{p}} \|\nabla w\|_T \quad \text{and} \quad \|w - \{w\}_\Gamma\|_\Gamma \le C_\Gamma^{\mathrm{Tr}} \|\nabla w\|_T. \tag{31}$$

Embeddings (31) are justified by the equivalence of quantity $\|w\|_l := \|\nabla w\|_T + \big|\int_\Gamma w \,\mathrm{d}s\big|$ to the norm of $H^1(T)$, so that the existence of C_Γ^{p} and C_Γ^{Tr} follows automatically.

Numerical results presented below are obtained with two different codes based on MATLAB Symbolic Math Toolbox [33] and The FEniCS Project [21]. Table 1 demonstrates that the ratios between exact constants and their approximate values (for the selected ρ and α) are quite close to 1 (as it is expected) even for relatively small N. Therefore, we select $N = 6$ or 7 in tests discussed below.

		$\alpha=\frac{\pi}{2}$, $\rho=1$		$\alpha=\frac{\pi}{4}$, $\rho=\frac{\sqrt{2}}{2}$	
N	$M(N)$	$\underline{c}^M_{\mathrm{p},\pi/2}$	$\underline{c}^M_{\mathrm{Tr},\pi/2}$	$\underline{c}^M_{\mathrm{p},\pi/4}$	$\underline{c}^M_{\mathrm{Tr},\pi/4}$
1	3	0.8801	0.9561	0.8647	1.0000
2	8	0.9945	0.9898	0.9925	1.0000
3	15	0.9999	0.9998	0.9962	1.0000
4	24	1.0000	0.9999	1.0000	1.0000
5	35	1.0000	1.0000	1.0000	1.0000
6	48	1.0000	1.0000	1.0000	1.0000

Table 1: Ratios of $\underline{c}^M_{\mathrm{p},\pi/2}$, $\underline{c}^M_{\mathrm{Tr},\pi/2}$ and $\underline{c}^M_{\mathrm{p},\pi/4}$, $\underline{c}^M_{\mathrm{Tr},\pi/4}$ with respect to increasing N and $M(N)$.

In Figs. 3a and 3c, we depict $\underline{C}_\Gamma^{M,\mathrm{p}}$ for $M(N) = 48$ (thin line) for different T with $\rho = \frac{\sqrt{2}}{2}$, $\rho = 1$, and $\alpha \in (0,\pi)$. Guaranteed upper bounds $\overline{C}^{\mathrm{p}}_{\pi/2} = c^{\mathrm{p}}_{\pi/2}\, C^{\mathrm{p}}_{\widehat{\Gamma},\pi/2}$ and $\overline{C}^{\mathrm{p}}_{\pi/4} = c^{\mathrm{p}}_{\pi/4}\, C^{\mathrm{p}}_{\widehat{\Gamma},\pi/4}$ are depicted by dashed lines. By the bold line, we emphasize on $\overline{C}^{\mathrm{p}}_{\pi/2}$ and $\overline{C}^{\mathrm{p}}_{\pi/4}$, which present $\overline{C}^{\mathrm{p}}_\Gamma$ as it is defined in Lemma 1. Analogously in Figs. 4a and 4b, the lower bound $\underline{C}_\Gamma^{M,\mathrm{Tr}}$ (for $M(N) = 48$) of the constant C_Γ^{Tr} is presented together with the upper bound $\overline{C}_\Gamma^{\mathrm{Tr}}$ (which is defined as minimum of $\overline{C}^{\mathrm{Tr}}_{\pi/2} = c^{\mathrm{Tr}}_{\pi/2}\, C^{\mathrm{Tr}}_{\widehat{\Gamma},\pi/2}$ and $\overline{C}^{\mathrm{Tr}}_{\pi/4} = c^{\mathrm{Tr}}_{\pi/4}\, C^{\mathrm{Tr}}_{\widehat{\Gamma},\pi/4}$; dashed lines. Parameter M is fixed to 48 since the difference of order 1e-8 between $\underline{C}_\Gamma^{M,\mathrm{Tr}}$ (for bigger M) becomes unnoticeable. In the digital form, the information is represented in Table 2.

Fig. 3a corresponds to the case $\rho = \frac{\sqrt{2}}{2}$. It is worth noting that for $\alpha = \frac{\pi}{4}$ the lower bound $\underline{C}_\Gamma^{M,\mathrm{p}}$ coincides with constant C_Γ^{p} ($\overline{C}^{\mathrm{p}}_{\pi/4}$). This happens because for $\alpha = \frac{\pi}{4}$ the mapping $\mathcal{F}_{\pi/4}$ is identical (see, e.g., Fig. 3b). Analogous coincidence can be observed for C_Γ^{Tr} ($\overline{C}^{\mathrm{Tr}}_{\pi/4}$) in Fig. 4a. In Fig. 3c, the curve corresponding to $\underline{C}_\Gamma^{M,\mathrm{p}}$ coincides with the line of C_Γ^{p} ($\overline{C}^{\mathrm{p}}_{\pi/2}$) at the point $\alpha = \frac{\pi}{2}$ (due to the fact for this angle $\mathcal{F}$ is the identical mapping and T coincides with $\widehat{T}_{\pi/2}$ (see Fig. 3d)). Fig. 4b exposes similar results for $\underline{C}_\Gamma^{M,\mathrm{Tr}}$ and C_Γ^{Tr} ($\overline{C}^{\mathrm{Tr}}_{\pi/2}$). Figs. 5 and 6 demonstrate the same bounds for more interesting $\rho = \frac{\sqrt{3}}{2}$ and $\frac{3}{2}$, which stay quite efficient even for the cases unrelated to the reference triangles, e.i., if $\rho = \frac{\sqrt{3}}{2}$, $I_{\mathrm{eff}} \approx 1.0463 \div 0.1300$ for C_Γ^{p} and $I_{\mathrm{eff}} \approx 1.0363 \div 1.3388$ for C_Γ^{Tr}, and if $\rho = \frac{3}{2}$, $I_{\mathrm{eff}} \approx 1.0249 \div 0.1634$ for C_Γ^{p} and $I_{\mathrm{eff}} \approx 1.2917 \div 1.7643$ for C_Γ^{Tr}.

	$\rho=\frac{\sqrt{2}}{2}$				$\rho=1$			
α	$\underline{C}_\Gamma^{48,\mathrm{P}}$	$\overline{C}_\Gamma^{\mathrm{P}}$	$\underline{C}_\Gamma^{48,\mathrm{Tr}}$	$\overline{C}_\Gamma^{\mathrm{Tr}}$	$\underline{C}_\Gamma^{48,\mathrm{P}}$	$\overline{C}_\Gamma^{\mathrm{P}}$	$\underline{C}_\Gamma^{48,\mathrm{Tr}}$	$\overline{C}_\Gamma^{\mathrm{Tr}}$
$\pi/18$	0.2429	0.2657	1.2786	1.5386	0.3245	0.3486	1.2572	1.6971
$\pi/9$	0.2414	0.2627	0.9289	1.0838	0.3248	0.3493	0.9058	1.2116
$\pi/6$	0.2389	0.2577	0.7919	0.8792	0.3268	0.3527	0.7632	1.0118
$2\pi/9$	0.2379	0.2507	0.7259	0.7543	0.3339	0.3636	0.6906	0.9201
$5\pi/18$	0.2632	0.2722	0.6945	0.7503	0.3514	0.3884	0.6529	0.9003
$\pi/3$	0.3008	0.3220	0.6829	0.8348	0.3809	0.4269	0.6362	0.8634
$7\pi/18$	0.3382	0.3694	0.6840	0.8432	0.4173	0.4721	0.6332	0.7840
$4\pi/9$	0.3740	0.4140	0.6947	0.7973	0.4556	0.5187	0.6404	0.7162
$\pi/2$	0.4075	0.4554	0.7136	0.7801	0.4929	0.4929	0.6560	0.6560
$5\pi/9$	0.4382	0.4933	0.7409	0.7973	0.5280	0.5340	0.6797	0.7162
$11\pi/18$	0.4660	0.5165	0.7779	0.8432	0.5600	0.5710	0.7125	0.7840
$2\pi/3$	0.4905	0.5361	0.8274	0.9118	0.5884	0.6037	0.7569	0.8634
$13\pi/18$	0.5115	0.5552	0.8948	1.0040	0.6129	0.6318	0.8175	0.9607
$7\pi/9$	0.5289	0.5720	0.9898	1.1292	0.6332	0.6550	0.9033	1.0874
$5\pi/6$	0.5426	0.5856	1.1334	1.3107	0.6492	0.6733	1.0332	1.2673
$8\pi/9$	0.5524	0.5956	1.3796	1.6118	0.6607	0.6865	1.2565	1.5623
$17\pi/18$	0.5583	0.6017	1.9436	2.2851	0.6676	0.6944	1.7692	2.2179

Table 2: Lower and upper bounds of C_Γ^{P} and C_Γ^{Tr} with respect to α and for $\rho=\frac{\sqrt{2}}{2}$ and 1.

3.2 Two-sided bounds of C_T^{P}

The spaces V_1^N and V_2^N are also used for analysis of the quotient $\mathcal{R}_T[w]=\frac{\|\nabla w\|_T}{\|w-\{w\}_T\|_T}$, which yields guaranteed lower bounds of the constant in (1) denoted by $\underline{C}_T^{M,\mathrm{P}}$. Obtained numerical results are compared with above presented estimate $\overline{C}_T^{\mathrm{P}}$, $\overline{C}_T^{\mathrm{P},\oplus}:=\frac{\mathrm{diam}(T)}{j_{1,1}}$, and $\underline{C}_T^{\mathrm{P}}:=\max\left\{\frac{\mathrm{diam}(T)}{2\,j_{0,1}},\frac{P}{4\,\pi}\right\}$, which follow from (7) and (8), respectively.

In Figs. 7a, 7b, 7d, and 7e, we illustrate $\underline{C}_T^{M,\mathrm{P}}(M(N)=48)$ together with $\overline{C}_T^{\mathrm{P}}$, $\overline{C}_T^{\mathrm{P},\oplus}$ and $\underline{C}_T^{\mathrm{P}}$ with respect to $\alpha\in(0,\pi)$ for T with $\rho=\frac{\sqrt{2}}{2}$, $\frac{\sqrt{3}}{2}$, $\frac{3}{2}$, and 2. We see that $\underline{C}_T^{48,\mathrm{P}}$ indeed lies within the admissible two-sided bound. From these figures, it is obvious that obtained upper bounds $\overline{C}_T^{\mathrm{P}}$ are sharper than existing estimates $\overline{C}_T^{\mathrm{P},\oplus}$ for T with $\rho\neq 1$. True values of the constant lie between the bold and dashed lines, but closer to the bold line, which practically illustrates the constant (this follows from the fact that increasing $M(N)$ does not provide a noticeable change for the line, e.g., for $M(N)=63$ maximal difference with respect to Fig. does not exceed 1e-8). Also, we note that, the lower bound $\underline{C}_T^{\mathrm{P}}$ is quite efficient, and, moreover, asymptotically exact for $\alpha\to\pi$.

Due to [19], we know the improved upper bound $\overline{C}_T^{\mathrm{P},\triangle}$ (cf. (6)) for isosceles triangles. In Fig. 7c, we compare $\underline{C}_T^{M,\mathrm{P}}$ $(M(N)=48)$ with both upper bounds $\overline{C}_T^{\mathrm{P}}$ (from the Lemma 2) and $\overline{C}_T^{\mathrm{P},\triangle}$. It is easy to see that $\overline{C}_T^{\mathrm{P},\triangle}$ is rather accurate and for $\alpha\to 0$ and $\alpha\to\pi$ provide almost exact estimates. $\overline{C}_T^{\mathrm{P}}$ improves $\overline{C}_T^{\mathrm{P},\triangle}$ only for some α. Moreover, the lower bound $\underline{C}_T^{48}$ indeed converges to $\overline{C}_T^{\mathrm{P},\triangle}$ as T degenerates when α tends to 0 (see [19]).

3.3 Shape of the minimizer

Exact constants in (2) and (4) are generated by minimal positive eigenvalues of (3) and (5). This section presents results related to the respective eigenfunctions. In order to depict all of them in a unified form, we use barycentric coordinates $\lambda_i\in(0,1)$, $i=1,2,3$, $\sum_{i=1}^{3}\lambda_i=1$. Figs. 8 and 9 show eigenfunctions computed for isosceles triangles with different angles α between two legs (zero mean condition is imposed on one of the legs). They are constructed in the process of finding $\underline{C}_\Gamma^{M,\mathrm{p}}$ and $\underline{C}_\Gamma^{M,\mathrm{Tr}}$ and normalized such that the maximal value of a function is equal to 1. For $\alpha=\frac{\pi}{2}$, the exact eigenfunction associated with the smallest positive eigenvalue $\lambda_\Gamma^{\mathrm{p}}=\left(\frac{z_0}{h}\right)^2$, is known (see [25]). It is (see Fig 8d)

$$u_\Gamma^{\mathrm{p}}=\cos(\tfrac{z_0x_1}{h})+\cos\left(\tfrac{z_0(x_2-h)}{h}\right),$$

where z_0 is the root of the first equation in (10). We can compare it with the approximate eigenfunction $u_\Gamma^{M,\mathrm{p}}$ computed by minimization of $\mathcal{R}_\Gamma^{\mathrm{p}}[w]$. It is depicted in Fig. 8c.

Analogous results for eigenfunctions related to the constant $\underline{C}_\Gamma^{M,\mathrm{Tr}}$ are presented in Fig. 9. Again, for $\alpha=\frac{\pi}{2}$ we know the exact one

$$u_\Gamma^{\mathrm{Tr}}=\cos(\hat{z}_0x_1)\,\cosh\left(\hat{z}_0(x_2-h)\right)+\cosh(\hat{z}_0\,x_1)\,\cos\left(\hat{z}_0(x_2-h)\right),$$

where $\hat{z}_0$ is the root of second equation in (10) (see Fig. 9d), which minimizes the quotient $\mathcal{R}_\Gamma^{\mathrm{Tr}}[w]$ associated with the smallest positive eigenvalue $\lambda_\Gamma^{\mathrm{Tr}}=\frac{\hat{z}_0\tanh(\hat{z}_0)}{h}$. It is easy to see that with the assigned value of $M(N)$ numerical

approximation practically coincides with the exact one.

In the above considered examples, the eigenfunctions associated with minimal positive eigenvalues expose a continuous evolution with respect to α. However, this is not true for the quotient $\mathcal{R}_T[w]$, where the minimizer may cardinally change the profile. Fig. 7c indicates a possibility of such rapid change at $\alpha = \frac{\pi}{3}$, where the curve (related to $\underline{C}_T^{48}$) obviously becomes non-smooth. This happens because equilateral triangle has double eigenvalue and the function, minimizing $\mathcal{R}_T[w]$ over V_1^N, changes its profile. Figs. 10d–10i show three eigenfunctions $u_{T,1}^{48}$, $u_{T,2}^{48}$, and $u_{T,3}^{48}$ related to three minimal eigenvalues $\lambda_{T,1}^{48}$, $\lambda_{T,2}^{48}$, and $\lambda_{T,3}^{48}$ computed in the process of minimization of $\mathcal{R}_T[w]$. All functions are computed for isosceles triangles and are sorted in accordance with increasing values of the respective eigenvalues. Fig. 10 illustrates these three eigenfunctions for T with angles $\alpha = \frac{\pi}{3}$, $\frac{\pi}{3} + \varepsilon$, and $\frac{\pi}{3} - \varepsilon$, where $\varepsilon = \frac{\pi}{36}$. It is easy to see that at $\alpha = \pi/3$ the first and the second eigenfunctions change places. Table 3 presents the corresponding results in the digital form.

It is worth noting that for equilateral triangles two minimal eigenfunctions are known (see [23]):

$$u_1 = \cos\Big(\tfrac{2\,\pi}{3}(2\,x_1 - 1)\Big) - \cos\Big(\tfrac{2\,\pi}{\sqrt{3}}x_2\Big)\cos\Big(\tfrac{\pi}{3}(2\,x_1 - 1)\Big),$$
$$u_2 = \sin\Big(\tfrac{2\,\pi}{3}(2\,x_1 - 1)\Big) + \cos\Big(\tfrac{2\,\pi}{\sqrt{3}}x_2\Big)\sin\Big(\tfrac{\pi}{3}(2\,x_1 - 1)\Big).$$

These functions practically coincide with the functions $u_{T,1}^{48}$ and $u_{T,2}^{48}$ presented in Fig. 10d. Finally, we note that this phenomenon (change of the minimal eigenfunction) does not appear for, e.g., $\rho = \frac{\sqrt{2}}{2}$ or $\rho = \frac{3}{2}$, due to the fact that non-quadrilateral triangles have simple lowest eigenvalue. The eigenvalues as well as the constants corresponding to the eigenfunctions presented in Figs 10 are summarized in the Table 3.

		$\frac{\pi}{3} - \varepsilon$		$\frac{\pi}{3}$		$\frac{\pi}{3} + \varepsilon$	
	$u_{T,i}^{M}$	$\underline{C}_{T,i}^{48}$	$\lambda_{T,i}^{48}$	$\underline{C}_{T,i}^{48}$	$\lambda_{T,i}^{48}$	$\underline{C}_{T,i}^{48}$	$\lambda_{T,i}^{48}$
$\rho = 1$	$u_{T,1}^{48}$	0.2419	17.0951	0.2387	17.5463	0.2537	15.5404
	$u_{T,2}^{48}$	0.2229	20.1216	0.2387	17.5463	0.2355	18.0309
	$u_{T,3}^{48}$	0.1353	54.6024	0.1378	52.6396	0.1422	49.4818
$\rho = \frac{\sqrt{2}}{2}$	$u_{T,1}^{48}$	0.23137	18.6804	0.23671	17.8471	0.24336	16.8850
	$u_{T,2}^{48}$	0.17082	34.2707	0.17435	32.8970	0.17642	32.1295
	$u_{T,3}^{48}$	0.1229	66.2058	0.12789	61.1402	0.13298	56.5493
$\rho = \frac{3}{2}$	$u_{T,1}^{48}$	0.34714	8.2983	0.35523	7.9247	0.3648	7.5143
	$u_{T,2}^{48}$	0.24485	16.6801	0.24885	16.1482	0.25125	15.8412
	$u_{T,3}^{48}$	0.18258	29.9981	0.19084	27.4575	0.19845	25.3921

Table 3: $\underline{C}_T^{M,\mathrm{P}}$ and λ_T^M corresponding to the first three eigenfunctions in Fig. 10.

4 Two-sided bounds of C_Γ^{P} and C_Γ^{Tr} for tetrahedrons

A nondegenerate tetrahedron $T \in \mathbb{R}^3$ can be presented in the form

$$T = \mathrm{conv}\big\{(0,0,0),(h_1,0,0),(0,0,h_3),(D_{x_1},D_{x_2},D_{x_3})\big\}, \tag{32}$$

where $(D_{x_1},D_{x_2},D_{x_3}) = (h_1\rho\cos\alpha\,\sin\theta,\ h_1\rho\sin\alpha\,\sin\theta,\ h_1\rho\cos(\theta))$, h_1 and h_3 are the scaling parameters along axises O_{x_1} and O_{x_3}, respectively, α is a polar angle, and θ is an azimuthal angle (see Fig. 2). Let zero mean condition be imposed on

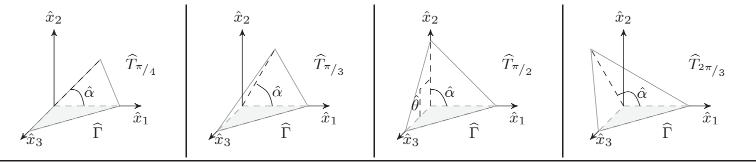
$$\Gamma = \mathrm{conv}\big\{(0,0,0),(h_1,0,0),(0,0,h_3)\big\},$$

and $\widehat{T}_{\hat\theta,\hat\alpha}$ denote the reference tetrahedron, where $\hat\theta$ and $\hat\alpha$ are fixed angles. Then, by $\mathcal{F}_{\hat\theta,\hat\alpha}$ we denote the respective mapping $\mathcal{F}_{\hat\theta,\hat\alpha} : \widehat{T}_{\hat\theta,\hat\alpha} \to T$.

It is possible that these results could be generalized to the other spectral problems that authors consider. To the best of our knowledge, exact values of constants in Poincaré-type inequalities for simplexes in $\mathbb{R}^3$ are unknown. Therefore, we first consider several reference tetrahedrons with $\rho = 1$, $\hat\theta = \frac{\pi}{2}$, and $\hat\alpha_1 = \frac{\pi}{4}$, $\hat\alpha_2 = \frac{\pi}{3}$, $\hat\alpha_3 = \frac{\pi}{2}$, and $\hat\alpha_4 = \frac{2\pi}{3}$, and find the constants numerically with high accuracy. Table 4 shows convergence of the constants with respect to increasing $M(N)$. Then, for an arbitrary tetrahedron T, we have

$$\begin{aligned}
&\|v\|_T \le \widetilde{C}_\Gamma^{\mathrm{P}}\,h_1\,h_3\,\|\nabla v\|_T, \quad \widetilde{C}_\Gamma^{\mathrm{P}} = \min_{\hat\alpha=\{\pi/4,\pi/3,\pi/2,2\pi/3\}}\Big\{c^{\mathrm{P}}_{\pi/2,\hat\alpha}\,C^{\mathrm{P}}_{\widehat\Gamma,\pi/2,\hat\alpha}\Big\},\\
&\|v\|_\Gamma \le \widetilde{C}_\Gamma^{\mathrm{Tr}}\,(h_1\,h_3)^{\frac12}\,\|\nabla v\|_T, \quad \widetilde{C}_\Gamma^{\mathrm{Tr}} = \min_{\hat\alpha=\{\pi/4,\pi/3,\pi/2,2\pi/3\}}\Big\{c^{\mathrm{Tr}}_{\pi/2,\hat\alpha}\,C^{\mathrm{Tr}}_{\widehat\Gamma,\pi/2,\hat\alpha}\Big\},
\end{aligned} \tag{33}$$

	$\hat\theta=\frac{\pi}{2},\hat\alpha=\frac{\pi}{4}$		$\hat\theta=\frac{\pi}{2},\hat\alpha=\frac{\pi}{3}$		$\hat\theta=\frac{\pi}{2},\hat\alpha=\frac{\pi}{2}$		$\hat\theta=\frac{\pi}{2},\hat\alpha=\frac{2\pi}{3}$	



$M(N)$	$C^{\mathrm{p},M}_{\widehat\Gamma,\pi/2,\hat\alpha}$	$C^{\mathrm{Tr},M}_{\widehat\Gamma,\pi/2,\hat\alpha}$	$C^{\mathrm{p},M}_{\widehat\Gamma,\pi/2,\hat\alpha}$	$C^{\mathrm{Tr},M}_{\widehat\Gamma,\pi/2,\hat\alpha}$	$C^{\mathrm{p},M}_{\widehat\Gamma,\pi/2,\hat\alpha}$	$C^{\mathrm{Tr},M}_{\widehat\Gamma,\pi/2,\hat\alpha}$	$C^{\mathrm{p},M}_{\widehat\Gamma,\pi/2,\hat\alpha}$	$C^{\mathrm{Tr},M}_{\widehat\Gamma,\pi/2,\hat\alpha}$
7	0.32431	0.760099	0.325985	0.654654	0.360532	0.654654	0.4152099	0.686161
26	0.338539	0.829445	0.340267	0.761278	0.373669	0.751615	0.4274757	0.863324
63	0.341122	0.831325	0.342556	0.762901	0.375590	0.751994	0.4286444	0.864595
124	0.341147	0.831335	0.342589	0.762905	0.375603	0.751999	0.4286652	0.864630
215	**0.341147**	**0.831335**	**0.342589**	**0.762905**	**0.375603**	**0.751999**	**0.4286652**	**0.864630**

Table 4: $C^{\mathrm{p},M}_{\widehat\Gamma,\pi/2,\hat\alpha}$ and $C^{\mathrm{Tr},M}_{\widehat\Gamma,\pi/2,\hat\alpha}$ with respect to $M(N)$ for $\widehat T_{\hat\theta,\hat\alpha}$ with $\rho=1$, $\hat\theta=\frac{\pi}{2}$, and several $\hat\alpha$.

where $C^{\mathrm{p}}_{\widehat\Gamma,\pi/2,\hat\alpha}$ and $C^{\mathrm{Tr}}_{\widehat\Gamma,\pi/2,\hat\alpha}$ are the constants related to four reference thetrahedron from Table 4 and $c^{\mathrm{p}}_{\pi/2,\hat\alpha}$ and $c^{\mathrm{Tr}}_{\pi/2,\hat\alpha}$ are the ratios of the mapping $\mathcal{F}_{\pi/2,\hat\alpha}\colon \widehat T_{\pi/2,\hat\alpha}\to T$. Here, $\widehat T_{\pi/2,\hat\alpha} := \mathrm{conv}\big\{(0,0,0),(1,0,0),(0,0,1),(\cos\hat\alpha,\sin\hat\alpha,0)\big\}$ with $\hat\alpha=\{\frac{\pi}{4},\frac{\pi}{3},\frac{\pi}{2},\frac{2\pi}{3}\}$, T is defined in (32), and $\mathcal{F}_{\pi/2,\hat\alpha}(\hat x)$ is presented by the relation

$$x=\mathcal{F}_{\pi/2,\hat\alpha}(\hat x)=B_{\pi/2,\hat\alpha}\hat x,\quad B_{\pi/2,\hat\alpha}=\begin{pmatrix} h_1 & \frac{h_1}{\sin\hat\alpha}(\rho\cos\alpha\sin\theta-\cos\hat\alpha) & 0\\ 0 & h_1\rho\frac{\sin\alpha\sin(\theta)}{\sin\hat\alpha} & 0\\ 0 & h_1\rho\frac{\cos\theta}{\sin\hat\alpha} & h_3\end{pmatrix}. \tag{34}$$

The mapping ratios of (34) depend on the maximum eigenvalue of the matrix

$$A_{\pi/2,\hat\alpha} := \begin{pmatrix} h_1^2+b_{12}^2 & b_{12}b_{22} & b_{12}b_{32}\\ b_{12}b_{22} & b_{22}^2 & b_{22}b_{32}\\ b_{12}b_{32} & b_{22}b_{32} & h_3^2+b_{32}^2\end{pmatrix}$$

$$= h_1^2\cdot\begin{pmatrix} 1+\frac{1}{\sin^2\hat\alpha}\,(\rho\cos\alpha\sin\theta-\cos\hat\alpha)^2 & \frac{\rho\sin\alpha\sin\theta}{\sin^2\hat\alpha}(\rho\cos\alpha\sin\theta-\cos\hat\alpha) & \frac{\rho\cos\theta}{\sin^2\hat\alpha}(\rho\cos\alpha\sin\theta-\cos\hat\alpha)\\ \frac{\rho\sin\alpha\sin\theta}{\sin^2\hat\alpha}(\rho\cos\alpha\sin\theta-\cos\hat\alpha) & \frac{\rho^2\,\sin^2\alpha\,\cos^2\theta}{\sin^2\hat\alpha} & \frac{\rho^2\sin\alpha\sin2\theta}{2\sin^2\hat\alpha}\\ \frac{\rho\cos\theta}{\sin^2\hat\alpha}(\rho\cos\alpha\sin\theta-\cos\hat\alpha) & \frac{\rho^2\sin\alpha\sin2\theta}{2\sin^2\hat\alpha} & \frac{h_3^2}{h_1^2}+\rho^2\frac{\cos^2\theta}{\sin^2\hat\alpha}\end{pmatrix},$$

where b_{12}, b_{22}, b_{32} are elements of $B_{\pi/2,\hat\alpha}$ in (34). Accordingly, $\lambda_{\max}(A_{\pi/2,\hat\alpha})$ is defined by the relation

$$\lambda_{\max}(A_{\pi/2,\hat\alpha})=\mu_{\pi/2,\hat\alpha}=\mathcal{E}_4^{1/3}-\mathcal{E}_2\mathcal{E}_4^{-1/3}+\tfrac{1}{2}\mathcal{E}_1,$$

where

$$\begin{aligned}
\mathcal{E}_1 &= b_{12}^2+b_{22}^2+b_{32}^2+h_1^2+h_3^2,\\
\mathcal{E}_2 &= \tfrac{1}{3}\Big(h_3^2\big(b_{12}^2+b_{22}^2\big)+h_1^2\big(b_{22}^2+b_{32}^2\big)-\tfrac{1}{3}\mathcal{E}_1^2+h_1^2h_3^2\Big),\\
\mathcal{E}_3 &= \tfrac{1}{27}\mathcal{E}_1^3-\tfrac{1}{6}\mathcal{E}_1\Big(h_3^2\big(b_{12}^2+b_{22}^2\big)+h_1^2\big(b_{22}^2+b_{32}^2\big)+h_1^2\,h_3^2\Big)+\tfrac{1}{2}b_{22}^2\,h_1^2\,h_3^2,\\
\mathcal{E}_4 &= \mathcal{E}_3+\Big(\Big(\tfrac{h_3^2}{3}\big(b_{12}^2+b_{22}^2\big)+\tfrac{h_1^2}{3}\big(b_{22}^2+b_{32}^2\big)-\tfrac{1}{9}\mathcal{E}_1^2+\tfrac{1}{3}h_1^2h_3^2\Big)^3+\mathcal{E}_3^2\Big)^{1/2}.
\end{aligned}$$

Therefore, $c^{\mathrm{p}}_{\pi/2,\hat\alpha}$ and $c^{\mathrm{Tr}}_{\pi/2,\hat\alpha}$ in (33) reads as follows

$$c^{\mathrm{p}}_{\pi/2,\hat\alpha}=\frac{\mu^{1/2}_{\pi/2,\hat\alpha}}{h_1h_3},\quad c^{\mathrm{Tr}}_{\pi/2,\hat\alpha}=\Big(\frac{h_3\sin\hat\alpha}{\rho\sin\alpha\sin\theta}\Big)^{1/2}c^{\mathrm{p}}_{\pi/2,\hat\alpha}.$$

Lower bounds of the constants C^{p}_{Γ} and C^{Tr}_{Γ} are computed by minimization of $\mathcal{R}^{\mathrm{p}}_{\Gamma}[w]$ and $\mathcal{R}^{\mathrm{Tr}}_{\Gamma}[w]$ over the set $V^N_3\subset H^1(T)$, where

$$V^N_3 := \Big\{\,\varphi_{ijk}=x^iy^jz^k,\quad i,j,k=0,\dots,N,\ \ (i,j,k)\neq(0,0,0)\,\Big\},\qquad \mathrm{dim}V^N_3=M(N):=(N+1)^3-1.$$

The respective results are presented in Tables 5 and 6 for T with $h_1 = 1$, $h_3 = 1$, and $\rho = 1$. We note that exact values of constants are probably closer to the numbers presented in left-hand side columns. For a fixed angle $\theta = \pi/2$, we also present estimates of $\underline{C}_\Gamma^{M,\mathrm{P}}$ and $\underline{C}_\Gamma^{M,\mathrm{Tr}}$ graphically in Fig. 11.

	$\alpha = \frac{\pi}{6}$		$\alpha = \frac{\pi}{4}$		$\alpha = \frac{\pi}{3}$		$\alpha = \frac{\pi}{2}$	
θ	$\underline{C}_\Gamma^{M,\mathrm{P}}$	$\widetilde{C}_\Gamma^{\mathrm{P}}$	$\underline{C}_\Gamma^{M,\mathrm{P}}$	$\widetilde{C}_\Gamma^{\mathrm{P}}$	$\underline{C}_\Gamma^{M,\mathrm{P}}$	$\widetilde{C}_\Gamma^{\mathrm{P}}$	$\underline{C}_\Gamma^{M,\mathrm{P}}$	$\widetilde{C}_\Gamma^{\mathrm{P}}$
$\pi/6$	0.23883	0.49035	0.24621	0.49841	0.25870	0.51054	0.29484	0.51308
$\pi/4$	0.23883	0.45388	0.24621	0.46173	0.25870	0.47683	0.29484	0.49075
$\pi/3$	0.29666	0.41958	0.31194	0.42259	0.33489	0.43724	0.38976	0.46002
$\pi/2$	0.34302	0.35667	0.34112	0.34115	0.34256	0.34259	0.37559	0.37560
$2\pi/3$	0.40428	0.41958	0.40562	0.42259	0.40927	0.43724	0.42867	0.46002
$3\pi/4$	0.42890	0.45388	0.43110	0.46173	0.43505	0.47683	0.45017	0.49075
$5\pi/6$	0.44964	0.49035	0.45193	0.49841	0.45539	0.51054	0.46607	0.51308
	$\alpha = \frac{\pi}{2}$		$\alpha = \frac{2\pi}{3}$		$\alpha = \frac{3\pi}{4}$		$\alpha = \frac{5\pi}{6}$	
θ	$\underline{C}_\Gamma^{M,\mathrm{P}}$	$\widetilde{C}_\Gamma^{\mathrm{P}}$	$\underline{C}_\Gamma^{M,\mathrm{P}}$	$\widetilde{C}_\Gamma^{\mathrm{P}}$	$\underline{C}_\Gamma^{M,\mathrm{P}}$	$\widetilde{C}_\Gamma^{\mathrm{P}}$	$\underline{C}_\Gamma^{M,\mathrm{P}}$	$\widetilde{C}_\Gamma^{\mathrm{P}}$
$\pi/6$	0.29484	0.51308	0.33069	0.51792	0.34468	0.52253	0.35499	0.52694
$\pi/4$	0.29484	0.49075	0.33069	0.50261	0.34468	0.51308	0.35499	0.52253
$\pi/3$	0.38976	0.46002	0.43880	0.48413	0.45742	0.50261	0.47106	0.51792
$\pi/2$	0.37559	0.37560	0.42865	0.42867	0.45017	0.45731	0.46607	0.47811
$2\pi/3$	0.42867	0.46002	0.45997	0.48413	0.47457	0.50261	0.48598	0.51792
$3\pi/4$	0.45017	0.49075	0.47204	0.50261	0.48239	0.51308	0.49064	0.52253
$5\pi/6$	0.46607	0.51308	0.47972	0.51792	0.48607	0.52253	0.49115	0.52694

Table 5: $\underline{C}_\Gamma^{M,\mathrm{P}}(M(N) = 124)$ and $\widetilde{C}_\Gamma^{\mathrm{P}}$.

	$\alpha = \frac{\pi}{6}$		$\alpha = \frac{\pi}{4}$		$\alpha = \frac{\pi}{3}$		$\alpha = \frac{\pi}{2}$	
θ	$\underline{C}_\Gamma^{M,\mathrm{Tr}}$	$\widetilde{C}_\Gamma^{\mathrm{Tr}}$	$\underline{C}_\Gamma^{M,\mathrm{Tr}}$	$\widetilde{C}_\Gamma^{\mathrm{Tr}}$	$\underline{C}_\Gamma^{M,\mathrm{Tr}}$	$\widetilde{C}_\Gamma^{\mathrm{Tr}}$	$\underline{C}_\Gamma^{M,\mathrm{Tr}}$	$\widetilde{C}_\Gamma^{\mathrm{Tr}}$
$\pi/6$	1.09760	3.78259	0.96245	2.71866	0.91255	2.27382	0.93123	2.05449
$\pi/4$	1.09760	2.43897	0.96245	1.78094	0.91255	1.50166	0.93123	1.38951
$\pi/3$	0.89122	1.74467	0.79146	1.31130	0.75950	1.12431	0.78904	1.06349
$\pi/2$	0.98017	1.22920	0.83132	0.83133	0.76290	0.76291	0.75199	0.75200
$2\pi/3$	1.17698	1.74467	0.99473	1.31130	0.90578	1.12431	0.86463	1.06349
$3\pi/4$	1.35195	2.43897	1.14144	1.78094	1.03737	1.50166	0.98220	1.38951
$5\pi/6$	1.65317	3.78259	1.39424	2.71866	1.26490	2.27382	1.19017	2.05449
	$\alpha = \frac{\pi}{2}$		$\alpha = \frac{2\pi}{3}$		$\alpha = \frac{3\pi}{4}$		$\alpha = \frac{5\pi}{6}$	
θ	$\underline{C}_\Gamma^{M,\mathrm{Tr}}$	$\widetilde{C}_\Gamma^{\mathrm{Tr}}$	$\underline{C}_\Gamma^{M,\mathrm{Tr}}$	$\widetilde{C}_\Gamma^{\mathrm{Tr}}$	$\underline{C}_\Gamma^{M,\mathrm{Tr}}$	$\widetilde{C}_\Gamma^{\mathrm{Tr}}$	$\underline{C}_\Gamma^{M,\mathrm{Tr}}$	$\widetilde{C}_\Gamma^{\mathrm{Tr}}$
$\pi/6$	0.93123	2.05449	1.07244	2.39471	1.21573	2.95902	1.47044	4.21999
$\pi/4$	0.93123	1.38951	1.07244	1.64324	1.21573	2.01841	1.47044	2.80588
$\pi/3$	0.78904	1.06349	0.91773	1.27423	1.04309	1.50833	1.26357	2.11790
$\pi/2$	0.75199	0.75200	0.86459	0.86463	0.98220	1.12971	1.19017	1.67033
$2\pi/3$	0.86463	1.06349	0.96174	1.27423	1.08134	1.50833	1.30191	2.11790
$3\pi/4$	0.98220	1.38951	1.07921	1.64324	1.20686	2.01841	1.44721	2.80588
$5\pi/6$	1.19017	2.05449	1.29582	2.39471	1.44268	2.95902	1.72383	4.21999

Table 6: $\underline{C}_\Gamma^{M,\mathrm{Tr}}$ $(M(N) = 124)$ and $\widetilde{C}_\Gamma^{\mathrm{Tr}}$.

5 Example

Constants in the Friedrichs', Poincaré, and other functional inequalities arise in various problems of numerical analysis, where we need to know values of the respective constants associated with particular domains. For example, results related to extension and projection type estimates for FEM can be found in [24, 7] (and many other publications). Concerning constants in the trace inequalities associated with polygonal domain, we mention the paper [5]. Constants in functional (embedding) inequalities arise in various error estimates. We deduce an advanced version of the estimate (46) in [31], which uses constants in Poincaré-type inequalities for functions with zero mean traces on inter-element boundaries in order to maximally extend the space of admissible fluxes. Below we address the latter case and first explain reasons that invoke the constants in general terms.

Let u denote the exact solution of an elliptic boundary value problem generated by the pair of conjugate operators grad and $-$div (e.g., the problem (38)–(41) considered below) and v be a function in the energy space satisfying the prescribed (Dirichlét) boundary conditions. Typically, the error $e := u - v$ is measured in terms of the energy

norm $\|\nabla e\|$ (or some other equivalent norm), which square is bounded from above by the quantities

$$\int_\Omega R(v, \mathrm{div} q)\, e \ \mathrm{d}x, \quad \int_\Omega D(\nabla v, q) \cdot \nabla e \ \mathrm{d}x, \text{ and } \int_{\Gamma_N} R_{\Gamma_N}(v, q \cdot n)\, e \, ds,$$

where Γ_N is the Neumann part of the boundary ∂T, n is the outward unit normal, and q is an approximation of the dual variable (flux). The terms R, D, and R_{Γ_N} represent residuals of the differential (balance) equation, constitutive (duality) relation, and Neumann boundary condition, respectively. Since v and q are known from a numerical solution, fully computable estimates can be obtained if these integrals are estimated by the Hölder, Friedrichs, and trace inequalities (which involve the corresponding constants). However, for a Lipschitz domain Ω with piecewise smooth (e.g., polynomial) boundaries these constants may be unknown. A way to avoid these difficulties is suggested by modifications of the estimates using ideas of domain decomposition. Assume that Ω is polygonal (polyhedral) domain decomposed into a collection of non-overlapping convex polygonal sub-domains Ω_i, i.e.,

$$\overline{\Omega} := \bigcup_{\Omega_i \in \mathcal{O}_\Omega} \overline{\Omega}_i, \quad \mathcal{O}_\Omega := \Big\{ \ \Omega_i \in \Omega \ \Big| \ \Omega_{i'} \cap \Omega_{i''} = \emptyset, \ i' \neq i'', \ i = 1, \dots, N \ \Big\}.$$

We denote the set of all edges (faces) by $\mathcal{G}$ and the set of all interior faces by $\mathcal{G}_{\mathrm{int}}$ (i.e., $\Gamma_{ij} \in \mathcal{G}_{\mathrm{int}}$, if $\Gamma_{ij} = \overline{\Omega}_i \cap \overline{\Omega}_j$). Analogously, $\mathcal{G}_N$ denotes the set of edges on Γ_N. The latter set is decomposed into $\Gamma_{N_k} := \Gamma_N \cap \partial\Omega_k$ (the number of faces that belongs to Γ_{N_k} is K_N). Now, the integrals associated with R and R_{Γ_N} can be replaced by sums of local quantities

$$\sum_{i=1}^{N} \int_{\Omega_i} R_\Omega(v, \mathrm{div} q)\, e \ \mathrm{d}x, \quad \text{and} \quad \sum_{k=1}^{K_N} \int_{\Gamma_{N_k}} R_{\Gamma_N}(v, q \cdot n)\, e \ \mathrm{d}s.$$

If the residuals satisfy the conditions

$$\int_{\Omega_i} R_{\Omega_i}(v, \ \mathrm{div} q) \, \mathrm{d}x = 0, \quad \forall i = 1, \dots, N,$$

and

$$\int_{\Gamma_{N_k}} R_{\Gamma_N}(v, \ q \cdot n) \, \mathrm{d}s = 0, \quad \forall k = 1, \dots, K_N,$$

then

$$\int_{\Omega_i} R_\Omega(v, \mathrm{div} q)\, e \ \mathrm{d}x \le C^{\mathrm{P}}_{\Omega_i} \| R_{\Omega_i}(v, \mathrm{div} q) \|_{\Omega_i} \, \| \nabla e \|_{\Omega_i}, \tag{35}$$

and

$$\int_{\Gamma_{N_k}} R_{\Gamma_N}(v, q \cdot n)\, e \ \mathrm{d}s \le C^{\mathrm{Tr}}_{\Gamma_{N_k}} \| R_{\Gamma_N}(v, q \cdot n) \|_{\Gamma_{N_k}} \| \nabla e \|_{\Omega_k}. \tag{36}$$

Hence, we can deduce a computable upper bound of the error that contains local constants $C^{\mathrm{P}}_{\Omega_i}$ and $C^{\mathrm{Tr}}_{\Gamma_{N_k}}$ for simple subdomains (e.g., triangles or tetrahedrons) instead of the global constants associated with Ω.

The constant C^{P}_{Ω} may arise if, e.g., nonconforming approximations are used. For example, if v does not exactly satisfy the Dirichlét boundary condition on Γ_{D_k}, then in the process of estimation it may be necessary to evaluate terms of the type

$$\int_{\Gamma_{D_k}} G_D(v)\, e \ \mathrm{d}s, \quad k = 1, \dots, K_D,$$

where Γ_{D_k} is a part of Γ_D associated with a certain Ω_k, and $G_D(v)$ is a residual generated by inexact satisfaction of the boundary condition. If we impose the requirement that the Dirichlét boundary condition is satisfied in a weak sense, i.e., $\{|G_D(v)|\}_{\Gamma_{D_k}} = 0$, then each boundary integral can be estimated as follows:

$$\int_{\Gamma_{D_k}} G_D(v)\, e \ \mathrm{d}s \le C^{\mathrm{P}}_{\Gamma_{D_k}} \| G_D(v) \|_{\Gamma_{D_k}} \| \nabla e \|_{\Omega_k}. \tag{37}$$

After summing (35), (36), and (37), we obtain a product of weighted norms of localized residuals (which are known) and $\|\nabla e\|_\Omega$. Since the sum is bounded from below by the squared energy norm, we arrive at computable error majorant.

Now, we discuss elaborately these questions with the paradigm of the following boundary value problem: find u such that

$$-\mathrm{div} p + \varrho^2 u = f, \qquad \text{in } \Omega, \tag{38}$$

$$p = A\nabla u, \qquad \text{in } \Omega, \tag{39}$$

$$u = u_D, \qquad \text{on } \Gamma_D, \tag{40}$$

$$A\nabla u \cdot n = F \qquad \text{on } \Gamma_N. \tag{41}$$

Here $f \in L^2(\Omega)$, $F \in L^2(\Gamma_N)$, $u_D \in H^1(\Omega)$, and A is a symmetric positive definite matrix with bounded coefficients satisfying the condition $\lambda_1|\xi|^2 \le A\xi \cdot \xi$, where λ_1 is a positive constant independent of ξ. The generalized solution of (38)–(41) exists and is unique in the set $V_0 + u_D$, where $V_0 := \Big\{ w \in H^1(\Omega) \ \big| \ w = 0 \text{ on } \Gamma_D \Big\}$.

Assume that $v \in V_0 + u_D$ is a conforming approximation of u. We wish to find a computable majorant of the error norm

$$\| e \|^2 := \|\nabla e\|_A^2 + \|\varrho\, e\|^2, \tag{42}$$

where $\|\nabla e\|_A^2 := \int\limits_\Omega A\nabla e \cdot \nabla e \,\mathrm{d}x$. First, we note that the integral identity that defines u can be rewritten in the form

$$\int\limits_\Omega A\nabla e \cdot \nabla w \ \mathrm{d}x + \int\limits_\Omega \varrho^2 e\, w \ \mathrm{d}x = \int\limits_\Omega (f w - \varrho^2 v\, w - A\nabla v \cdot \nabla w) \,\mathrm{d}x + \int\limits_{\Gamma_N} F w \,\mathrm{d}s, \quad \forall w \in V_0. \tag{43}$$

It is well known (see [31, Section 4.2]) that this relation yields computable majorant of $\| e \|^2$, if we introduce a vector valued function $q \in H(\Omega, \mathrm{div})$ and transform (43) by means of integration by parts relations. The majorant has the form

$$\| e \| \le \ \|D_\Omega(\nabla v, q)\|_{A^{-1}} + C_1 \|R(v, \mathrm{div} q)\|_\Omega + C_2 \|R_{\Gamma_N}(v, q\cdot n)\|_{\Gamma_N}, \tag{44}$$

where C_1 and C_2 are positive constants explicitly defined by λ_1, the Friedrichs' inequality C_Ω^{F} in $\|v\|_\Omega \le C_\Omega^{\mathrm{F}} \|\nabla v\|_\Omega$ for functions vanishing on Γ_D, and constant $C_{\Gamma_N}^{\mathrm{Tr}}$ in the trace inequality associated with Γ_N. The integrands are defined by the relations

$$D(\nabla v, q) := A\nabla v - q, \quad R(v, \mathrm{div} q) := \mathrm{div} q + f - \varrho^2 v, \quad \text{and} \quad R_{\Gamma_N}(v, q\cdot n) := q \cdot n - F.$$

In general, finding C_Ω^{F} and $C_{\Gamma_N}^{\mathrm{Tr}}$ may be not an easy task. We can exclude C_2 if q additionally satisfies the condition $q \cdot n = F$. Then, the last term in (44) vanishes. However, this condition is difficult to satisfy, if F is a complicated nonlinear function. In order to exclude C_1, we can apply domain decomposition and use (35) instead of the global estimate. Then, the estimate will operate with the constants $C_{\Omega_i}^{\mathrm{P}}$ (which upper bounds are known for convex domains). Moreover, it is shown below that using the inequalities (2) and (4), we can essentially weaken the assumptions required for the variable q.

Define the space of vector valued functions

$$\begin{aligned} \hat{H}(\Omega, \mathcal{O}_\Omega, \mathrm{div}) := \Big\{ q \in L^2(\Omega, \mathbb{R}^d) \ \big| \ & q = q_i \in H(\Omega_i, \mathrm{div}), \\ & \big\{ \big| \mathrm{div} q_i + f - \varrho^2\, v \big| \big\}_{\Omega_i} = 0, \quad \forall\, \Omega_i \in \mathcal{O}_\Omega, \\ & \big\{ | (q_i - q_j)\cdot n_{ij} | \big\}_{\Gamma_{ij}} = 0, \quad \forall\, \Gamma_{ij} \in \mathcal{G}_{\mathrm{int}}, \\ & \big\{ | q_i \cdot n_k - F | \big\}_{\Gamma_{N_k}} = 0, \quad \forall\, k = 1, \dots, K_N \Big\}. \end{aligned}$$

We note that the space $\hat{H}(\Omega, \mathcal{O}_\Omega, \mathrm{div})$ is wider then $H(\Omega, \mathrm{div})$ (so that we have more flexibility in determination of optimal reconstruction of numerical fluxes). Indeed, the vector valued functions in $H(\Omega, \mathrm{div})$ must have continuous normal components on all $\Gamma_{ij} \in \mathcal{G}_{\mathrm{int}}$ and satisfy the Neumann boundary condition in the pointwise sense. The functions in $\hat{H}(\Omega, \mathcal{O}_\Omega, \mathrm{div})$ satisfy much weaker conditions: namely, the normal components are continuous only in terms of mean values (integrals) and the Neumann condition must hold in the integral sense only.

We reform (43) by means of the integral identity

$$\sum_{\Omega_i \in \mathcal{O}_\Omega} \int_{\Omega_i} (q \cdot \nabla w + \mathrm{div} q\, w) \ \mathrm{d}x = \sum_{\Gamma_{ij} \in \mathcal{G}_{\mathrm{int}}} \int\limits_{\Gamma_{ij}} (q_i - q_j)\cdot n_{ij}\, w \,\mathrm{d}s + \sum_{\Gamma_{N_k} \in \Gamma_N} \int\limits_{\Gamma_{N_k}} q_i \cdot n_i \, w \,\mathrm{d}s,$$

which holds for any $w \in V_0$ and $q \in \hat{H}(\Omega, \mathcal{O}_\Omega, \mathrm{div})$. Setting $w = e$ in (43) and applying the Hölder inequality, we find that

$$\begin{aligned}\mathbb{|\!|\!|} e \mathbb{|\!|\!|}^2 \le \|D(\nabla v, q)\|_{A^{-1}} \|\nabla e\|_A + \sum_{\Omega_i \in \mathcal{O}_\Omega} \|R(v, \mathrm{div} q)\|_{\Omega_i} \Big\| e - \{e\}_{\Omega_i} \Big\|_{\Omega_i} \\ + \sum_{\Gamma_{ij} \in \mathcal{G}_{\mathrm{int}}} r_{ij}(q) \Big\| e - \{e\}_{\Gamma_{ij}} \Big\|_{\Gamma_{ij}} + \sum_{\Gamma_{N_k} \in \Gamma_N} \rho_k(q) \Big\| e - \{e\}_{\Gamma_{N_k}} \Big\|_{\Gamma_{N_k}},\end{aligned}$$

where

$$r_{ij}(q) := \|(q_i - q_j) \cdot n_{ij}\|_{\Gamma_{ij}}, \quad \rho_k(q) := \|q_k \cdot n_k - F\|_{\Gamma_{N_k}}.$$

In view of (1) and (4), we obtain

$$\begin{aligned}\mathbb{|\!|\!|} e \mathbb{|\!|\!|}^2 \le \|D(\nabla v, q)\|_{A^{-1}} \|\nabla e\|_A + \sum_{\Omega_i \in \mathcal{O}_\Omega} \|R(v, \mathrm{div} q)\|_{\Omega_i} C^{\mathrm{P}}_{\Omega_i} \|\nabla e\|_{\Omega_i} \\ + \sum_{\Gamma_{ij} \in \mathcal{G}_{\mathrm{int}}} r_{ij}(q) C^{\mathrm{Tr}}_{\Gamma_{ij}} \|\nabla e\|_{\Omega_i} + \sum_{\Gamma_{N_k} \in \Gamma_N} \rho_k(q) C^{\mathrm{Tr}}_{\Gamma_{N_k}} \|\nabla e\|_{\Omega_i}. \quad (45)\end{aligned}$$

The second term in the right hand side is estimated by the quantity $\mathfrak{R}_1(v, q)\, \|\nabla e\|_\Omega$, where

$$\mathfrak{R}_1^2(v, q) := \sum_{\Omega_i \in \mathcal{O}_\Omega} \tfrac{(\mathrm{diam}\, \Omega_i)^2}{\pi^2} \|R(v, \mathrm{div} q)\|^2_{\Omega_i}.$$

We can represent any $\Omega_i \in \mathcal{O}_\Omega$ as a sum of simplexes such that each simplex has one edge on $\partial\Omega_i$. Let $C^{\mathrm{Tr}}_{i,\max}$ denote the largest constant in the respective Poincaré-type inequalities (4) associated with all edges of $\partial\Omega_i$. Then, the last two terms of (45) can be estimated by the quantity $\mathfrak{R}_2(v, q)\, \|\nabla e\|_\Omega$, where

$$\mathfrak{R}_2^2(q) := \sum_{\Omega_i \in \mathcal{O}_\Omega} (C^{\mathrm{Tr}}_{i,\max})^2\, \eta_i^2, \quad \text{with} \quad \eta_i^2 = \sum_{\substack{\Gamma_{ij} \in \mathcal{G}_{\mathrm{int}}, \\ \Gamma_{ij} \cap \partial\Omega_i \neq \varnothing}} \tfrac{1}{4} r_{ij}^2(q) + \sum_{\substack{\Gamma_k \in \mathcal{G}_N, \\ \Gamma_k \cap \partial\Omega_i \neq \varnothing}} \rho_k^2(q).$$

Then, (45) yields the estimate

$$\mathbb{|\!|\!|} e \mathbb{|\!|\!|}^2 \le \|D(\nabla v, q)\|_{A^{-1}} \|\nabla e\|_A + (\mathfrak{R}_1(v, q) + \mathfrak{R}_2(q))\, \|\nabla e\|_\Omega,$$

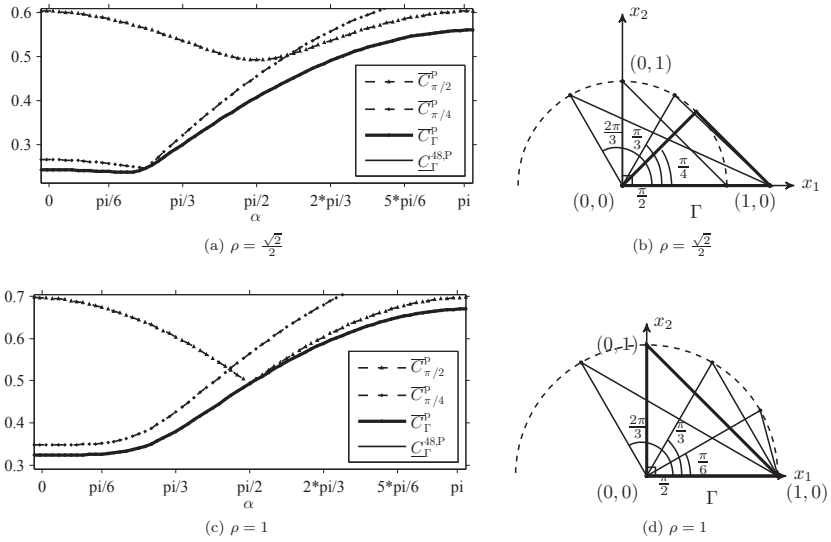
which shows that

$$\mathbb{|\!|\!|} e \mathbb{|\!|\!|} \le \|D(\nabla v, q)\|_{A^{-1}} + \tfrac{1}{\lambda_1} \Big(\mathfrak{R}_1(v, q) + \mathfrak{R}_2(q) \Big). \quad (46)$$

Here, the term $\mathfrak{R}_2(q)$ controls violations of conformity of q (on interior edges) and inexact satisfaction of boundary conditions (on edges related to Γ_N). It is easy to see that $\mathfrak{R}_2(q) = 0$, if and only if $q \cdot n$ is continuous on $\mathcal{G}_{\mathrm{int}}$ and exactly satisfy the boundary condition. Hence, it can be viewed as a measure of the "flux nonconformity". Other terms have the same meaning as in known a posteriori estimates of the functional type, namely, the first term measures the violation in the relation $q = A\nabla v$ (cf. (39)), and $\mathfrak{R}_1(v, q)$ measures inaccuracy in the equilibrium (balance) equation (38). The right-hand side of (46) contains known functions (approximation v and the reconstruction of the flux q) and constants that can be easily computed using results of Section 2-4. Finally, we note that estimates similar to (46) were derived in [32] for elliptic variational inequalities and in [22] for a class of parabolic problems.

6 Appendix

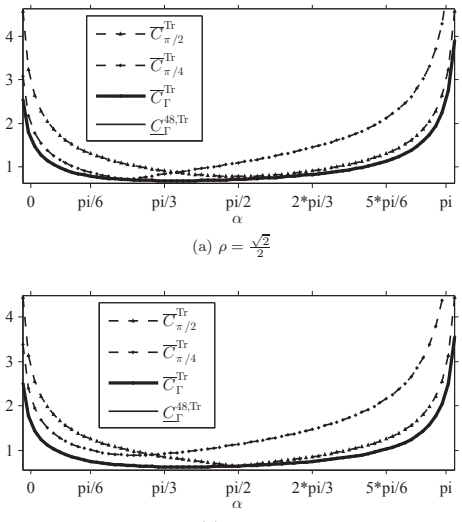
For convenience, we collect in this part the graphics cited in Section 3 and 4.



(a) $\rho = \frac{\sqrt{2}}{2}$ (b) $\rho = \frac{\sqrt{2}}{2}$

(c) $\rho = 1$ (d) $\rho = 1$

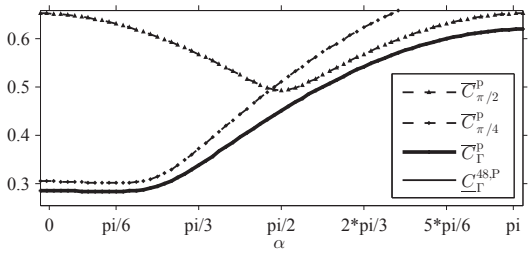
Figure 3: Lower and upper bounds of C^{p}_{Γ} for $T \in \mathbb{R}^2$ (a)-(b) $\rho = \frac{\sqrt{2}}{2}$ and (c)-(d) $\rho = 1$.



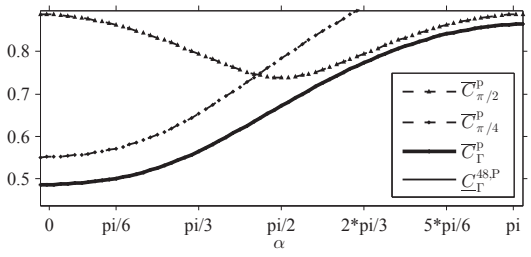
(a) $\rho = \frac{\sqrt{2}}{2}$

(b) $\rho = 1$

Figure 4: Lower and upper bound of C^{Tr}_{Γ} for $T \in \mathbb{R}^2$ (a) $\rho = \frac{\sqrt{2}}{2}$ and (b) $\rho = 1$.

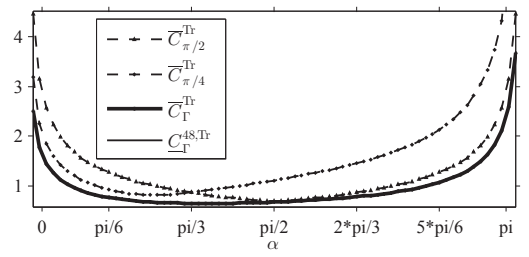


(a) $\rho = \frac{\sqrt{3}}{2}$

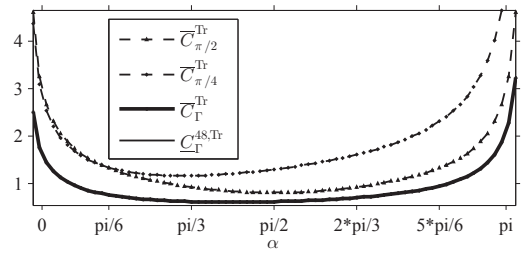


(b) $\rho = \frac{3}{2}$

Figure 5: Lower and upper bounds of C^{p}_{Γ} for $T \in \mathbb{R}^2$ (a) $\rho = \frac{\sqrt{3}}{2}$ and (b) $\rho = \frac{3}{2}$.

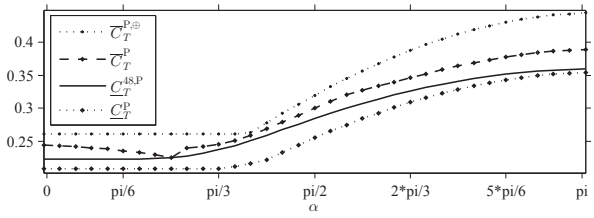


(a) $\rho = \frac{\sqrt{3}}{2}$

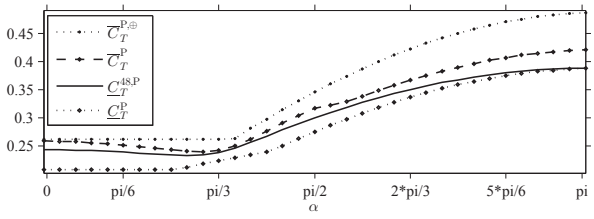


(b) $\rho = \frac{3}{2}$

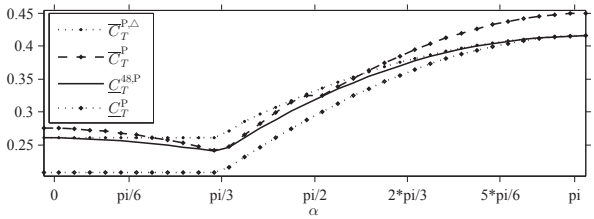
Figure 6: Lower and upper bounds of C^{Tr}_{Γ} for $T \in \mathbb{R}^2$ (a) $\rho = \frac{\sqrt{3}}{2}$ and (b) $\rho = \frac{3}{2}$.



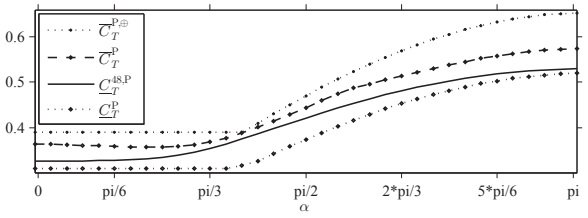
(a) $\rho = \frac{\sqrt{2}}{2}$



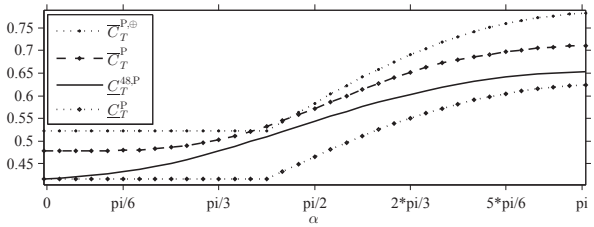
(b) $\rho = \frac{\sqrt{3}}{2}$



(c) $\rho = 1$

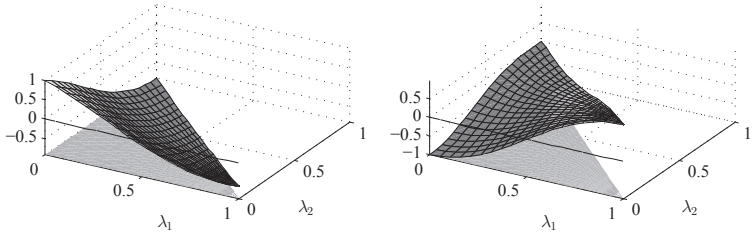


(d) $\rho = \frac{3}{2}$



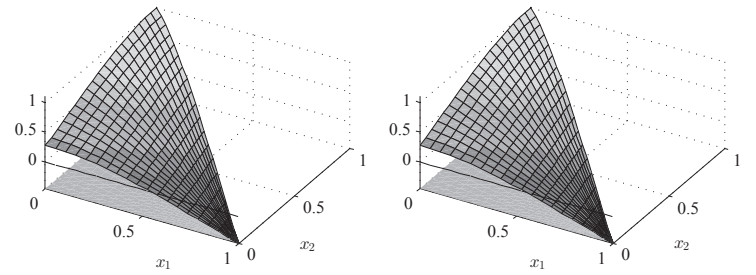
(e) $\rho = 2$

Figure 7: $\underline{C}_T^{48}$ with upper bounds $\overline{C}_T^{\mathrm{P},\triangle}$ and $\overline{C}_T^{\mathrm{P}}$ and lower bound with respect to α on T with (a) $\rho = \frac{\sqrt{2}}{2}$, (b) $\frac{\sqrt{3}}{2}$, (c) 1, (d) $\frac{3}{2}$, and (e) 2.



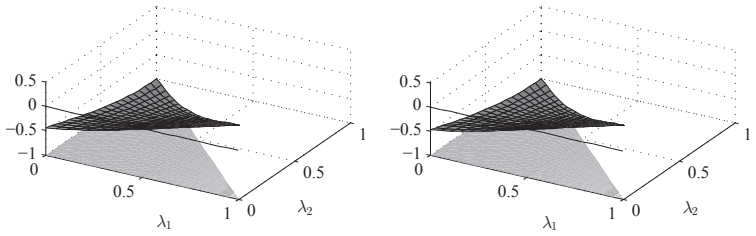
(a) $u_\Gamma^{48,\mathrm{p}}$, $\alpha = \frac{\pi}{6}$

(b) $u_\Gamma^{48,\mathrm{p}}$, $\alpha = \frac{\pi}{3}$



(c) $u_\Gamma^{48,\mathrm{p}}$, $\alpha = \frac{\pi}{2}$

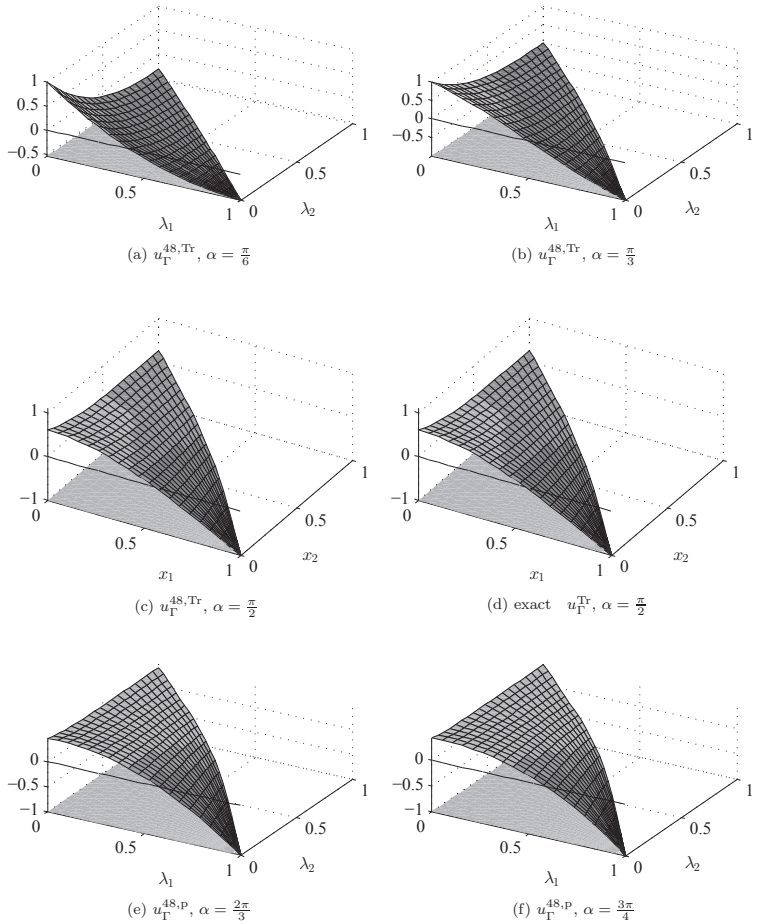
(d) exact u_Γ^{p}, $\alpha = \frac{\pi}{2}$



(e) $u_\Gamma^{48,\mathrm{p}}$, $\alpha = \frac{2\pi}{3}$

(f) $u_\Gamma^{48,\mathrm{p}}$, $\alpha = \frac{3\pi}{4}$

Figure 8: Eigenfunctions corresponding to $\underline{C}_\Gamma^{M,\mathrm{p}}$ and for $M = 48$ on simplex T with $\rho = 1$ and different α.



(a) $u_\Gamma^{48,\mathrm{Tr}}$, $\alpha = \frac{\pi}{6}$

(b) $u_\Gamma^{48,\mathrm{Tr}}$, $\alpha = \frac{\pi}{3}$

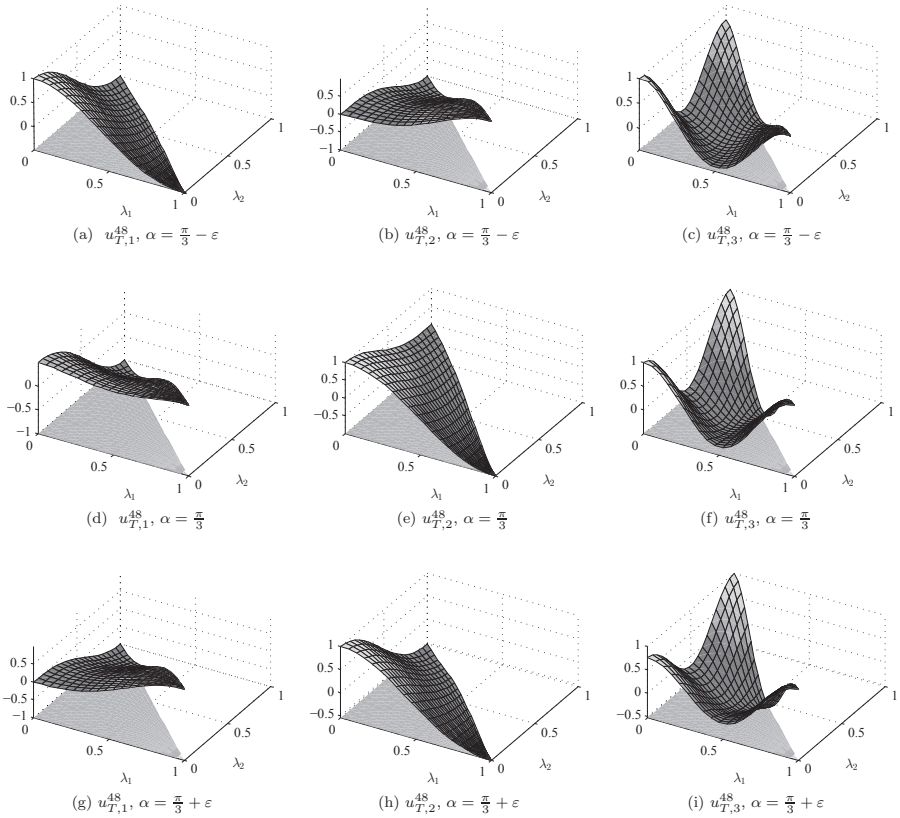
(c) $u_\Gamma^{48,\mathrm{Tr}}$, $\alpha = \frac{\pi}{2}$

(d) exact u_Γ^{Tr}, $\alpha = \frac{\pi}{2}$

(e) $u_\Gamma^{48,\mathrm{p}}$, $\alpha = \frac{2\pi}{3}$

(f) $u_\Gamma^{48,\mathrm{p}}$, $\alpha = \frac{3\pi}{4}$

Figure 9: Eigenfunctions corresponding to $\underline{C}_\Gamma^{M,\mathrm{Tr}}$ for $M = 48$ on simplex T with $\rho = 1$ and different α.



(a) $u^{48}_{T,1}$, $\alpha = \frac{\pi}{3} - \varepsilon$ (b) $u^{48}_{T,2}$, $\alpha = \frac{\pi}{3} - \varepsilon$ (c) $u^{48}_{T,3}$, $\alpha = \frac{\pi}{3} - \varepsilon$

(d) $u^{48}_{T,1}$, $\alpha = \frac{\pi}{3}$ (e) $u^{48}_{T,2}$, $\alpha = \frac{\pi}{3}$ (f) $u^{48}_{T,3}$, $\alpha = \frac{\pi}{3}$

(g) $u^{48}_{T,1}$, $\alpha = \frac{\pi}{3} + \varepsilon$ (h) $u^{48}_{T,2}$, $\alpha = \frac{\pi}{3} + \varepsilon$ (i) $u^{48}_{T,3}$, $\alpha = \frac{\pi}{3} + \varepsilon$

Figure 10: Eigenfunctions corresponding to $\underline{C}_T^{M,\mathrm{P}}$ with $M = 48$ on isosceles triangles $T \in \mathbb{R}^2$ with $\alpha = \frac{\pi}{3}$, $\frac{\pi}{3} - \varepsilon$, and $\frac{\pi}{3} + \varepsilon$ in barycentric coordinates.

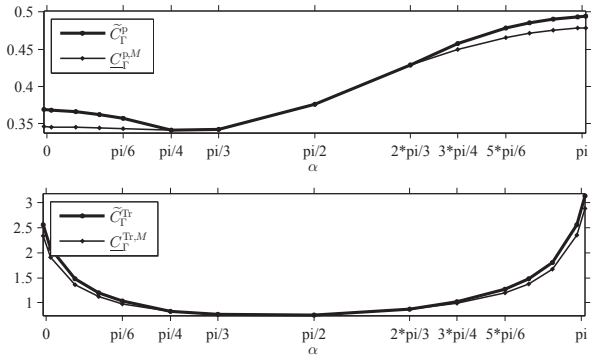


Figure 11: C_Γ^{p} and C_Γ^{Tr} for $T \in \mathbb{R}^3$ with $H = 1$, $\rho = 1$ with estimate based on four reference tetrahedrons.

References

[1] R. Bañuelos and K. Burdzy. On the "hot spots" conjecture of J. Rauch. *J. Funct. Anal.*, 164(1):1–33, 1999.

[2] R. Bañuelos, T. Kulczycki, I. Polterovich, and B. Siudeja. Eigenvalue inequalities for mixed Steklov problems. In *Operator theory and its applications*, volume 231 of *Amer. Math. Soc. Transl. Ser. 2*, pages 19–34. Amer. Math. Soc., Providence, RI, 2010.

[3] P. H. Bérard. Spectres et groupes cristallographiques. I. Domaines euclidiens. *Invent. Math.*, 58(2):179–199, 1980.

[4] C. Carstensen and J. Gedicke. Guaranteed lower bounds for eigenvalues. *Math. Comp.*, 83(290):2605–2629, 2014.

[5] C. Carstensen and S. A. Sauter. A posteriori error analysis for elliptic PDEs on domains with complicated structures. *Numer. Math.*, 96(4):691–721, 2004.

[6] S. Y. Cheng. Eigenvalue comparison theorems and its geometric applications. *Math. Z.*, 143(3):289–297, 1975.

[7] P. G. Ciarlet. *The finite element method for elliptic problems*. North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978. Studies in Mathematics and its Applications, Vol. 4.

[8] C. R. Dohrmann, A. Klawonn, and O. B. Widlund. Domain decomposition for less regular subdomains: overlapping Schwarz in two dimensions. *SIAM J. Numer. Anal.*, 46(4):2153–2168, 2008.

[9] D. W. Fox and J. R. Kuttler. Sloshing frequencies. *Z. Angew. Math. Phys.*, 34(5):668–696, 1983.

[10] M. Fuchs. Computable upper bounds for the constants in Poincaré-type inequalities for fields of bounded deformation. *Math. Methods Appl. Sci.*, 34(15):1920–1932, 2011.

[11] A. Girouard and I. Polterovich. Spectral geometry of the steklov problem. *arXiv.org*, math/1411.6567, 2014.

[12] Y. Hoshikawa and H. Urakawa. Affine Weyl groups and the boundary value eigenvalue problems of the Laplacian. *Interdiscip. Inform. Sci.*, 16(1):93–109, 2010.

[13] F. Kikuchi and X. Liu. Estimation of interpolation error constants for the P_0 and P_1 triangular finite elements. *Comput. Methods Appl. Mech. Engrg.*, 196(37-40):3750–3758, 2007.

[14] A. Klawonn, O. Rheinbach, and O. B. Widlund. An analysis of a FETI-DP algorithm on irregular subdomains in the plane. *SIAM J. Numer. Anal.*, 46(5):2484–2504, 2008.

[15] V. Kozlov and N. Kuznetsov. The ice-fishing problem: the fundamental sloshing frequency versus geometry of holes. *Math. Methods Appl. Sci.*, 27(3):289–312, 2004.

[16] V. Kozlov, N. Kuznetsov, and O. Motygin. On the two-dimensional sloshing problem. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, 460(2049):2587–2603, 2004.

[17] N. Kuznetsov, T. Kulczycki, M. Kwaśnicki, A. Nazarov, S. Poborchi, I. Polterovich, and B. Siudeja. The legacy of Vladimir Andreevich Steklov. *Notices Amer. Math. Soc.*, 61(1):9–22, 2014.

[18] R. S. Laugesen and B. A. Siudeja. Maximizing Neumann fundamental tones of triangles. *J. Math. Phys.*, 50(11):112903, 18, 2009.

[19] R. S. Laugesen and B. A. Siudeja. Minimizing Neumann fundamental tones of triangles: an optimal Poincaré inequality. *J. Differential Equations*, 249(1):118–135, 2010.

[20] X. Liu and S. Oishi. Guaranteed high-precision estimation for P_0 interpolation constants on triangular finite elements. *Jpn. J. Ind. Appl. Math.*, 30(3):635–652, 2013.

[21] A. Logg, K.-A. Mardal, and G. N. Wells, editors. *Automated solution of differential equations by the finite element method*, volume 84 of *Lecture Notes in Computational Science and Engineering*. Springer, Heidelberg, 2012. The FEniCS book.

[22] S. Matculevich, P. Neittaanmäki, and S. Repin. A posteriori error estimates for time-dependent reaction-diffusion problems based on the Payne–Weinberger inequality. *AIMS*, 35(6), 2015.

[23] B. J. McCartin. Eigenstructure of the equilateral triangle. II. The Neumann problem. *Math. Probl. Eng.*, 8(6):517–539, 2002.

[24] S. G. Mikhlin. *Constants in some inequalities of analysis*. A Wiley-Interscience Publication. John Wiley and Sons, Ltd., Chichester, 1986. Translated from the Russian by Reinhard Lehmann.

[25] A. I. Nazarov and S. I. Repin. Exact constants in Poincare type inequalities for functions with zero mean boundary traces. *Mathematical Methods in the Applied Sciences*, 2014. Puplished in arXiv.org in 2012, math/1211.2224.

[26] D. Pauly. On Maxwell's and Poincaré's constants. *Discrete Contin. Dyn. Syst. Ser. S*, 8(3):607–618, 2015.

[27] L. E. Payne and H. F. Weinberger. An optimal Poincaré inequality for convex domains. *Arch. Rational Mech. Anal.*, 5:286–292 (1960), 1960.

[28] M. A. Pinsky. The eigenvalues of an equilateral triangle. *SIAM J. Math. Anal.*, 11(5):819–827, 1980.

[29] H. Poincare. Sur les Equations aux Derivees Partielles de la Physique Mathematique. *Amer. J. Math.*, 12(3):211–294, 1890.

[30] H. Poincare. Sur les Equations de la Physique Mathematique. *Rend. Circ. Mat. Palermo*, 8:57–156, 1894.

[31] S. Repin. *A posteriori estimates for partial differential equations*, volume 4 of *Radon Series on Computational and Applied Mathematics*. Walter de Gruyter GmbH & Co. KG, Berlin, 2008.

[32] S. Repin. Estimates of deviations from exact solutions of variational inequalities based upon payne-weinberger inequality. *J. Math. Sci. (N. Y.)*, 157(6):874–884, 2009.

[33] Inc. ©1994-2015 The MathWorks. Mathworks. products and services, 2015.

[34] A. Toselli and O. Widlund. *Domain decomposition methods—algorithms and theory*, volume 34 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2005.