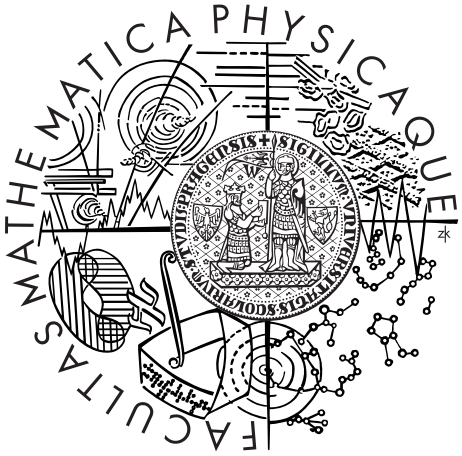
Charles University in Prague
Faculty of Mathematics and Physics

DOCTORAL THESIS



Jan Pech

Numerical modelling of unstable fluid flow past heated bodies

Mathematical Institute of Charles University

Supervisor of the doctoral thesis:	prof. Ing. František Maršík, DrSc.
Study programme:	Physics
Specialization:	Mathematical and Computer Modelling

Prague 2016

Acknowledgements

I would like to express my thanks to Prof. Ing. František Maršík, DrSc., who gave me opportunity to work on this thesis. I would like to thank to the Institute of Thermomechanics for the support in this research and allowing my stay at National Taiwan University. A special thank belongs to my mother, who supported me during whole the time of my studies.

I declare that I carried out this doctoral thesis independently, and only with the cited sources, literature and other professional sources.

I understand that my work relates to the rights and obligations under the Act No. 121/2000 Coll., the Copyright Act, as amended, in particular the fact that the Charles University in Prague has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 paragraph 1 of the Copyright Act.

In ..... date ..... signature of the author

Název práce: Numerické modelování nestabilit při obtékání zahřívaných těles

Autor: Jan Pech

Katedra: Matematický ústav UK

Vedoucí disertační práce: prof. Ing. František Maršík, DrSc., Matematický ústav UK

Abstrakt: Práce přináší nové výsledky z oboru numerických výpočtů proudění ovlivněného změnami teploty. Navržený výpočetní algoritmus zohledňuje proměnné koeficienty v diferenciálních operátorech systému nestlačitelných Navier-Stokesových rovnic s rovnicí teploty. Prostorová diskretizace problému cílí na uplatnění metod vysokého řádu, metody spektrálních elementů. Jevy spojené s aproximacemi vysokého řádu jsou diskutovány na řadě příkladů a srovnání s více rozšířenými metodami nižšího řádu. Výsledků bylo dosaženo pro 2 tekutiny s odlišnou teplotní odezvou, vzduch a vodu. Sledovaným parametrem proudění je zejména frekvence odtrhávání vírů, Strouhalovo číslo, v závislosti na teplotě a rychlosti proudění. Vypočtené hodnoty byly porovnány s výsledky experimentů a vykazují dobrou shodu. Numerická analýza závislosti úhlu odtržení proudu při obtékání rotačního válce na teplotě, může dát nový podnět k ověření přesnosti a spolehlivosti vypracované metody.

Klíčová slova: Navier-Stokes-Fourierovy rovnice, metoda spectrálních elementů, teplotně závislé proudění, úhel odtržení

Title: Numerical modeling of unstable fluid flow past heated bodies

Author: Jan Pech

Department: Mathematical Institute of Charles University

Supervisor: prof. Ing. František Maršík, DrSc., Mathematical Institute of Charles University

Abstract: Presented work brings new results to numerical computations of flow influenced by temperature changes. Constructed numerical algorithm takes into account variable coefficients of the differential operators in the system of incompressible Navier-Stokes equations coupled with thermal heat equation. The spatial discretisation of the problem targets to application of high order method, the spectral element method. Phenomenons connected with high order approximations are discussed on a number of examples and comparisons with methods of lower order, which are more common. Results were achieved for two fluids with opposite response to heating, air and water. The observed quantity is particularly a frequency of vortex shedding, the Strouhal number, as dependent on temperature and Reynolds number. The calculated values were compared with experimental results and exhibit a good coincidence. Numerical analysis of separation angle in flow around heated circular cylinder may give a new impulse to verification of accuracy and reliability of the developed method.

Keywords: Navier-Stokes-Fourier, spectral element methods, heated flow, separation angle

Contents

Introduction

This work concerns two problems

- design of algorithm for computations of incompressible flow influenced by temperature changes
- high order approximation to the solution in spatial coordinates (at single time step).

Motivation for study of the heated flow was a missing numerical simulation to the work (Vít et al. [40]), where the frequency of vortex shedding in flow around heated cylinder was investigated experimentally. Empirical formula for the dependence of the Strouhal number on the Reynolds number and heating was derived in (Maršík et al. [27]) and our numerical results are compared with this result.

The studied flow is in the range of transition from laminar to turbulent state, when the Kármán vortex street occurs. The quantities describing this regime of flow evolve in time, but the evolution is relatively slow and the fields seems to be smooth in view of physical experiments. Expectation of the quantities smoothness motivates us to use high order methods to represent its spatial distribution (Canuto [8]-[9], Karniadakis [25], Peyret [32], Šolín [38]). The computational algorithm is designed such, that use of the high order spatial approximation is advantageous.

The high order methods are capable to reach the computer precision approximation for smooth problems. But various types of singularities occur in mathematical solutions of differential equations. The high order methods show these mathematical aspects and connect theory of the mathematical analysis with computational praxis. The high order methods also provide insight to function representation in a transformed space, its coefficient spectra, which contain information about quality of the approximation and which are not available in the low order methods. However, setting the problem and the time discretisation algorithm to work with mathematically smooth problem is still demanding work and an automatic process is missing. Therefore the high order methods got more attention in academic area and up to the authors knowledge, they are not used in commercial engineering codes.

The text is organized to three chapters. The problem is formulated in sense of physics and continuum mechanics in the first chapter. The final system of differential equations is derived and present results from mathematical analysis of this problem are collected.

Second chapter contains construction of the computational algorithm and discussion of the high order methods.

Practical applications of the developed numerical algorithms are provided in the third chapter.

Nomenclature

The symbols are used interchangeably in the physical formulation concerning the SI units and the dimensionless, mathematical form.

Symbol	Meaning	Definition or SI units
Ω	(computational) domain	
$\partial\Omega$	boundary of the domain Ω	
$\mathbf{n}$	normal vector	
$\mathbf{t}$	tangential vector	
$\cdot$	dot product	$\mathbf{a} \cdot \mathbf{b} = \sum_i a_i \, b_i$
$:$	Frobenius product	$\mathbb{A} : \mathbb{B} = \sum_{ij} a_{ij} b_{ij}$
$\otimes$	tensor product	$\mathbf{a} \otimes \mathbf{b} = \begin{pmatrix} a_1 b_1 & a_1 b_2 & \dots \\ a_2 b_1 & a_2 b_2 & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}$
$\frac{\partial}{\partial t}$	partial derivative (in time)	
$\frac{\mathrm{d}}{\mathrm{d}t}$	total derivative	
∇	gradient	$\nabla \mathbf{f} = (\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \dots)^{\mathrm{T}}$
$\nabla\cdot$	divergence	$\nabla \cdot \mathbf{a} = \sum_i \frac{\partial a_i}{\partial x_i}$
$\nabla\times$	curl	
$\|\cdot\|_{\mathcal{X}}$	norm in space $\mathcal{X}$	
$\mathbf{a}$	a vector quantity	$\mathbf{a} = (a_1, a_2, \dots)^{\mathrm{T}}$
A	operator	
$\mathbf{A}$	vector operator	
$\mathbb{A}$	tensor quantity/matrix	$\mathbb{A} = \begin{pmatrix} a_{11} & a_{12} & \dots \\ a_{21} & a_{22} & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}$
D	dimension of the space in spatial coordinates	
x_i	spatial coordinates	$i = 1, \dots, D$
$\vec{x}$	position vector	$\vec{x} = (x_1, \dots, x_D)^{\mathrm{T}}$
$\mathcal{C}$	complex numbers	
c_p	specific heat at constant pressure	$[J \, kg^{-1} \, K^{-1}]$

c_V	specific heat at constant volume	$[J\,kg^{-1}\,K^{-1}]$
$\mathbb{D}$	strain rate tensor	
e	internal energy	$[J]$
E	total energy	$[J]$
$\mathbf{g}$	acceleration vector due to Earth's gravity	$m\,s^{-2}$
h	specific enthalpy	$[J]$
$\mathfrak{I}c$	imaginary part of number c	
$J_n^{(\alpha,\beta)}$	Jacobi polynomial of n-th order	
L	characteristic length	$[m]$
p	pressure	$[Pa]$
$\tilde{p}$	kinematic pressure	$p = \tilde{p}/\rho$
$\mathbf{v}$	velocity	$[m\,s^{-1}]$
t	time	[s]
Δt	time step	
T	temperature	$[K]$
$\mathbb{T}$	stress tensor	
κ	thermal conductivity	$[W\,m^{-1}\,K^{-1}]$
μ	dynamic viscosity	$[Pa\,s]$
ν	kinematic viscosity	$[m^2\,s^{-1}]$
ρ	density	$[kg\,m^{-3}]$
$\mathcal{R}$	real numbers	
$\mathfrak{R}c$	real part of number c	
Pr	Prandtl number	
Re	Reynolds number	
Ri	Richardson number	
St	Strouhal number	
t	time	$[s]$
$U_n^{(\alpha)}$	n-th order ultraspherical polynomial	$U_n^{(\alpha)} = J_n^{(\alpha,\alpha)}$

1. Heated flow: physical and mathematical model

The aim of this work is in numerical model of a flow under change of a fluids temperature. In this chapter, we will formulate the system of basic equations and conclude with the mathematical formulation.

Properties of fluids depend mainly on the temperature, density and the deformation rate. In the heated flows, it may seem, that the most significant is the influence of heating in the change of density, which results in buoyant flows. This is not the case in our study, since the forced convection generating the von Kármán vortex street will dominate in comparison to the buoyancy. However, we will primarily examine the balance equations in the sense of Boussinesque approximation, which introduces a linear dependence of the buoyancy on the temperature in the incompressible flow. The buoyancy will be finally neglected, but the Boussinesque approach will be used to establish incompressibility in case of the heated flow.

It is the thermal dependence of viscosity, what is responsible for changes in the vortex structures of the flow in our settings. It do not cause the fluids motion itself, as buoyancy does, but results in change of the frequency of vortex shedding, as measured in (Vít [40]). To preserve the physical characteristics of a real fluid, we have to discuss also temperature dependence of other quantities acting in the model. This is mostly the case of the thermal conductivity. Its change with temperature, together with change of viscosity, moderate locally the Prandtl number, which is a scaling factor in the dimensionless form of the governing equations (c.f. 1.2).

Since the accurate experimental data as for air as for water are available, we decided to examine both these fluids.

1.1 Basic equations

Our formulation of the governing equations follows the Eulerian approach, which describes the motion of the continuum as a distribution of velocity $\mathbf{v}$ in particular domain of observation Ω.
Let $\Omega \subset \mathbb{R}^D, D = 1, 2, 3$ be a domain occupied by the fluid. If not explicitly written, we will simplify writing of variables by omitting the dependencies on the time t and position vector $\vec{x}$, we denote a scalar functions $s = s(\vec{x}, t)$, vectors (in D-dimensional space) $\mathbf{v} = (v_1(\vec{x}, t), \ldots, v_N(\vec{x}, t))^{\mathrm{T}}$ and tensors (or objects represented by a matrix) $\mathbb{T} = \mathbb{T}(\mathbf{v}, t)$.

The notion of physical quantities will be as follows:
ρ density, $\mathbf{v}$ velocity, $\mathbb{T}$ stress tensor, $\mathbf{f}$ (specific)[1] volumetric force, E (specific) total energy and $\mathbf{q}$ heat flux. The total energy E is a sum of the specific internal energy e and specific mechanical energy

$$E = e + \frac{1}{2}(\mathbf{v} \cdot \mathbf{v}) \ . \tag{1.1}$$

[1] A *specific* quantity is one per unit mass

The balances of mass, momentum and energy give the system of equations for unknowns ρ, $\mathbf{v}$ and E

- Continuity equation (conservation of mass)

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{v}) = 0 \tag{1.2}$$

- The equation of motion (balance of momentum)[2]

$$\rho \frac{\mathrm{D}\mathbf{v}}{\mathrm{D}t} = \nabla \cdot \mathbb{T} + \rho \mathbf{f} \tag{1.3}$$

- The energy equation (balance of total energy)

$$\rho \frac{\mathrm{D}E}{\mathrm{D}t} = \rho \mathbf{f} \cdot \mathbf{v} + \nabla \cdot (\mathbb{T}\mathbf{v}) - \nabla \cdot \mathbf{q}$$

  may be simplified by subtracting the whole balance of mechanical energy, what results in balance of internal energy [3]

$$\rho \frac{De}{Dt} = \mathbb{T} : \nabla \mathbf{v} - \nabla \cdot \mathbf{q}\,. \tag{1.4}$$

The last equation, sometimes noted as the *thermal energy equation* (e.g. Kundu [26]), expresses the I. Law of Thermodynamics.

To complete the governing equations (1.2)-(1.4) we have to define the stress tensor $\mathbb{T}$, the term representing volumetric forces $\mathbf{f}$ and vector of the heat flux $\mathbf{q}$. Definitions of these quantities introduce material properties (coefficients or functions) which complement the governing equations to specific problem.

In our study, we will constrain to the fluids of Newtonian type, whose rheological equation[4] fulfils a linear dependence between the strain rate $\nabla \mathbf{v}$ and the stress tensor. The coefficient of the linearity relation, however, is dependent on the temperature and therefore is variable both in time and space. We will restrict to those materials, for which the balance of angular momentum implies symmetry of the strain rate tensor[5]

$$\nabla \mathbf{v} = \mathbb{D} = \frac{1}{2}[\nabla \mathbf{v} + (\nabla \mathbf{v})^{\mathrm{T}}]\,.$$

Collecting these properties, we arrive to the rheological equation of a generalized Newtonian fluid

$$\mathbb{T} = -p^* \mathbb{I} + \lambda \nabla \cdot \mathbf{v} + 2\mu(T)\mathbb{D}\,. \tag{1.5}$$

[2] We introduce the *material derivative* $\dfrac{\mathrm{D}}{\mathrm{D}t} = \dfrac{\partial}{\partial t} + (\mathbf{v} \cdot \nabla)$ to simplify the notation; $\mathbf{a} \otimes \mathbf{b}$ is a tensor with components $a_i b_j$

[3] $\nabla \cdot (\mathbb{T}\mathbf{v}) = \mathbf{v} \cdot (\nabla \cdot \mathbb{T}) + \mathbb{T} : \nabla \mathbf{v}$, resp. in components $\nabla \cdot (\mathbb{T}\mathbf{v}) = \dfrac{\partial}{\partial x_k}(t_{kl} v_l) = \dfrac{\partial t_{kl}}{\partial x_k} v_l + t_{kl} \dfrac{\partial v_l}{\partial x_k}$, where we use the Einsteins summation convention.

[4] Relations between the stress tensor and material properties are the *rheological equations.*

[5] $\mathbb{D}_{kl} = \dfrac{1}{2}\left(\dfrac{\partial v_k}{\partial x_l} + \dfrac{\partial v_l}{\partial x_k}\right)$

In the above formula, p^* denotes pressure ($[p^*] = \mathrm{Pa}$; star is used to simplify later notation, since p will be used for the *kinematic pressure*[6]), μ is non-constant $\mu = \mu(T) = \mu(\vec{x}, t)$ dynamic viscosity, defining the linear dependence of $\mathbb{T}$ on $\mathbb{D}$ and *second viscosity* λ describes partly the resistance of material to change its volume.

Concerning the energy equation, we introduce the *Fourier law*

$$\mathbf{q} = -\kappa \nabla T \tag{1.6}$$

to define $\mathbf{q}$. The material specific function $\kappa = \kappa(T)$ represents the *thermal conductivity*.

Incompressibility and Boussinesque approximation

Propagation of sound is an intrinsic property of material. It manifests the compressibility, change of materials density. However, under certain conditions the compressibility may be neglected, what have striking consequences especially in design of computational schemes (c.f. Section 2.1).
The compressibility of fluids is negligible in wide range of thermodynamic states, regardless, whether the fluid is a liquid or a gas. For example the change of water volume with pressure change of 10^5Pa (one atmosphere) is only about 5%. Similarly the compressibility of gases is negligible in case of flow velocities which are much lower than the speed of sound. Denoting by M the Mach number[7], the incompressible approach is usually used when $\mathrm{M} \ll 1$. Avoiding extreme situations, speed of motion in water is much lower than the speed of sound, so the incompressible approach is acceptable to a wide range of models. The situation differs in case of gas motion, when compressibility plays crucial role in many technical applications and incompressible model must be used carefully.

In both studies, air and water flows, the velocity magnitude will be safely under the above mentioned limit, given by the Mach number.

Using the material derivative, we rewrite equation (1.2) into the form

$$\frac{1}{\rho}\frac{\mathrm{D}\rho}{\mathrm{D}t} + \nabla \cdot \mathbf{v} = 0\,.$$

Following the Boussinesque approximation (e.g. Kundu [26]), which introduces incompressible model with buoyancy, we neglect the term with density change and arrive to

$$\nabla \cdot \mathbf{v} = 0\,. \tag{1.7}$$

The last equation is obtained also in the case, when $\rho(\vec{x}, t) = \text{const.}$, but for wider range of problems its meaning is, that the magnitude of the term $\dfrac{1}{\rho}\dfrac{\mathrm{D}\rho}{\mathrm{D}t}$ is negligible in comparison with magnitude of $\nabla \cdot \mathbf{v}$.

The condition (1.7) is a constraint to the velocity field and its geometric meaning is the zero volumetric strain rate. This constraint influences those terms in the momentum and energy equation, where acts on the stress tensor $\mathbb{T}$ and on the forcing $\mathbf{f}$.

[6] $p = p^*/\rho$

[7] Denoting by c the speed of sound and $|\mathbf{v}|$ the flow speed, Mach number is defined as $\mathrm{M} = |\mathbf{v}|/c$

In those cases, when the Boussinesque approximation is valid, the forcing term takes form

$$\rho\mathbf{f} = \rho_\infty[1 + \beta(T - T_\infty)]\mathbf{g}\,, \tag{1.8}$$

since the density is expected to depend linearly on temperature, $\rho \simeq \rho_\infty[1 + \beta(T - T_\infty)]$. In (1.8), the subscript ∞ denotes a constant reference values and $\mathbf{g}$ is the acceleration vector due to the Earth's gravity, β is the *volumetric thermal expansion coefficient*.

Assumption of incompressibility reduces the stress tensor (1.5) to

$$\mathbb{T} = -p^*\mathbb{I} + 2\mu\mathbb{D} \tag{1.9}$$

and the term $\lambda\nabla\cdot\mathbf{v}$ is neglected.

Equation (1.3) under the Boussinesque approximation takes form

$$\rho_\infty\left(\frac{\partial\mathbf{v}}{\partial t} + \mathbf{v}\cdot\nabla\mathbf{v}\right) = -\nabla p^* + \nabla\cdot(2\mu\,\mathbb{D}) + \rho_\infty\beta(T - T_\infty)\mathbf{g}\,, \tag{1.10}$$

where we used also (1.9).

Substituting the stress tensor in form (1.9) to the thermal energy equation (1.4), the deformation work term becomes

$$\mathbb{T} : \nabla\mathbf{v} = -p^*\nabla\cdot\mathbf{v} + 2\mu\mathbb{D} : \nabla\mathbf{v}\,, \tag{1.11}$$

or[8]

$$\mathbb{T} : \mathbb{D} = -p^*\nabla\cdot\mathbf{v} + 2\mu\mathbb{D} : \mathbb{D}\,, \tag{1.12}$$

as a result of symmetry in $\mathbb{D}$. The dissipative term $2\mu\mathbb{D} : \mathbb{D}$ represents the viscous heating[9]. Magnitude of this term in comparison to the left side of (1.4) is usually very small ($\sim 10^{-7}$) and we can neglect it for this reason.

The term $p^*\nabla\cdot\mathbf{v}$ represents the change of temperature due to the fluids expansion/compression and may be neglected for the incompressible liquids. The deformation work of incompressible liquids is then negligible and we can set

$$\mathbb{T} : \mathbb{D} = 0\,. \tag{1.13}$$

However, the term $p^*\nabla\cdot\mathbf{v}$ stays significant for gases also in the case of incompressible approximation. We remind now, that the gas of our interest is air, which well satisfy the state equation of the ideal gas[10]. Using the continuity equation (1.2), we can rewrite the mentioned term

$$-\,p^*\nabla\cdot\mathbf{v} = \frac{p^*}{\rho}\frac{D\rho}{Dt} \simeq \frac{p^*}{\rho}\left(\frac{\partial\rho}{\partial T}\right)_p\frac{DT}{Dt} = -p^*\beta\frac{DT}{Dt}\,, \tag{1.14}$$

where $\left(\dfrac{\partial\rho}{\partial T}\right)_p$ is a temperature variation of density at constant pressure. We get

$$\beta = -\frac{1}{T} \tag{1.15}$$

[8] the product of symmetric and antisymmetric tensor is zero

[9] In 2D we have particularly $2\mu\mathbb{D} : \mathbb{D} = \mu\left\{2\left[\left(\dfrac{\partial v_1}{\partial x_1}\right)^2 + \left(\dfrac{\partial v_2}{\partial x_2}\right)^2\right] + \left(\dfrac{\partial v_1}{\partial x_2} + \dfrac{\partial v_2}{\partial x_1}\right)^2\right\}$

[10] $p^* = \rho RT$; R denotes the gas constant, which is universal gas constant divided by the molecular mass

for ideal gases. Finally, the deformation work for ideal gases states

$$\mathbb{T} : \nabla \mathbf{v} = -\rho(c_p - c_V)\frac{\mathrm{D}T}{\mathrm{D}t}\,, \tag{1.16}$$

since $R = c_p - c_V$ (c_p and c_V are specific heat capacities at constant pressure and constant volume).

The influence of heating will be studied in temperature ranges of $T \in [16; 23]^\circ C$ for water and $T \in [298; 537]K$ for air. In both cases, change of the c_p and c_V can be neglected. Only in case of the highest temperatures in the air flow the difference reaches values around 5%.

If the specific heat capacities are constant, the fluid may be examined as *calorically perfect*, what results in linear dependence between internal energy and temperature

$$e = c_V T \tag{1.17}$$

and specific enthalpy h and temperature

$$h = c_p T\,. \tag{1.18}$$

Concerning the Fourier law (1.6), expression (1.13) and (1.17) in the energy equation (1.4) we arrive to thermal energy equation for water

$$\rho c_V \frac{\mathrm{D}T}{\mathrm{D}t} = \nabla \cdot \kappa \nabla T\,, \tag{1.19}$$

which describes evolution of the temperature field.

Similarly, substituting (1.16) and (1.17) to (1.4), we get temperature evolution for flow of air in incompressibility approximation

$$\rho c_p \frac{\mathrm{D}T}{\mathrm{D}t} = \nabla \cdot \kappa \nabla T\,. \tag{1.20}$$

Remark: Both the (1.19) and (1.20) differ only in the coefficient in the left side, but its derivation differ substantially. In view of (1.17) and (1.18), (1.19) is equation describing evolution of specific internal energy, while (1.20) is evolution of specific enthalpy.

Following the experimental setting presented in (Vít [40]), we neglect an influence of the buoyancy. Therefore the temperature effects in the flow are solely described by temperature variations of μ and κ in our model.

1.1.1 Temperature dependent properties

Viscosity

The dynamic viscosity μ expresses mutual sliding of layers in fluids. It is the coefficient relating velocity gradient and tangential stress. It is generally related to the substance, temperature[11], shear rate, time (strain history), pressure, etc. Avoiding extreme cases, viscosity of gases do not depend on pressure, but viscosity of liquids slightly increase with pressure. For the most of the fluids, viscosity

[11] Strong dependencies of viscosity on temperature is not rare. Several mineral oils lose about 10% of their viscosity with every Kelvin grade of increase in temperature

also do not depend on the shear rate, this is the case of the Newtonian fluids. Viscosity of gases increases with temperature, but viscosity of liquids has opposite tendency. Impacts of the temperature dependence results in change of the flow structures in the heated flow. As we will observe later in computational results, it is responsible for moderation of the Strouhal number (St, c.f. table 1.1) in flow around a heated/cooled cylinder.

For computations, we need a functional dependence $\mu = \mu(T)$, which is to be imposed in the governing equations. Various empirical formulas relating the viscosity of liquids to temperature are widely used

- Arrhenius Law
$$\mu(T) = C_1\, e^{C_2/(C_3+T)} \tag{1.21}$$

- Andrade's Law:
$$\mu(T) = C_1\, e^{C_2/T} \tag{1.22}$$

For water in temperature range $T \in [10; 100]^\circ C$ the least-squares fitting for Andrade's law results in constants $C_1 = \exp(-12.9896)$ and $C_2 = 1780.622$.

Widely used formula for dynamic viscosity of gases as dependence on absolute temperature is Sutherland's Law:

$$\mu(T) = \mu_\infty \left(\frac{T}{T_\infty}\right)^{3/2} \frac{T_\infty + C_1}{T + C_2} \tag{1.23}$$

where for air the constants $C_1 = C_2 = 110.5K$, if $T_\infty = 273K$ and $\mu_\infty = \mu(T_\infty)$. Formula (1.23) results from theory of ideal gases and idealised potential of intermolecular forces. For a limited temperature range a power law approximation may be used

$$\mu(T) = \mu_\infty \left(\frac{T}{T_\infty}\right)^{\omega} . \tag{1.24}$$

We will follow the data presented in (Gebhart [14]), so the powers in (1.24) are ([27])

- air:
$$\omega_a = 0.7774 \tag{1.25}$$

- water:
$$\omega_w = -7\,. \tag{1.26}$$

The power law form was used in computations presented in chapter 3, since it was used in the work (Maršík [27]), whose results are also compared with present numerical simulations.

Temperature dependences of dynamical viscosities of water and air are plotted in Figure 1.1.

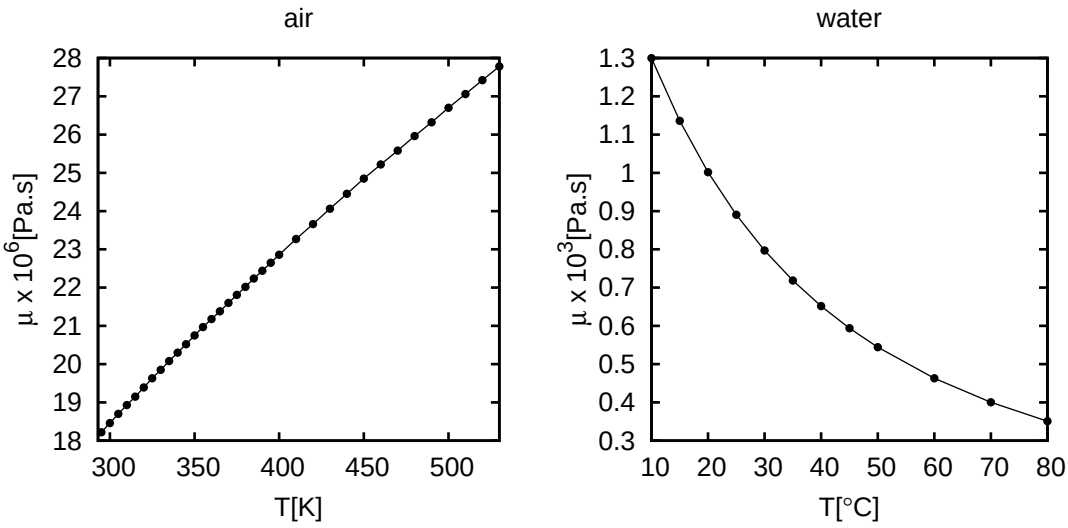


Figure 1.1: Dependence of dynamical viscosity of air and water. Data taken from (Gebhart [14]) and fitted by power law (1.24).

Thermal conductivity

Thermal conductivity of both air and water exhibits temperature dependence as well. Omitting this fact in a model may result, at least locally, in incorrect Prandtl number Pr (c.f. table 1.1).The experimental data for the thermal conductivity of water have been fitted to a quadratic functional form in Ramires [35]:

$$\kappa^* = -1.48445 + 4.12292T^* - 1.63866T^{*2} \,. \tag{1.27}$$

In the above expression $\kappa^* = \kappa(T)/\kappa(298.15)$ and $T^* = T/298.15$ are dimensionless thermal conductivity and temperature, respectively. The value of κ is the adopted standard value of thermal conductivity of water at $298.15\,K$ and $0.1\,MPa$. Recommended value from Ramires [35] is

$$\kappa(298.15) = 0.6065 \pm 0.0036\,W\,m^{-1}\,K^{-1}\,.$$

In our study, we follow again the fitting to power law and the data from Gebhart [14]:

$$\kappa = \kappa_\infty \left(\frac{T}{T_\infty}\right)^\omega , \tag{1.28}$$

where $\omega = 0.71$ for water and $\omega = 0.85$ for air.

Plots of these temperature dependencies for both air and water is in Fig. 1.2.

1.2 Mathematical formulation

The equations (1.10), (1.7), (1.20) or (1.19) are the balance equations of the physical quantities in the physical units. But mathematical formulation is independent of physical units and it works with equations scaled by numerical values only.

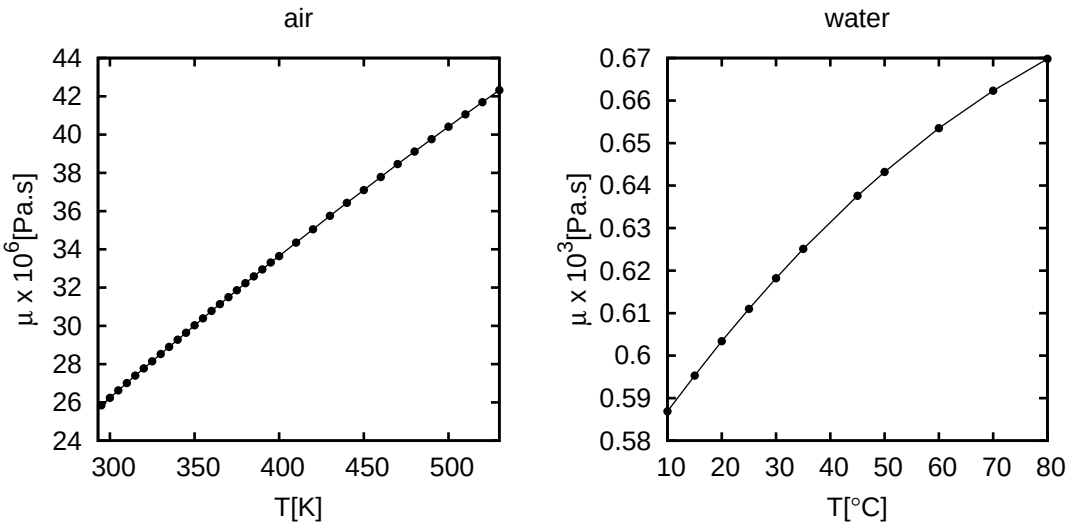


Figure 1.2: Temperature dependence of thermal conductivity for water and air. Data taken from Gebhart [14].

Redefinition of the coordinate system

$$t = \tau \tilde{t}\,,\ \vec{x} = L\tilde{\vec{x}}\,, \tag{1.29}$$

allows to write the equations independently to the physical units. Value L is the *characteristic length* (eg. diameter of an obstacle in the flow), significant dimension in the studied problem (e.g. diameter of channel in case of channel flow). Value of τ results from other quantities, as seen on the case of velocity, $\mathbf{v} = |\mathbf{v}_\infty|\tilde{\mathbf{v}}$, scaled by its inlet magnitude $|\mathbf{v}_\infty|$

$$\mathbf{v} = \frac{\partial \vec{x}}{\partial t} = \frac{L}{\tau}\frac{\partial \tilde{\vec{x}}}{\partial \tilde{t}} = |\mathbf{v}_\infty|\tilde{\mathbf{v}} \Rightarrow \tau = \frac{L}{|\mathbf{v}_\infty|}\,. \tag{1.30}$$

We use the subscript ∞ to denote constant-in-time values scaled in physical units, which are used for the time average and boundary values. Quantities with tilde are numerical values, independent of physical units. Other quantities are defined accordingly in the new coordinates

$$\rho = \rho_\infty \tilde{\rho}\,,\ \mu = \mu_\infty \tilde{\mu}\,,\ T = T_\infty\, \tilde{T} \text{ or } T = (\Delta T)\tilde{T} + T_\infty$$

$$\kappa(T) = \kappa(T_\infty)\tilde{\kappa}(T) = \kappa_\infty \tilde{\kappa}(T)\,.$$

Two definition of $\tilde{T}$ were presented, the former will be used in computations in chapter 3, while the latter is traditionally used since it introduces the characteristic temperature difference ΔT, which acts in the definitions of the Richardson and the Grashof numbers (c.f. table 1.1).

The system of governing equations (1.7, 1.10, 1.19 or 1.20) with the stress tensor $\mathbb{T}$ defined in (1.9) becomes

$$\tilde{\nabla} \cdot \tilde{\mathbf{v}} = 0 \tag{1.31a}$$

$$\frac{\partial \tilde{\mathbf{v}}}{\partial \tilde{t}} + \tilde{\mathbf{v}} \cdot \tilde{\nabla}\tilde{\mathbf{v}} = -\tilde{\nabla}\tilde{p} + \frac{1}{\mathrm{Re}}\tilde{\nabla} \cdot \left[\tilde{\mu}\left(\tilde{\nabla}\tilde{\mathbf{v}} + (\tilde{\nabla}\tilde{\mathbf{v}})^{\mathrm{T}}\right)\right] - \left\{\mathrm{Ri}\,\tilde{T}\tilde{\mathbf{g}}\right\} \tag{1.31b}$$

Name	label	definition	meaning
Eckert	Ec	$\|\mathbf{v}_\infty\|^2 / c_p \Delta T$	mech. energy/char. enthalpy difference
Grashof	Gr	$\|\mathbf{g}\| \beta \Delta T L^3 / \nu^2$	buoyancy/viscous forces
Péclet	Pe	Re Pr	advective/diffusive transport rate
Prandtl	Pr	$\mu c_p / \kappa_\infty$	momentum/thermal diffusivity
Reynolds	Re	$\|\mathbf{v}_\infty\| L / \nu_\infty$	inertial forces/viscous forces
Richardson	Ri	$\mathrm{Gr}/\mathrm{Re}^2$	buoyancy/flow gradient
Strouhal	St	$fL / \|\mathbf{v}_\infty\|$	frequency

Table 1.1: Nondimensional parameters.

$$\frac{\partial \tilde{T}}{\partial \tilde{t}} + \tilde{\mathbf{v}} \cdot \tilde{\nabla} \tilde{T} = \frac{1}{\mathrm{Pe}} \tilde{\nabla} \left(\tilde{\kappa} \tilde{\nabla} \tilde{T} \right) + \left\{ \frac{\mathrm{Ec}}{\mathrm{Re}} \tilde{\mathbb{D}} : \tilde{\mathbb{D}} \right\}, \tag{1.31c}$$

where we denote $\tilde{\nabla}$ the gradient operator in coordinates $\tilde{\vec{x}}$, $\tilde{p} = \dfrac{p^*}{\rho_\infty |\mathbf{v}_\infty|^2} = p/|\mathbf{v}_\infty|^2$ (p is the *kinematic pressure*), $\tilde{\mathbf{g}}$ dimensionless acceleration due to Earth's gravity, $\tilde{\mathbb{D}}$ symmetric part of the strain rate tensor in dimensionless form and the dimensionless parameters as listed in Table 1.1. Phenomenons described by the terms in the braces (buoyancy and viscous heating) are negligible in the following study of the flow around heated cylinder (c.f. chapter 3).

Irrespective of particular values of the quantities, two experiments exhibit the same phenomenons, if the nondimensional parameters are same in both cases. Hypothesis of the continuum together with the dimensionless numbers in equations allows us to set the mathematical description to particular physical phenomena independently on the physical units.

Henceforward, we will work with dimensionless formulation, but we will omit writing the $\sim$ symbol to simplify the notation. The final system of equations, which will be discussed and approximatively solved in the following, consists of the momentum equation[12]

$$\frac{\partial \mathbf{v}}{\partial t} + (\mathbf{v} \cdot \nabla)\mathbf{v} = -\nabla p + \frac{1}{Re} \nabla \cdot (\mu(T)\mathbb{D}) \tag{1.32}$$

under the constraint of incompressibility

$$\nabla \cdot \mathbf{v} = 0 \tag{1.33}$$

and the equation for temperature distribution, which takes form

$$\frac{\mathrm{D}T}{\mathrm{D}t} = \gamma \frac{1}{Pe} \nabla \cdot (\kappa(T) \nabla T) \tag{1.34a}$$

for incompressible liquid (water) and

$$\frac{\mathrm{D}T}{\mathrm{D}t} = \frac{1}{Pe} \nabla \cdot (\kappa(T) \nabla T) \tag{1.34b}$$

[12] $\nabla \cdot (\mu(T)\mathbb{D}) = \dfrac{\partial}{\partial x_k} \left[\mu(T) \left(\dfrac{\partial v_k}{\partial x_l} + \dfrac{\partial v_l}{\partial x_k} \right) \right]$

for gas in incompressible approximation. Coefficient $\gamma = c_p/c_V$ was introduced in (1.34a) to satisfy definition of the Péclet number. Definition of γ coincides with the (Poisson) *adiabatic constant*.

1.2.1 Mathematical analysis

In terms of mathematics, the system (1.32)-(1.34) consists of the Incompressible Navier-Stokes equations (INS) with the energy equation, noted also as the Navier-Stokes-Fourier system. It forms a set of coupled non-linear differential equations.

Up to the authors knowledge, the analysis of this problem, concerning physically realistic setting (boundary condition), is not closed.

The subject of this work stays in numerical solution of the system and extending results of the mathematical analysis is beyond its scope. However, to underlay the presented discrete solutions, we collect some known results from analysis on this place. Especially, we will be concerned of existence of a unique solution and its continuous dependence on the data, since we need to establish wellposedness (Appendix A, Definition 4) to ensure convergence of the discretisation methods.

Both cited results, (Pérez [30]) and (Bulíček [4]), analyze more general systems, than is our objective (buoyancy included in [30], viscous heating included in [4]). But concerning boundary conditions and shape of the computational domain, both of them meet our setting only partially.

Existence of the weak and *suitable weak solution*[13] has been proven to a system in fully thermodynamic setting in (Bulíček [4]). Limiting for us in view of the velocity $\mathbf{v}$, is the absence of other than impermeable boundary condition ($\mathbf{v}\cdot\mathbf{n} = 0$) and the *Navier's slip* boundary condition, which excludes the no-slip condition. This result is limited also in view of temperature, since it expects *no heat exchange* ($\nabla\cdot T = 0$ on $(0,T)\times\partial\Omega$). On the other side, the system analyzed in [4] contains the viscous-heating term, is analysed in 3D and continuously in time, what is not the case of the result in (Pérez [30]).

Work [30] analyse the system (1.32)-(1.34) with the buoyancy in the Boussinesque approximation. The analysis is simplified by discretisation in time, when the existence, uniqueness and continuous dependence on data is analysed for single time layer, resp. steady problem. However, the idea coincides with the problem arising from discretisation by backward difference formulas in time, what is the technique used in sec. 2.1. Boundary conditions in the discussed problem encompass both the flow through the domain and heat exchange, therefore we will describe this result in more detail.

Let Ω is an open, bounded, convex domain with a Lipschitz boundary $\partial\Omega$ and $\partial\Omega$ is composed of the part Γ_D, on which the Dirichlet condition is prescribed and Γ_N with Neumann type condition ($\partial\Omega = \Gamma_D \cup \Gamma_N$). $L^2(\Omega)$ is the standard Lebesgue space of square integrable functions with norm denoted by $\|\cdot\|_0$ and scalar product $(\cdot,\cdot)$. We define the standard functional spaces

$$\mathcal{H}^1(\Omega) = \{v \in L^2(\Omega), \nabla v \in \boldsymbol{\mathcal{L}}^2(\Omega)\} \tag{1.35}$$

$$\mathcal{H}^1_{0,\Gamma_D} = \{v \in H^1(\Omega), v|_{\Gamma_D} = 0\} \tag{1.36}$$

[13] The suitable weak solution is defined in [4] as a weak solution satisfying the weak form of the entropy inequality.

$$\mathcal{H}^{1/2}(\Gamma_D) = \{\eta \in L^2(\Gamma_D), \exists v \in \mathcal{H}^1(\Omega, v|_{\Gamma_D}) = \eta\}\,. \tag{1.37}$$

Bold symbol is used for vector-valued spaces, e.g. $\boldsymbol{\mathcal{H}}^1_{0,\Gamma_D} = \mathcal{H}^1_{0,\Gamma_D} \times \mathcal{H}^1_{0,\Gamma_D}$ and $\|\cdot\|_{m,\Omega}$ is the norm of the Hilbert space $\mathcal{H}^m(\Omega)$.

The variable properties are expected to be bounded

$$\exists \mu_1, \mu_2 = \text{const.}\, \forall T \in (0,\infty) :\ 0 < \mu_1 \leq \mu(T) \leq \mu_2 \tag{1.38}$$

$$\exists \kappa_1, \kappa_2 = \text{const.}\, \forall T \in (0,\infty) :\ 0 < \kappa_1 \leq \kappa(T) \leq \kappa_2 \tag{1.39}$$

Introduction of the "frozen" values of temperature $\hat{T}$ and velocity $\hat{\mathbf{v}}$ (e.g. values at particular time level of the temporal discretisation) allows to decouple the momentum and energy equations and consider existence and uniqueness of the two separate steady problems

1. the steady incompressible Navier-Stokes problem with fixed temperature $\hat{T}$

$$\begin{gathered} \mathbf{v}\cdot\nabla\mathbf{v} - \frac{1}{\mathrm{Re}}\nabla\cdot(\mu(\hat{T})\mathbb{D}(\mathbf{v})) + \nabla p = \mathrm{Ri}\,\hat{T}\mathbf{g} + \mathbf{F} \\ \nabla\cdot\mathbf{v} = 0\,, \end{gathered} \tag{1.40a}$$

   suject to boundary conditions

$$\begin{gathered} \mathbf{v} = \mathbf{v}_\infty \text{ ...at inflow,} \\ \mathbf{v} = \mathbf{0} \text{ ...no-slip at walls,} \\ \mathbb{T} : \mathbf{n} = \mathbf{0} \text{ ...zero normal traction forces at outflow} \end{gathered} \tag{1.40b}$$

2. steady advection-diffusion problem with fixed divergence-free velocity field $\hat{\mathbf{v}}$ and fixed temperature $\hat{T}$ for linearisation of the diffusive term

$$\hat{\mathbf{v}}\cdot\nabla T - \frac{1}{\mathrm{Pe}}\nabla\cdot(\kappa(\hat{T})\nabla T) = 0\,, \tag{1.41a}$$

   with following boundary conditions

$$\begin{gathered} T = T_\infty \text{ ...at inflow} \\ T = T_W \text{ ...at wall of the body} \\ (\kappa(T)\nabla T)\cdot\mathbf{n} = 0 \text{ ...at the outflow.} \end{gathered} \tag{1.41b}$$

Both problems are analysed in its weak form, but the a priori estimates are still hard to obtain without any considerations on the solution values on the outflow boundary Γ_N. Operator denoting a weak form of the convective term in the momentum equation

$$\mathbf{b}(\mathbf{v},\mathbf{w},\boldsymbol{\phi}) = \int_\Omega (\mathbf{v}\cdot\nabla)\mathbf{w}\cdot\boldsymbol{\phi} \tag{1.42}$$

becomes

$$\mathbf{b}(\mathbf{v},\mathbf{w},\boldsymbol{\phi}) = \frac{1}{2}\mathbf{b}(\mathbf{v},\mathbf{w},\boldsymbol{\phi}) - \frac{1}{2}\mathbf{b}(\mathbf{v},\boldsymbol{\phi},\mathbf{w}) + \frac{1}{2}\int_{\Gamma_N}\mathbf{v}\cdot\mathbf{n}\mathbf{w}\cdot\boldsymbol{\phi} \tag{1.43}$$

for $\mathbf{v}$ satisfying $\nabla \cdot \mathbf{v} = 0$ in Ω, $\boldsymbol{\phi} = 0$ on Γ_D. The scalar version

$$b(\mathbf{v}, T, \phi) = \frac{1}{2} b(\mathbf{v}, T, \phi) - \frac{1}{2} b(\mathbf{v}, \phi, T) + \frac{1}{2} \int_{\Gamma_N} \mathbf{v} \cdot \mathbf{n}\, T\, \phi \tag{1.44}$$

is used in the weak form of the equation for temperature ($\nabla \cdot \mathbf{v} = 0$ in Ω, $\phi = 0$ on Γ_D). The boundary terms in the above definitions can't be well controlled on the outflow boundary, since the values of velocity are a priory unknown on Γ_N.

Constraint, leading to proof of the weak solutions existence, stays in prohibiting the re-entrant flow. Introducing a term $[\mathbf{u} \cdot \mathbf{n}]^- = \sup\{-(\mathbf{u} \cdot \mathbf{n}), 0\}$, which is non-zero only if the flow on the outflow is re-entrant, we arrive to the modified convective terms

$$\tilde{\mathbf{b}}(\mathbf{v}, \mathbf{w}, \boldsymbol{\phi}) = \mathbf{b}(\mathbf{v}, \mathbf{w}, \boldsymbol{\phi}) + \frac{1}{2} \int_{\Gamma_N} [\mathbf{v} \cdot \mathbf{n}]^- \mathbf{w} \cdot \boldsymbol{\phi} \tag{1.45}$$

$$\tilde{b}(\mathbf{v}, T, \phi) = b(\mathbf{v}, T, \phi) + \frac{1}{2} \int_{\Gamma_N} [\mathbf{v} \cdot \mathbf{n}]^- T \phi\,, \tag{1.46}$$

where the scalar version is applied in the energy equation.

However, introducing the terms (1.45) and (1.46), the solution $(\mathbf{v}, p, T)$ of the weak problem (formulated below) is formally a solution of the equation system (1.40a), (1.41a), but subject to a different (outflow) boundary conditions, than (1.40b),(1.41b)

$$\begin{gathered} \mathbf{v} = \mathbf{v}_D \text{ on } \Gamma_D, \\ \mathbb{T} : \mathbf{n} + \tfrac{1}{2}[\mathbf{v} \cdot \mathbf{n}]^- \mathbf{v} = 0 \text{ on } \Gamma_N \end{gathered} \tag{1.47a}$$

$$\begin{gathered} T = T_D \text{ on } \Gamma_D, \\ \kappa(T) \nabla T \cdot \mathbf{n} + \tfrac{1}{2}[\mathbf{v} \cdot \mathbf{n}]^- T = 0 \text{ on } \Gamma_N\,. \end{gathered} \tag{1.47b}$$

Two weak problems are analysed, while decoupling the system using the frozen fields $\hat{T} \in \mathcal{H}^1(\Omega)$ and $\hat{\mathbf{v}} \in \boldsymbol{\mathcal{H}}^1(\Omega) : \nabla \cdot \hat{\mathbf{v}} = 0$[14]:

1. Let $T_D \in \mathcal{H}^{1/2}(\Gamma_D)$. We say that $T \in \mathcal{H}^1(\Omega)$, such that $T|_{\Gamma_D} = T_D$, is a weak solution to (1.41a), (1.47b) if

$$a_{\hat{T}}(T, \phi) + \tilde{b}(\hat{\mathbf{v}}, T, \phi) = 0 \qquad \forall \phi \in \mathcal{H}^1_{0,\Gamma_D}(\Omega)\,, \tag{1.48}$$

2. Let $\mathbf{v}_D \in \boldsymbol{\mathcal{H}}^{1/2}(\Gamma_D)$. We say that couple $(\mathbf{v}, p) \in \boldsymbol{\mathcal{H}}^1(\Omega) \times \mathcal{L}^2(\Omega)$ is a weak solution of (1.40a), (1.41a) if $\mathbf{v}|_{\Gamma_D} = \mathbf{v}_D$ and

$$\begin{aligned} \mathbf{a}_{\hat{T}}(\mathbf{v}, \boldsymbol{\phi}) + \tilde{\mathbf{b}}(\mathbf{v}, \mathbf{v}, \boldsymbol{\phi}) - (\operatorname{div} \boldsymbol{\phi}, p)_{0,\Omega} \\ = (\mathrm{Ri}\, T \mathbf{g}, \boldsymbol{\phi})_{0,\Omega} + (\mathbf{f}, \boldsymbol{\phi})_{0,\Omega} \qquad \forall \boldsymbol{\phi} \in \boldsymbol{\mathcal{H}}^1_{0,\Gamma_D}(\Omega) \end{aligned} \tag{1.49}$$

[14]Standard lifting technique is applied to the solutions T and $\mathbf{v}$ in order to satisfy the Dirichlet boundary conditions. More precisely, functions

$$T^* \in \mathcal{H}^1(\Omega) : T^*|_{\Gamma_D} = T_D$$

and

$$\mathbf{v}^* \in \boldsymbol{\mathcal{H}}^1(\Omega) : \mathbf{v}^*|_{\Gamma_D} = \mathbf{v}_D, \nabla \cdot \mathbf{v}^* = 0 \text{ in } \Omega$$

are introduced. The unknowns are then $\Theta = T - T^* \in \mathcal{H}^1_{0,\Gamma_D}(\Omega)$ and $\mathbf{u} = \mathbf{v} - \mathbf{v}^* \in \boldsymbol{\mathcal{H}}^1_{0,\Gamma_D}(\Omega)$, what coincides with test spaces in (1.48) and (1.49).

where subscript $\hat{T}$ indicates temperature dependence of the operator $\mathbf{a}_{\hat{T}}$ (and $a_{\hat{T}}$), $\mathbf{g} = (0, -g, 0)^{\mathrm{T}}$ denotes the gravity and

$$\mathbf{a}_{\hat{T}}(\mathbf{v}, \boldsymbol{\phi}) = 2 \int_{\Omega} \frac{\mu(\hat{T})}{\mathrm{Re}} \mathbb{D}(\mathbf{v}) : \mathbb{D}(\boldsymbol{\phi}) \tag{1.50a}$$

$$a_{\hat{T}}(T, \phi) = \int_{\Omega} \frac{\kappa(\hat{T})}{Pe} \nabla T \cdot \nabla \phi \tag{1.50b}$$

Following theorems are collected from results in (Pérez [30])

Theorem 1. *Let $\hat{\mathbf{v}} \in \boldsymbol{\mathcal{H}}^1(\Omega)$ be a given divergence-free velocity field. If $T_D \in \mathcal{H}^{1/2}(\Gamma_D)$ (Dirichlet data) is such that*

$$T_1 \leq T_D(x) \leq T_2 \quad a.e.\ on\ \Gamma_D\,,$$

then, problem (1.48) *has an unique solution $T \in \mathcal{H}^1(\Omega)$ such that $T|_{\Gamma_D} = T_D$, which satisfies*

$$T_1 \leq T(x) \leq T_2 \quad a.e.\ in\ \Omega$$

and

$$\|T\|_{1,\Omega} \leq C \left(\frac{\kappa_2}{\kappa_1} + \frac{\mathrm{Pe}}{\kappa_1} \|\mathbf{v}\|_{1,\Omega} \right) \|T_D\|_{1/2,\Gamma_D} \tag{1.51}$$

Note, that the uniform bound (1.51) do not depend on $\hat{T}$ explicitly, but implicitly through $\mathbf{v} = \mathbf{v}(\hat{T})$.

For the weak form of the steady incompressible Navier-Stokes follows

Theorem 2. *For any $\mathbf{F} \in \boldsymbol{\mathcal{L}}^2(\Omega)$, for any $\hat{T}$ given in $\mathcal{L}^2(\Omega)$, with $T_1 \leq \hat{T} \leq T_2$ and for any $\mathbf{v}_D \in \boldsymbol{\mathcal{H}}^{1/2}(\Gamma_D)$*

- *exists at least one pair $(\mathbf{v}, p) \in \boldsymbol{\mathcal{H}}^1(\Omega) \times \mathcal{L}^2(\Omega)$, solution of* (1.49)*, such that $\mathbf{v} = \mathbf{v}_D$ on Γ_D.*
- *For sufficiently small Re and Pe, and small prescribed boundary values T_D and $\mathbf{v}_D$, the Navier-Stokes problem* (1.49) *admits an unique solution.*
- *$\exists\, C = \mathrm{const.}$, depending on Ω and Γ_D, such that any solution $(\mathbf{v}, p)$ of the Navier-Stokes problem* (1.49) *satisfies*

$$\|\mathbf{v}\|_{1,\Omega} \leq C \left(\|\mathbf{v}_D\|_{1/2,\Gamma_D} + \frac{\mathrm{Re}}{\mu_1} \|\mathrm{v}_D\|^2_{1/2,\Gamma_D} + \frac{\mathrm{Gr}}{\mu_1 \,\mathrm{Re}} \left\|\hat{T}\right\|_{0,\Omega} + \frac{\mathrm{Re}}{\mu_1} \|\mathrm{F}\|_{0,\Omega} \right) \tag{1.52}$$

and

$$\|p\|_{0,\Omega} \leq C \left(\mu_2 \|\mathrm{v}_D\|^2_{1/2,\Gamma_D} + \mathrm{Ri} \left\|\hat{T}\right\|_{0,\Omega} + \|\mathbf{F}\|_{0,\Omega} + \mu_2 \|\mathbf{v}\|^2_{1,\Omega} \right). \tag{1.53}$$

Remarks:

- In our case, we neglect the buoyancy term (Gr = 0 and Ri = 0), what in (1.52) results in uniform bound for $\mathbf{v}$, independently of the "frozen" temperature field $\hat{T}$. Both these facts then establish uniform bound for pressure. The uniform bound of velocity also implies uniform bound for the temperature field. The constant property problem without the buoyant term was analysed in former work of (Bruneau and Fabrie [3]). It can be shown, that the existence and uniqueness in case of variable property and no-buoyant flow follows from the same technique as in the constant property case.

- The solution of the coupled problem is finally approximated by the decoupled sub-problems using outer Picard iteration in the thermophysical properties (μ, κ)
$$\hat{T} \rightarrow (\mathbf{v}(\hat{T}), p(\hat{T})) \rightarrow T(\mathbf{v}(\hat{T}))\,. \tag{1.54}$$
It is also shown in (Pérez [30]), that (1.54) allows a fixed point even for the problem with the linear approximation of buoyancy.

- Modification of the convective terms in both the Navier-Stokes equations and Energy equation facilitates the proof of solution existence, but also motivates further discussion especially on the treatment of the outflow boundary condition (sec. 2.1.2). In other words, the modified form of the convective operator in weak form (1.45) leads to the *directional do-nothing* boundary condition discussed in (Braack [2]). Physically realistic situations, however, exhibit the reentrant flow, which is partially allowed by construction of (Dong [6], c.f. sec. 2.1.2).

2. Discretisation methods

This chapter concerns discretisation of the incompressible Navier-Stokes (INS) system coupled with the heat equation (HE), (1.32)-(1.34), where the temperature influence of the velocity field is solely given by the change in temperature dependent viscosity and the buoyancy is neglected. Unlike the buoyancy effects, which were successfully computed under Boussinesque approximation (e.g. Ren [36]) and act as a forcing term in the momentum equation, temperature dependent viscosity and thermal conductivity affect the second order differential operators, which are no more constant in time.

Two of the main contributions of this work are presented in this chapter. First is the construction of a reasonably fast algorithm for the systems coupled by temperature dependent properties, which allows us to use a high order method in space.

Second contribution is in application of a high order polynomial approximation to the spatial part of the solution. The low order methods approximate a function in a piecewise manner due to decomposition of the computational domain Ω to smaller parts. This is in contrast to the high order methods, which construct a global approximation, while keeping functions mathematical properties over the whole or substantial part of the computational domain. If the domain must be divided for some reason (e.g. complicated geometry of Ω), the high order approximation substantially lowers the number of elements Ω_e, needed to cover Ω ($\Omega \approx \cup_e \Omega_e$). High order approximation also carries information about (higher) functions derivatives. Rate of convergence to the exact solution is controlled by regularity of the approximated function, when the convergence rate is not limited for smooth functions, what again stays in contrast to the low order methods (sec. 2.2).

The chosen discretisation approach states the temporal discretisation as primal, what coincides with already presented mathematical analysis. This approach formulates the problem as a sequence of steady state sub-problems (c.f. sec. 1.2.1). The differential operator of the coupled problem is split on multiple levels. As a result of the temporal discretisation a set of linear differential equations is constructed and solved as separate boundary value problems, employing the high order method in space.

The problems solved in temporal and spatial discretisation differs fundamentally and will be described in a separate sections of this chapter.

2.1 Temporal discretization

In this section, we present methods for discrete solution of ordinary differential equations. Those appropriate for the computational algorithm of the system (1.32)-(1.34) are discussed in detail. The general structure of the time discretisation schemes is presented to establish terminology and idea of the General Linear Method, which has important consequences in programming techniques and is described in sec. 2.1.3 in more detail.

An ordinary differential equation (ODE) in time, written in the implicit form

$$F(t, u, \dot{u}) = 0 \tag{2.1a}$$

$$u_0 = u(0)\,, \tag{2.1b}$$

encompasses both the momentum (1.32) and energy (1.34) equations, which are in form

$$\dot{u} = f(t, u)\,. \tag{2.2}$$

We denote the solution $u = u(t)$, $\dot{u}$ its time derivative and f a function, which may include a differential operator in space coordinates. The initial condition (IC) denoted by u_0 completes the problem.

Solution methods for ODEs form two basic families, the *one-step/multi-stage* methods and *multi-step* methods.

Higher order **one-step** methods are equivalently noted as *multi-stage*, since the time step is divided internally to multiple stages. An M-stage method can be written in form

$$F(t_{n-1} + c_i\,\Delta t, u_{n-1} + \Delta t \sum_{j=1}^{M} a_{ij}\dot{Y}_j, \dot{Y}_i) = 0\,, \tag{2.3}$$

where $\dot{Y}_i$ are estimates for $\dot{u}(t_{n-1} + c_i\Delta t)$ called the *stage derivatives*. The terms $u_{n-1} + \Delta t \sum_{j=1}^{M} a_{ij}\dot{Y}_j \equiv Y_i$ are the *stage values*, estimates to $u(t_{n-1} + c_i\,\Delta t)$. Coefficients c_i and a_{ij} are method-specific. Into this group belong the widely used Runge-Kutta schemes. Its advantage is in simple initialisation, which requires knowledge of only one previous time level of solution (possibly initial condition), regardless the order of the method. They are recommended for algorithms, where the scheme is frequently restarted (e.g. multigrid methods). This kind of methods is not used in computations presented in this work, but we can meet with results concerning spectral element method with temporal discretisation as composed of mixture of a multi-step method and Runge-Kutta schemes (see Karniadakis [25], pg. 428).

The **multi-step** methods omit evaluations on the stage level, but take information from the multiple steps of computational history. All the computed values are used repeatedly and only evaluation of the new time level quantities is performed, regardless the order of the method. This significantly reduce the computational cost. Initialization of an N-step scheme with $N > 1$ employs a sequence of $1, 2, \ldots, N-1$-step methods or the $N-1$ values are calculated using a multi-stage method. Generally, the multi step methods are not recommended for algorithms performing frequent restart. Representatives of multi-step methods are especially the *Adams methods* and *backward differentiation formulae* (BDF). We will use an equidistant time stepping, but schemes with variable time step are available (see Appendix E).

Stiffly stable methods

The particular steps of our computational algorithm will involve the second order diffusion operator in the spatial coordinates. However, discretisation of the diffusion operator employing the high order/spectral methods leads to the *stiff problem*, what may cause instability of some time-integration methods.

Definition 1. *(Stiff problem)*
A linear differential system

$$\dot{u} = Au + \phi(x) \tag{2.4}$$

where $A \in \mathcal{R}^{n\times n}$, u, $\phi \in \mathcal{R}^n$ *is said to be stiff if and only if*

- *For all* $i, \Re\lambda_i < 0$,
- *(Stiffness ratio)* $\dfrac{max\,|\Re\lambda_i|}{min\,|\Re\lambda_i|} >> 1$

where λ_i, $i = 1, ..., n$ *are eigenvalues of* A.

It is worthy to note, that stiffness is exhibited by more general problems than (2.4). The above definition fits to the case of the heat equation

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2 u(x,t)}{\partial x^2}$$

$$u(0,t) = u(1,t) = 0$$

$$u(x,0) = g(x)$$

As noted in almost all works about the spectral methods the eigenvalues of the diffusion operator discretised through N-th order polynomial basis grow as $O(N^4)$ and are negative.

Following the definition 1, a method is *stiffly stable* if it is accurate around origin for all the components in the stability diagram and absolutely stable in the half complex plane with a negative real component. Stability properties of these methods remain almost constant, while the accuracy of integration (together with the order) increases. The coefficients of stiffly stable schemes were already found up to 11-th order.

Definition 2. *(Stiffly stable method; Gear [13])*
A method is stiffly stable if in the region

$$R_1 := \{z \in \mathcal{C};\, \Re(z) \leq a < 0\} \tag{2.5}$$

it is absolutely stable, and in

$$R_2 := \{z \in \mathcal{C};\, a < \Re(z) < b, |\Im(z)| < c\}, \tag{2.6}$$

it is accurate.

In the above theorem $a, b, c \in \mathcal{R}$, $a < 0 < b, c$.

Theorem concerning existence of the stiffly stable schemes is provided in Gear [13]

Theorem 3. *(Existence of stiffly stable methods)*

- *The k-step methods of order k-1 with characteristic polynomial* $\sigma(\xi) = \beta_0\xi^k$ *are stiffly stable for* $k = 1, \ldots, 6$ *(These methods are backward differentiation formulae, see folowing paragraph).*
- *Above methods are not stiffly stable for* $k = 7, \ldots, 15$.
- *There are stiffly stable multistep methods of orders up to 11.*

The characteristic polynomial of a general k-step method

$$\sum_{q=0}^{k}(\alpha_q u_{n-q} + h\beta_q f_{n-q}) = 0$$

is $\rho(\xi) + h\lambda\sigma(\xi) = 0$, where $\rho(\xi) = \sum_{q=0}^{k} \alpha_q \xi^{k-q}$ and $\sigma(\xi) = \sum_{q=0}^{k} \beta_q \sigma^{k-q}$.

The highest order of stiffly stable Adams-Moulton method is 2. Higher order schemes (order ≤ 6), which exhibit stability against the stiff problems, are based on the *backward differentiation formulae* (BDF). These are all implicit and use the values of multiple old solution values $\{u_{n-q}\}_{q=0}^{Q-1}$. BDF's are based on multi-step approximation of the temporal derivative in (2.2) at the node t_{n+1}, so it can be generally written as

$$\sum_{q=0}^{Q} \alpha_q u_{n+1-q} = \Delta t\, f_{n+1} \tag{2.7}$$

or with requirement $\alpha_0 = 1$ as

$$\sum_{q=0}^{Q} \alpha_q u_{n+1-q} = \Delta t\, \beta_0\, f_{n+1}\,,$$

where $f_{n+1} = f(t_{n+1}, u_{n+1}))$. In comparison to the Adams methods, which have the simplest first characteristic polynomial $\rho(\xi) = \xi^n - \xi^{n-1}$, BDF have a simplest second characteristic polynomial $\sigma(\xi) = \dfrac{1}{\alpha_0}\xi^n$ and form a complementary approach to the Adams methods.

To derive the coefficients α_q the Lagrange interpolation polynomial of $u(t)$ is expressed using the *backward differences* $\nabla u_n = u_n - u_{n-1}$. Having Q points $\{[t_{n+1-q}, u_{n+1-q}]\}_{q=0}^{Q-1}$ available, we obtain

$$u(t) \approx \Pi_u^{Q-1} = u_{n+1} + \frac{(t-t_{n+1})}{\Delta t}\nabla u_{n+1} + \frac{(t-t_{n+1})(t_{n+1}-t_n)}{2\,\Delta t^2}\nabla^2 u_{n+1} + \dots$$
$$\dots + \frac{(t-t_{n+1})(t_{n+1}-t_n)\dots(t_{n+1}-t_{n+1-(Q-1)})}{(Q-1)!\,\Delta t^{Q-1}}\nabla^{Q-1}\, u_{n+1}$$

where $\nabla^q u_n = \nabla(\nabla^{q-1} u_n)$. Coefficients of the approximation then follow from the derivative of this interpolant $\dfrac{\partial u}{\partial t} \approx \dfrac{\mathrm{d}\Pi_u^{Q-1}}{\mathrm{d}t} = \dfrac{\sum_{q=0}^{Q-1} \alpha_q\, u_{n+1-q}}{\Delta t}$.

As mentioned in Theorem 3, the BDF are zero-stable for $Q \leq 5$ (see also [34]). Its properties suit the stiff stability definition 2. Only the first order BDF is unconditionally stable[1]. The higher-order BDF schemes include a small region around the imaginary axis, where they are unstable. Due to this stability properties they are suitable for diffusion equations, but may become unstable in case of convective operators.

Detail discussion and illustrations of the stability regions for BDF may be found in (Gear [13], Karniadakis [25] or Canuto [8]).

[1]First order BDF coincides with the first order Adams-Moulton scheme, which is often denoted as Euler Implicit scheme

k	1	2	3	4	5	6
α_0	1	1	1	1	1	1
α_1	-1	-4/3	-18/11	-48/25	-300/137	-360/147
α_2		1/3	9/11	36/25	300/137	450/147
α_3			-2/11	-16/25	-200/137	-400/147
α_4				3/25	75/137	225/147
α_5					-12/137	-72/147
α_6						10/147
β_0	1	2/3	6/11	12/25	60/137	60/147

Table 2.1: Coefficients of the k-step Backward-Differencing schemes (normalized to $\alpha_0 = 1$).

If the spatial differential operators in f_{n+1} are time dependent (convection, temperature dependent diffusive terms), the implicit evaluation employing matrix inversion, if possible, is not effective. However, an explicit evaluation of f_{n+1} in the BDF scheme is possible (see also Karniadakis [23] or Peyret [32]). Requirement of consistency restricts the coefficients of the BDF scheme

$$\alpha_0 = \sum_{q=1}^{Q-1} \alpha_q \,, \tag{2.8}$$

therefore

$$\begin{aligned} &\sum_{q=0}^{Q-1} \alpha_q \, u_{n+1-q} = \alpha_0 \, u_{n+1} - \sum_{q=1}^{Q-1} \alpha_q \, u_{n+1-q} = \sum_{q=1}^{Q-1} \alpha_q (u_{n+1} - u_{n+1-q}) = \\ &= \sum_{q=1}^{Q-1} \alpha_q \int_{t_{n+1-q}}^{t_{n+1}} \frac{\partial u}{\partial t} \mathrm{d}t = \sum_{q=1}^{Q-1} \alpha_q \int_{t_{n+1-q}}^{t_{n+1}} f(t, u(t)) \mathrm{d}t \,. \end{aligned} \tag{2.9}$$

Evaluating the integrals explicitly $\int_{t_i}^{t_{n+1}} f(t, t(t)) \mathrm{d}t \approx \Delta t \, f(t_n, u_n)$ and due to the uniformly increasing integration length $t_{n+1} - t_{n+1-q} = q\Delta t$ in the summed integrals in (2.9), we arrive to coefficients for explicit approximation

$$f_{n+1} \approx \sum_{q=1}^{Q-1} \beta_q f_{n+1-q} \,, \tag{2.10}$$

where $\beta_q = q \, \alpha_q$. Above formula represents a $Q-1$ extrapolation and its coefficients are summarized in the table 2.2.

IMEX schemes

As already noted, equations in the system (1.32)-(1.34) take form $\dot{u} = f(t, u)$, where f includes differential operators in the spatial coordinates. f then sums both the linear, constant in time operators f_L, which allow implicit evaluation,

Coefficient	1st order	2nd order	3rd order	4th order
β_0	1	2	3	4
β_1		-1	-3	-6
β_2			1	4
β_3				-1

Table 2.2: Extrapolation coefficients belonging to the BDFs (the values are derived from coefficients in table 2.1).

Coefficient	1st order	2nd order	3rd order	4th order
γ_0	1	3/2	11/6	25/12
α_0	1	2	3	4
α_1	0	-1/2	-3/2	-3
α_2	0	0	1/3	4/3
α_3	0	0	0	-1/4
β_0	1	2	3	4
β_1	0	-1	-3	-6
β_2	0	0	1	4
β_3	0	0	0	-1

Table 2.3: Coefficients of mixed stiffly-stable schemes ([25], [32]).

but also those non-linear and time dependent, f_N. The BDF scheme then results in a mixed form, denoted as *implicit-explicit* (IMEX)

$$\frac{\gamma_0 u_{n+1} - \sum_{q=0}^{Q-1} \alpha_q u_{n-q}}{\Delta t} = f_L(u_{n+1}) + \sum_{q=0}^{Q-1} \beta_q f_N(u_{n-q})\,, \tag{2.11}$$

since it consists of implicit problem

$$\frac{\gamma_0 u_{n+1} - \sum_{q=0}^{Q-1} \alpha_q u_{n-q}}{\Delta t} = f_L(u_{n+1})$$

enhanced of explicitly evaluated term $\sum_{q=0}^{Q-1} \beta_q f_N(u_{n-q})$. Coefficients of this type schemes are summarized in table 2.3.

Formulation of the IMEX schemes coincide with the *operator splitting* approach described in the Appendix B.

2.1.1 Time stepping algorithm

Decoupling of the momentum and heat equation

Introducing solution vector $\mathbf{u} = \begin{pmatrix} \mathbf{v} \\ T \end{pmatrix}$, whole system (1.32)-(1.34) may be written in form of initial value problem (IVP)

$$\begin{aligned} \frac{\partial \mathbf{u}}{\partial t} &= \mathbf{A}(\mathbf{u}) \\ \nabla \cdot \mathbf{u} &= 0 \end{aligned} \tag{2.12}$$

$$\mathbf{B}(\mathbf{u}) = 0\,, \quad \mathbf{u}(0, \vec{x}) = \mathbf{g}(\vec{x})\,, \quad \vec{x} \in \Omega\,, \quad t \in [0, T]\,,$$

where $\mathbf{B}$ is a boundary operator defining the boundary conditions and $\mathbf{A}(\mathbf{u}) = \begin{pmatrix} \mathbf{M}(\mathbf{u}) \\ H(\mathbf{u}) \end{pmatrix}$ represents a non-linear operator, which couples both the momentum

$$\mathbf{M}(\mathbf{v}, T) = -\mathbf{v} \cdot \nabla \mathbf{v} - \nabla p + \nabla \cdot \left[\nu(T)(\nabla \mathbf{v} + (\nabla \mathbf{v})^{\mathrm{T}})\right] + \mathbf{f}_M \tag{2.13}$$

and the heat

$$H(\mathbf{v}, T) = -\mathbf{v} \cdot \nabla T + \nabla \cdot (\kappa(T) \nabla T) + f_T \tag{2.14}$$

operators. The functions $\mathbf{f}_M = \begin{pmatrix} f_1 \\ \vdots \\ f_D \end{pmatrix}$ and f_T are known in (2.13) and (2.14). More precisely

$$\mathbf{A}(\mathbf{u}) = \begin{pmatrix} -\mathbf{v} \cdot \nabla v_1 - \dfrac{\partial p}{\partial x_1} + \nabla \cdot [\nu(T)(\nabla v_1 + \dfrac{\partial \mathbf{v}}{\partial x_1})] + f_1 \\ \vdots \\ -\mathbf{v} \cdot \nabla v_D - \dfrac{\partial p}{\partial x_D} + \nabla \cdot [\nu(T)(\nabla v_D + \dfrac{\partial \mathbf{v}}{\partial x_D})] + f_D \\ -\mathbf{v} \cdot \nabla T + \nabla \cdot [\kappa(T) \nabla T] + f_T \end{pmatrix}, \tag{2.15}$$

where we denote

$$\frac{\partial \mathbf{v}}{\partial x_i} = \begin{pmatrix} \dfrac{\partial v_1}{\partial x_i} \\ \vdots \\ \dfrac{\partial v_D}{\partial x_i} \end{pmatrix} \quad \forall i = 1, \ldots, D\,.$$

Reasons of computational efficiency motivate us to decouple the system to a sequence of linear equations with constant coefficients. We use the idea of the "frozen" values of both the velocity and temperature field as described in section 1.2.1. Values at time t_n, or other preceding time steps if higher order extrapolation is employed, are used for evaluation of the terms, which couple the equations.

This is the generalized diffusion operator

$$\nabla \cdot [\nu(T)(\nabla \mathbf{u} + (\nabla \mathbf{u})^{\mathrm{T}})] \tag{2.16}$$

in momentum equation and advection term

$$\mathbf{v} \cdot \nabla T \tag{2.17}$$

in the heat equation[2].

The viscous term (2.16) will be evaluated as in (Karamanos [22]), where the viscosity is divided to the constant $\bar{\nu}$ part and time dependent $\nu_s = \nu_s(t)$ part

$$\nu(T) = \bar{\nu} + \nu_s \,. \tag{2.20}$$

Then for all the velocity components ($i = 1, \ldots, D$), the viscous term splits to the linear part $\mathbf{D}_L$

$$(\mathbf{D}_L)_i(\mathbf{u}) = (\mathbf{D}_L)_i(\mathbf{v}) = \bar{\nu}\nabla^2 v_i \,,$$

which is treated implicitly and time dependent, non-linear term $\mathbf{D}_N$

$$(\mathbf{D}_N)_i(\mathbf{v}) = \nabla \nu_s \cdot (\nabla v_i + \frac{\partial \mathbf{v}}{\partial x_i}) + \nu_s \nabla^2 v_i \tag{2.21}$$

evaluated explicitly, using the known values from previous steps or the initial condition. Term (2.21) is already simplified by imposing $\nabla \cdot \mathbf{v} = 0$.

Note, that $\bar{\nu}$ is possibly variable in space $\bar{\nu} = \bar{\nu}(\vec{x})$. Successful computations were performed both taking $\bar{\nu} = \nu_\infty$ and $\bar{\nu}$ as the time average of values $\nu(T)$ taken from similar computation.

All the forcing terms (2.18), (2.19) and advective term in the heat equation (2.17) will be evaluated explicitly[3].

Now, $\mathbf{v}_{n+1} = \mathbf{v}(t_{n+1})$ and $T_{n+1} = T(t_{n+1})$ are evaluated independently in the algorithm and the system (2.12) separates to momentum equation

$$\begin{aligned} \frac{\partial \mathbf{v}}{\partial t} &= -\mathbf{v} \cdot \nabla \mathbf{v} - \nabla p + \bar{\nu}\nabla^2 \mathbf{v} + \mathbf{f}_M + \mathbf{D}_N(\mathbf{v}) \\ \nabla \cdot \mathbf{v} &= 0 \end{aligned} \tag{2.22}$$

[2]Functions $\mathbf{f}_M$ and f_T, however, may cause the coupling too. For example, the buoyancy term from the Boussinesque approximation (1.8) couples the momentum to the heat equation since

$$\mathbf{f}_M = -\mathrm{Ri}\, T\mathbf{g} \,. \tag{2.18}$$

In opposite direction, coupling of the heat to the momentum equation emerges in case of viscous heating

$$f_T = \frac{\mathrm{Ec}}{\mathrm{Re}}\, \mathbb{D} : \mathbb{D} \,. \tag{2.19}$$

[3]Forestier ([12]) studied the technique of splitting the velocity field in the convective term to constant $\bar{\mathbf{v}}$ and time-dependent $\mathbf{v}_s$ parts

$$(\bar{\mathbf{v}} + \mathbf{v}_s) \cdot \nabla \mathbf{v}$$

on the system of Navier-Stokes equations. This idea, coinciding with splitting the viscosity (2.20) is applicable also to the term (2.17), obtaining part solvable by linear solver and explicitly evaluated temporal deviation.

and

$$\frac{\partial T}{\partial t} = \nabla(\kappa(t)\nabla T) + f_T + \mathbf{v}\cdot\nabla T\,. \tag{2.23}$$

The operator splitting technique may be used to both the equations independently, arriving to the incompressible Navier-Stokes equations and non-linear heat equation.

Incompressible Navier-Stokes system

Both the $\mathbf{f}_M$ and $\mathbf{D}_N(\mathbf{v})$ are known and we arrive to the system

$$\begin{aligned} \frac{\partial \mathbf{v}}{\partial t} + \mathbf{M}(\mathbf{v}, T_n) &= \mathbf{0} \\ \nabla\cdot\mathbf{v} &= 0\,, \end{aligned} \tag{2.24}$$

which has the form of Incompressible Navier-Stokes equations with a "source" term $\mathbf{s} = \mathbf{f} + \mathbf{D}_N(\mathbf{v})$

$$\frac{\partial \mathbf{v}}{\partial t} + \mathbf{v}\cdot\nabla\mathbf{v} = -\nabla p + \bar{\nu}\nabla^2\mathbf{v} + \mathbf{s} \tag{2.25}$$

and coincides with the formulation with "frozen" temperature field presented in section 1.2.1. Now, the velocity components in the incompressible Navier-Stokes equations are still coupled in both the convective term $\mathbf{v}\cdot\nabla\mathbf{v}$ and the constraint $\nabla\cdot\mathbf{v} = 0$. In this part of the algorithm, we follow the approach of velocity-correction splitting scheme presented in (Karniadakis [23]). The operator splitting technique both decouples the variables and linearise the non-linear operators. The splitting is prior to the spatial discretisation, what is noted as *differential splitting* in literature and contrasts with *algebraic splitting* which acts to algebraic system emerging from spatial discretisation. This technique and its other variants are described in Appendix B and some recent schemes for the incompressible Navier-Stokes equations, based on this approach, are compared in Appendix C.

In view of operator splitting, two sub-problems are constructed

$$\frac{\partial \mathbf{v}^{(1)}}{\partial t} = -\mathbf{v}^{(1)}\cdot\nabla\mathbf{v}^{(1)} + \mathbf{s} \tag{2.26}$$

and the Stokes problem

$$\begin{aligned} \frac{\partial \mathbf{v}^{(2)}}{\partial t} &= -\nabla p + \bar{\nu}\nabla^2\mathbf{v}^{(2)} \\ \nabla\cdot\mathbf{v}^{(2)} &= 0 \end{aligned} \tag{2.27}$$

Backward difference formula approximating the time derivative combined with extrapolation of the convection term is used in the first sub-step (2.26)

$$\frac{\gamma_0\mathbf{v}^{(1)}_{n+1} - \sum_{q=0}^{J_i}\alpha_q\mathbf{v}_{n-q}}{\Delta t} = -\sum_{q=0}^{J_e}\beta_q(\mathbf{v}_{n-q}\cdot\nabla\mathbf{v}_{n-q} + \mathbf{s}_{n-q})\,. \tag{2.28}$$

Coefficients α_q and β_q correspond to those in the table 2.3. The source term $\mathbf{s}$ includes the variations in the velocity fields induced by the change of temperature.

Nevertheless the initial condition(s) of this step $\mathbf{v}_{n-q}$ may be solenoidal, the result $\mathbf{v}^{(1)}$ is generally not, since the divergence operator do not commute with the convection.

After explicit treatment of the nonlinearity, the Stokes problem (2.27) remains to be solved. The projection method will finally decouple the velocity components, while keeping the incompressibility condition $\nabla \cdot \mathbf{v} = 0$.

Stokes problem-Projection method

The Stokes problem can be solved for the velocity field without approximating the pressure field and there is no equation of state needed to complete the system. For more, this system of equations do not prescribe the time evolution of the pressure. Therefore the pressure looses its meaning of the state variable and gets more mathematical essence of variable enabling the velocity field to belong to the space of solenoidal functions.

Postulation of incompressibility is moderately significant approximation in view of physics of slow fluid flows, but has striking impact to construction of computational scheme. This results from the orthogonal decomposition of the space $\boldsymbol{\mathcal{L}}^2(\Omega)$ to the solenoidal

$$S(\Omega) = \{\mathbf{u} \in \boldsymbol{\mathcal{L}}^2(\Omega), \nabla \cdot \mathbf{u} = 0 \,\text{in}\, \Omega, \mathbf{u} \cdot \mathbf{n} = 0 \,\text{on}\, \partial\Omega\}$$

and irrotational part

$$R(\Omega) = \{\mathbf{w} \in \boldsymbol{\mathcal{L}}^2(\Omega), \mathbf{w} = \nabla\phi\}\,.$$

Velocity $\mathbf{v}^{(2)}$ in (2.27) is solution to the Stokes problem, which is divided into two steps in the projection method and may be seen as further operator splitting with sub-solutions $\mathbf{v}^{(2,1)}$ and $\mathbf{v}^{(2,2)}$ such, that

$$\frac{\partial \mathbf{v}^{(2,1)}}{\partial t} = -\nabla p_{n+1} \tag{2.29}$$

$$\frac{\partial v_i^{(2,2)}}{\partial t} = \nabla^2 v_i \quad i = 1, \ldots, D \tag{2.30}$$

and $\mathbf{v}_n^{(2,1)} = \mathbf{v}_{n+1}^{(1)}$, $\mathbf{v}_n^{(2,2)} = \mathbf{v}_{n+1}^{(2,1)}$, $\mathbf{v}_{n+1}^{(2)} = \mathbf{v}_{n+1}^{(2,2)}$. The purpose of writing (2.30) in non-vector form is to highlight the fact, that we get equations for the separate velocity components, so the system is now fully decoupled.

Concerning the scheme from (Karniadakis [23]), both the sub-steps are solved using Euler Implicit scheme. Equation (2.29) becomes

$$\frac{\mathbf{v}_{n+1}^{(2,1)} - \mathbf{v}_{n+1}^{(1)}}{\Delta t} = -\nabla p_{n+1}\,. \tag{2.31}$$

Divergence applied to equation (2.31) together with $\nabla \cdot \mathbf{v}_{n+1}^{(2,1)} = 0$ represents the projection step and results in the *pressure-Poisson* equation

$$\nabla^2 p_{n+1} = \nabla \cdot \mathbf{v}_{n+1}^{(1)} \tag{2.32}$$

Finally, the "heat equation" for velocity components, discretized by implicit Euler method reads

$$\frac{\mathbf{v}_{n+1}^{(2,2)} - \mathbf{v}_{n+1}^{(2,1)}}{\Delta t} = \nabla^2 \mathbf{v}_{n+1}^{(2,2)} \tag{2.33}$$

Evaluation of $\mathbf{v}_{n+1}^{(2,1)}$ follows from (2.31)

$$\mathbf{v}_{n+1}^{(2,1)} = -\Delta t \nabla p_{n+1} + \mathbf{v}_{n+1}^{(1)} \,. \tag{2.34}$$

The equation evaluated implicitly, using a spatial solver, forms the *Helmholtz* equation for all velocity components

$$\left(\nabla^2 - \frac{1}{\bar{\nu}\Delta t}\right)(v_i)_{n+1}^{(2,2)} = \frac{-1}{\bar{\nu}\Delta t}(v_i)_{n+1}^{(2,1)} \quad i = 1, \ldots, D \,. \tag{2.35}$$

Remark: In view of the Helmholtz decomposition, the equation (2.34) correlates elements of the solenoidal space, $\mathbf{v}_{n+1}^{(2,1)}$, irrotational space, $\Delta t \nabla p$, and result of the first sub-step, $\mathbf{v}_{n+1}^{(1)}$. In this sense, pressure p is such, that $\Delta t \nabla p$ represents that part of $\mathbf{v}_{n+1}^{(1)}$, which is not solenoidal. Subtracting this term from $\mathbf{v}_{n+1}^{(1)}$ results in solenoidal velocity $\mathbf{v}_{n+1}^{(2,1)}$, which is the initial condition for the second sub-problem (2.33). However, this initial condition stays in the RHS of the discretised problem, and if the boundary condition prescribed to the sub-problem is not "consistent" with this RHS (=initial condition), a steep gradient occurs in vicinity of $\partial\Omega$. C.f. section 2.1.2.

Non-linear heat equation

The equation for temperature with "frozen" velocity field, takes form

$$\frac{\partial T}{\partial t} + H(\mathbf{v}_n, T) = 0 \,. \tag{2.36}$$

after decoupling the whole system.

However, the generalized (anisotropic) diffusion operator

$$\nabla \cdot (\kappa(T) \nabla T) \tag{2.37}$$

is still strongly nonlinear in T. Nature of this term is different from (2.16), but the linearisation technique follow the same process as in the case of the viscous term. We introduce the structure of $\kappa(T)$

$$\kappa(T) = \bar{\kappa} + \kappa_s \,, \tag{2.38}$$

where $\kappa_s = \kappa_s(\vec{x}, t)$, while $\bar{\kappa} = \text{const.}$ or $\bar{\kappa} = \bar{\kappa}(\vec{x})$.

Operator splitting then results in two sub-problems

$$\frac{\partial T^{(1)}}{\partial t} = \nabla \kappa_s \cdot \nabla T^{(1)} + \kappa_s \nabla^2 T^{(1)} + f_H \tag{2.39}$$

$$\frac{\partial T^{(2)}}{\partial t} = \bar{\kappa} \nabla^2 T^{(2)} \,, \tag{2.40}$$

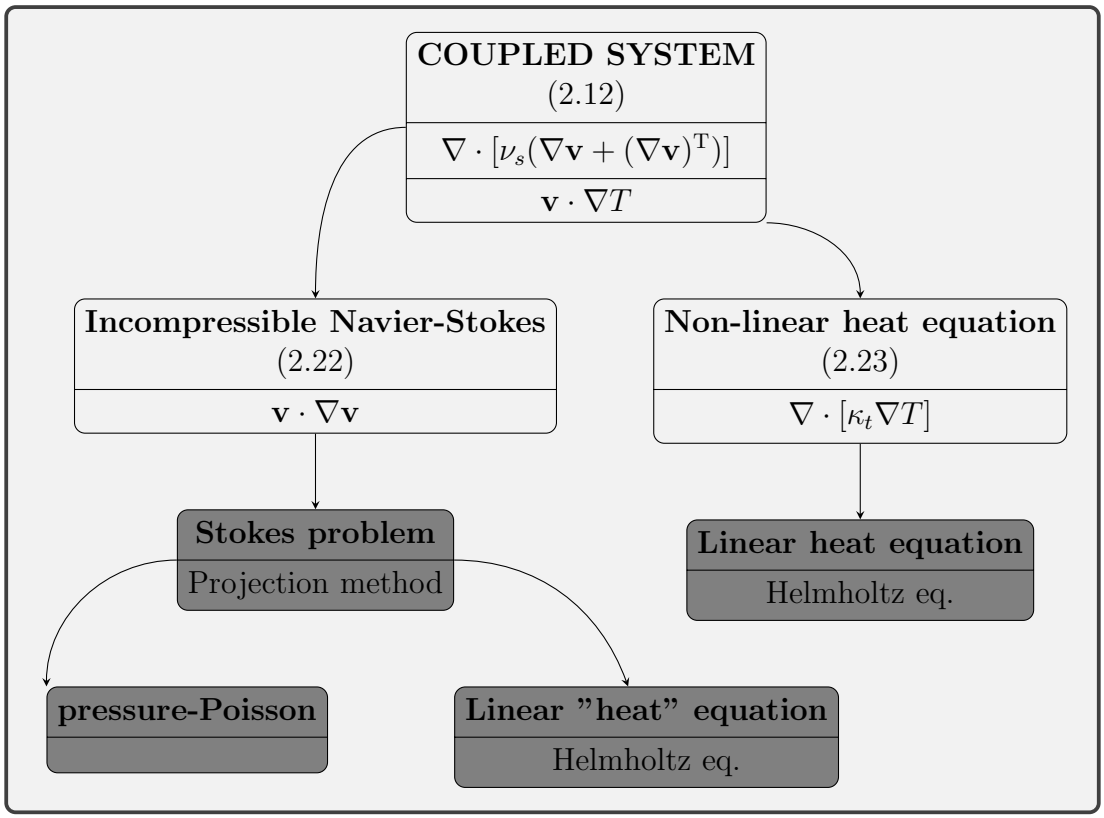


Figure 2.1: Scheme of decoupling the incompressible Navier-Stokes-Fourier system. The dark-grey blocks are solved by implicit method, while in the others an explicit evaluation of the presented terms is performed.

where the first contains the time dependent part of the diffusion operator, which is extrapolated, and the second sub-problem is linear constant-coefficient problem, solved by implicit method. The source term f_H represents now the advection term, which is evaluated through extrapolation in this step as a consequence of decoupling the momentum and heat equation.

Applying the backward difference scheme with extrapolation of the RHS, we get discretised version of the first sub-problem (2.39)

$$\frac{\gamma_0 T_{n+1}^{(1)} - \sum_{q=0}^{J_i-1} \alpha_q T^{n-q}}{\Delta t} = \sum_{q=0}^{J_e} \beta_q [\nabla \cdot \kappa_s \nabla T - \mathbf{v} \cdot \nabla T]_{n-q} \tag{2.41}$$

Implicit Euler method applied to the second sub-problem results in Helmholtz-type equation solved implicitly using a spatial solver

$$\left(\nabla^2 - \frac{1}{\bar{\kappa}\Delta t}\right) T_{n+1}^{(2)} = -\frac{1}{\bar{\kappa}\Delta t} T_{n+1}^{(1)} \,. \tag{2.42}$$

2.1.2 Boundary conditions

The boundary conditions (BC) play a crucial role both in definition of the intermediate steps of the computational algorithm and global setting of the physical model. We will show the problems, following from the imprecise setting of BC

inside the algorithm and collect some recent results for the open question of outflow BC. Attention will be paid also to compatibility between the BC and initial condition.

Every implicit step (2.32), (2.35) or (2.42), form an elliptic PDE and a boundary conditions must be specified to complete the problem. Setting of these boundaries define the model in view of mathematics. The solution than may be obtained for relatively arbitrary choice of BC, independently of particular definition of the equations RHS. However, not every such a solution may refer to a real physical problem as illustrates following 1D example.

We generate the RHS belonging to an exact solution u, e.g. $u = cos(x)$, for the Poisson equation (2.32 in 1D)

$$\frac{\mathrm{d}^2 u}{\mathrm{d}x^2} = f \rightarrow f = \frac{\mathrm{d}^2 cos(x)}{\mathrm{d}x^2} = -\cos(x)$$

and the Helmholtz equation (2.35 in 1D)

$$\left(\frac{\mathrm{d}^2}{\mathrm{d}x^2} - \lambda\right) u = f \rightarrow f = \left(\frac{\mathrm{d}^2}{\mathrm{d}x^2} - \lambda\right) \cos(x) = -(1+\lambda)\cos(x)\,.$$

Both the problems admit setting of boundary conditions, independently to the function f. However, we derived the function f from the "known" solution and the boundary condition for the known solution satisfy naturally $u|_{\partial\Omega} = cos|_{\partial\Omega}$. Solving the problem with accurate RHS and accurate BC leads to exact solution if sufficiently accurate method is employed (c.f. 2.2). However, the situation in the real computation is different, we prescribe a boundary condition resulting from considerations of physics and math. analysis, but the RHS is a result of a preceding sub-problem.

The inconsistency of BC and RHS influences differently the solution to the Poisson and the Helmholtz equation.

In case of the Poisson equation, the solution is influenced in the whole computational domain. However, the resulting function has similar demands to the approximation space as the function f [4]. But this is not the case of Helmholtz equation, where a "boundary layer" of thickness dependent on value of coefficient λ occurs. In 1D problem, the "boundary layer" refers to the solution of homogeneous problem, which has a known form

$$u_H = c_1 \exp(-\sqrt{\lambda}x) + c_2 \exp(\sqrt{\lambda}x)\,,$$

where c_1, c_2 are constants and the solution may be written as

$$\tilde{u} = u + u_H\,.$$

Comparison in the influence of the "RHS-inconsistent" boundary condition for the Poisson and Helmholtz problem is illustrated in figure 2.2 for $u = cos(x)$, $\lambda = 100$, $\tilde{u}(x_L) = u(x_L)$ and $\tilde{u}(x_R) \neq u(x_R)$ on interval $x \in [x_L, x_R]$.

It is also shown in figure 2.2, how the coefficient λ influences the function near the boundary. Among examples of the spectral methods (sec. 2.2.5), it will be

[4] If the spectral method is used for spatial approximation, same (or very similar) number of basis functions is needed to approximate both functions to a given precision

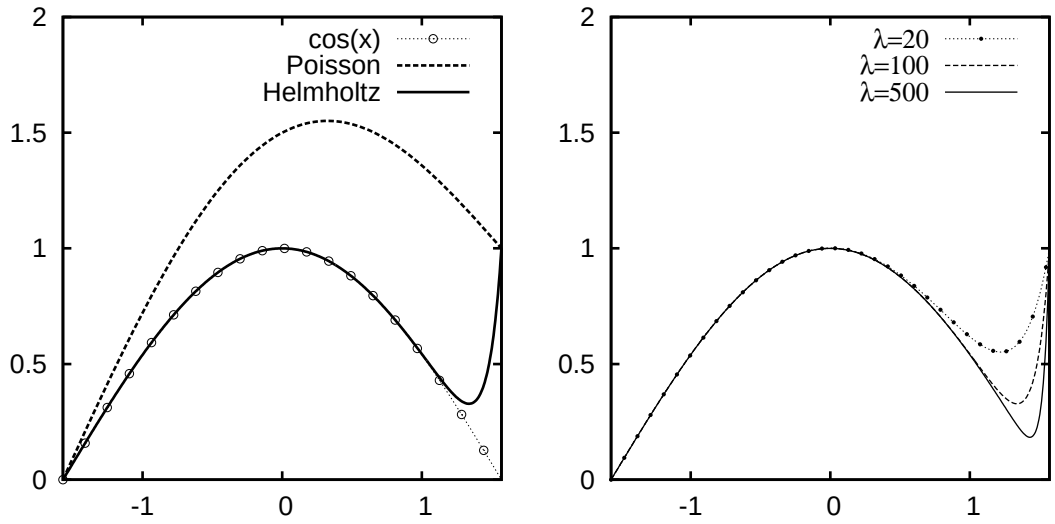


Figure 2.2: *left*: Comparison of solutions to the Poisson and Helmholtz ($\lambda = 100$) equations, when boundary condition is not consistent with the RHS. *right*: Dependence of the "boundary layer" on the value of *lambda* in the Helmholtz equation, when boundary condition inconsistent with the RHS is prescribed. Steep gradient occurs with increasing λ.

demonstrated, how a solution belonging to higher λ may become expensive for discrete approximation. Reminding the Helmholtz equation in the implicit step of the splitting algorithm, we find that

$$\lambda = \frac{1}{\bar{\nu}\,\Delta t}$$

in (2.35). If a fluid flow with higher Reynolds number is simulated, (e.g. $\mathrm{Re} = 10^4$), $\bar{\nu}$ becomes relatively small

$$\bar{\nu} \approx \frac{1}{\mathrm{Re}} = 10^{-4}\,.$$

Setting e.g. $\Delta t \approx 10^{-3}$ as a safe time step for stability of the scheme, we arrive to $\lambda \approx 10^{-7}$, what results in steep gradient ($\sim \exp(\sqrt{\lambda}x)$) in the solution towards the boundary if the boundary condition is not "RHS-consistent" (c.f. figure 2.19).

Now, we return to particular steps of the splitting algorithm. Those steps, where explicit time-stepping method is used, do not allow to prescribe implicit boundary conditions, since it is extrapolated together with the domain-interior values of the solution. Formally, the boundary condition is satisfied automatically in this case and strictly derived from the initial condition of the (sub-)problem. This is the case of (2.28), (2.41).

While setting the RHS-inconsistent boundary conditions do not result in lowering regularity of the solution in 1D problems, this is not the case in higher dimensions, where non-smooth geometry of domain Ω reflects in singularities as shown in more detail in sec. 2.2.5.

Neuman pressure BC

Since the pressure acts as a Lagrange multiplier in the INS system, its boundary condition may be calculated accordingly to the velocity field, if Dirichlet condition is not available.

Neumann pressure boundary condition is derived as the inner product of the momentum equation with the outward oriented normal vector $\mathbf{n}$ to the boundary $\partial\Omega$

$$\frac{\partial p_{n+1}}{\partial \mathbf{n}} = \mathbf{n} \cdot ([-\frac{\partial \mathbf{v}}{\partial t} - \mathbf{v} \cdot \nabla \mathbf{v} + \bar{\nu} \nabla \times \nabla \times \mathbf{v}]_{n+1}) \,,$$

where we used symbol $[\,]_n$ to represent evaluation at the n-th time step (isothermal case is presented for simplicity). At the pressure step, the velocity $\mathbf{v}_{n+1}$ is still unknown. Extrapolation is then used to evaluate the above formula, consistently with the extrapolation of the convective term in (2.28)

$$\frac{\partial p_{n+1}}{\partial \mathbf{n}} = \mathbf{n} \cdot \left\{ -\frac{\partial \mathbf{v}}{\partial t} + \sum_{q=0}^{J_e} \beta_q [\bar{\nu} \nabla \times \nabla \times \mathbf{v} - \mathbf{v} \cdot \nabla \mathbf{v}]_{n-q} \right\} . \tag{2.43}$$

The viscous term is in rotational form, what reduces the numerical boundary layer in velocity[5] (Orszag [28], Karniadakis [23]). The rotational form of the viscous term emerged directly in the computational algorithms presented in more recent works (Guermond [16], [17], Dong [6]). It has an impact not only to the behaviour near the boundary, but it has also relation to satisfaction of the inf-sup condition, what results in higher order estimates of the algorithm. This issue is discussed in Appendix C.

The same approach was used to derive boundary condition for pressure in case of the temperature dependent viscosity

$$\begin{aligned} \frac{\partial p_{n+1}}{\partial \mathbf{n}} &= \mathbf{n} \cdot \left\{ -\frac{\partial \mathbf{v}}{\partial t} + \right. \\ &\left. + \sum_{q=0}^{J_e} \beta_q [\nabla \nu_s \cdot (\nabla \mathbf{v} + (\nabla \mathbf{v})^{\mathrm{T}}) - \nu \nabla \times \nabla \times \mathbf{v} - \mathbf{v} \cdot \nabla \mathbf{v}]_{n-q} \right\} . \end{aligned} \tag{2.45}$$

Note, that viscosity multiplying the rotational term is "whole" viscosity ν, not only the time-constant part $\bar{\nu}$ as in previous cases.

This kind of pressure-Neumann boundary condition is evaluated using information only from the velocity field.

Outflow velocity BC

It is the third step (2.35), which allows (and requires) definition of the boundary conditions for velocity components $(v_i)_{n+1}^{(2,2)}$. Unfortunately, the boundary conditions prescribed a priory may be hardly consistent exactly with the RHS obtained

[5] Splitting of the operator results in the splitting error, which causes numerical boundary layer. The stiffly-stable method of third order (Karniadakis [23]) with the rotational form of the viscous term in pressure boundary condition produces numerical boundary layer, which vanishes as

$$\Delta t^3 \left[\frac{\mathrm{d}^3}{\mathrm{d}t^3} (\mathbf{v} \cdot \nabla) + 3\nu \frac{\mathrm{d}^3}{\mathrm{d}t^3} \nabla^2 \mathbf{v}) \right] \tag{2.44}$$

If the original form of the diffusive term is applied, constant numerical boundary layer is present (detail description is provided in (Karniadakis [25], pg. 424)).

employing extrapolation. The numerical boundary layer occurs, as demonstrated on the 1D situation on the beginning of this section. For more, the numerical boundary layer possibly lowers the solutions regularity, as a result of the non-smooth concatenating of domain boundaries (c.f. sec. 2.2.5).

While the values on inlet boundary are known and no-slip ($\mathbf{v} = \mathbf{0}$) condition may be prescribed on walls, values on the outflow boundary $\partial\Omega_O$ are still a subject of research. We briefly introduce some of the present results in this field.

One of the present works on this theme (Dong [6]) suggests its form

$$-p\mathbf{n} + \nu\mathbf{n}\cdot\nabla\mathbf{v} - \left[\frac{1}{2}|\mathbf{v}|^2 S_O(\mathbf{n}\cdot\mathbf{v})\right]\mathbf{n} = 0\,, \quad \text{on } \partial\Omega_O \tag{2.46}$$

where

$$S_O(\mathbf{n}\cdot\mathbf{v}) = \frac{1}{2}\left(1 - \tanh\frac{\mathbf{n}\cdot\mathbf{v}}{|\mathbf{v}_\infty|\,\delta}\right) \tag{2.47}$$

and $\mathbf{n}$ is the normal vector to the outflow part $\partial\Omega_O$ of the domain boundary. This kind of condition improves stability and may be modified by value of parameter δ, to find the best setting to the particular problem. The term 2.46 is an extension of other form, so called *directional do-nothing* condition (c.f. (1.47))

$$\mathbb{T}\mathbf{n} - \frac{1}{2}[\mathbf{v}\cdot\mathbf{n}]^- = 0\,. \tag{2.48}$$

This outflow condition was analysed presently in (Braack [2]), even so it was successfully used from its formulation in 1990's. In the above equations $[f(x)]^- = sup\{-f(x,0)\}$ is the term, which cuts off possible re-entrant flow and partly solves in this way the problems of the original *do-nothing* boundary condition

$$0 = \mathbb{T}\mathbf{n} = -p\mathbf{n} + \nu\mathbf{n}\cdot\nabla\mathbf{u} = 0\,. \tag{2.49}$$

In the case of pure outflow from the domain, both conditions 2.46 and 2.48 coincide with the classical do-nothing condition 2.49, which follows from the weak formulation of the (steady) Navier-Stokes system where

$$\begin{aligned} -(\nabla\cdot\mathbb{T},\phi) &= -\nu(\nabla^2\mathbf{v},\phi) + (\nabla p,\phi) \\ &= \nu(\nabla\mathbf{v},\nabla\phi) - (p,\nabla\cdot\phi) + \textstyle\int_{\partial\Omega}(-\nu\nabla\mathbf{v}+p)\mathbf{n}\,\phi\,\mathrm{d}s \\ &= \nu(\nabla\mathbf{v},\nabla\phi) - \textstyle\int_{\partial\Omega}\mathbb{T}\mathbf{n}\,\phi\,\mathrm{d}s \\ &\quad \forall\phi\in\{\mathbf{v}\in H^1(\Omega)^d : \nabla\cdot\mathbf{v}=0, \mathbf{v}|_{\partial\Omega_O} = 0\,a.e.\} \end{aligned} \tag{2.50}$$

2.1.3 General Linear Method

The general linear method (GLM) is a framework encapsulating schemes for time-stepping PDEs. It was originally proposed to unify stability analysis of Runge-Kutta (multi-stage) and multi-step methods (Butcher [5]). The common structure simplifies implementation of the time-stepping methods in computational codes. Recently, it emerged in implementations of software packages (e.g. Nektar++ [7], Hermes [19]).

We will focus only to the version implemented in Nektar++, which is enhanced to encompass the "IMEX" methods, which are employed in our computational algorithm.

Let the differential equation for $\mathbf{u} = [u_1, \ldots, u_D]^{\mathrm{T}}$ is in form

$$\begin{aligned} \frac{\mathrm{d}\mathbf{u}}{\mathrm{d}t} &= \mathbf{g}(\mathbf{u}) + \mathbf{f}(\mathbf{u}) \\ \mathbf{u}(t_0) &= \mathbf{u}_0 \,, \end{aligned} \tag{2.51}$$

where $\mathbf{g}(\cdot)$ and $\mathbf{f}(\cdot)$ are a differential operators in spatial coordinates. We employ the operator splitting technique (sec. 2.1.1) and apply implicit method to sub-problem with $\mathbf{g}$ and explicit method to the sub-problem with $\mathbf{f}$. Regarding algorithm for the incompressible Navier-Stokes equations (2.26-2.27), we can identify $\mathbf{g}(\mathbf{u}) = -\nabla p + \bar{\nu}\nabla^2\mathbf{u}$ and $\mathbf{f}(\mathbf{u}) = -\mathbf{u}\cdot\nabla\mathbf{u}$.

Method of our interest is single-stage, resp. strictly multi-step, but we will show its definition within the general GLM framework. A method using s-stages and r-steps is seen in form

$$\mathbf{Y}_i = \Delta t \sum_{j=1}^{s} a_{ij}^{IM}\mathbf{G}_j^{new} + \Delta t \sum_{j=1}^{s} a_{ij}^{EX}\mathbf{F}_j^{new} + \sum_{j=1}^{r} u_{ij}\mathbf{u}_j^{old},\ 1 \le i \le s, \tag{2.52a}$$

$$\mathbf{u}_i^{new} = \Delta t \sum_{j=1}^{s} b_{ij}^{IM}\mathbf{G}_j^{new} + \Delta t \sum_{j=1}^{s} b_{ij}^{EX}\mathbf{F}_j^{new} + \sum_{j=1}^{r} v_{ij}\mathbf{u}_j^{old},\ 1 \le i \le r\,, \tag{2.52b}$$

where $\mathbf{Y}_i$ are the *stage values* and $\mathbf{u}_i^{new}$ are elements of the output vector $\mathbf{u}^{new}$. $\mathbf{u}^{new}$ is an input ($\mathbf{u}^{old}$) for the next time step, carrying the actual solution together with all the other information needed for evaluation of the next time step as defined for a particular method. $\mathbf{G}_i = \mathbf{g}(\mathbf{Y}_i)$ denote implicitly evaluated *stage derivatives* and $\mathbf{F}_i = \mathbf{f}(\mathbf{Y}_i)$ denote explicitly evaluated stage derivatives. $\mathbf{Y}_i$, $\mathbf{G}_i$ and $\mathbf{F}_i$ ($i = 1, \ldots, s$) are computed within every time step Δt. These values are no more used in multi-stage methods, but some of them are stored in the output vector $\mathbf{u}^{new}$ if multi-step method is used. Runge-Kutta methods are recovered for $r = 1$, while linear multi-step methods are obtained with $s = 1$. The coefficients may be organized to matrices $\mathbb{B} = (b_{ij})$, $\mathbb{U} = (u_{ij})$, $\mathbb{V} = (v_{ij})$. Recognizing implicit $\mathbb{A}^{IM} = (a_{ij}^{IM})$ and explicit $\mathbb{A}^{EX} = (a_{ij}^{EX})$ evaluation of the stage derivatives allows the implementation of the IMEX schemes within the GLM framework.

An analogy to Butcher tableau (Butcher [5]) states

$$\left[\begin{array}{c} \mathbf{Y} \\ \hline \mathbf{u}^{new} \end{array}\right] = \left[\begin{array}{c|c||c} \mathbb{A}^{IM}\otimes\mathbb{I}_D & \mathbb{A}^{EX}\otimes\mathbb{I}_D & \mathbb{U}\otimes\mathbb{I}_D \\ \hline \mathbb{B}^{IM}\otimes\mathbb{I}_D & \mathbb{B}^{EX}\otimes\mathbb{I}_D & \mathbb{V}\otimes\mathbb{I}_D \end{array}\right] \left[\begin{array}{c} \Delta t\mathbf{G}^{new} \\ \Delta t\mathbf{F}^{new} \\ \hline \mathbf{u}^{old} \end{array}\right], \tag{2.53}$$

where the coefficient matrices are multiplied by $\mathbb{I}_D$, $D \times D$ identity matrix, since we work with D-dimensional problem (2.51).

$\mathbf{G}^{new}$ and $\mathbf{F}^{new}$ are evaluated within the new step of the algorithm. The input vector is known from previous time step(s)

$$\mathbf{u}^{old} = \left[\mathbf{u}_n, \ldots, \mathbf{u}_{n-(r-1)},\, \Delta t\mathbf{F}_n, \ldots, \mathbf{F}_{n-(r-1)}\right]^{\mathrm{T}},$$

where the subscript denote particular time step $u_n = u(t_n)$. The information is passed only through the output vector

$$\mathbf{u}^{new} = [\mathbf{u}_{n+1}, \ldots, \mathbf{u}_{n-r},\ \Delta t\mathbf{F}_{n+1}, \ldots, \mathbf{F}_{n-r}]^{\mathrm{T}} \ .$$

The computational algorithm is shown in Algorithm 1.

input: $\mathbf{u}^{old}$
output: $\mathbf{u}^{new}$

// Calculate stage values and stage derivatives
for $i = 1 \ldots s$ **do**
  //Calculate temporary variable $\mathbf{x}_i$
  $\mathbf{x}_i = \Delta t \sum_{j=1}^{i-1} a_{ij}^{IM} \mathbf{G}_j + \Delta t \sum_{j=1}^{i-1} a_{ij}^{EX} \mathbf{F}_j + \sum_{j=1}^{r} u_{ij} \mathbf{u}_j^{old}$

  //Calculate the stage value $\mathbf{Y}_i$ by solving
  $(\mathbf{Y}_i - a_{ii}^{IM} \Delta t \mathbf{g}(\mathbf{Y}_i)) = \mathbf{x}_i$

  //Calculate the explicit $\mathbf{F}_i$ and implicit $\mathbf{G}_i$ stage derivative
  $\mathbf{F}_i^{IM} = \mathbf{f}(\mathbf{Y}_i)$
  $\mathbf{G}_i^{IM} = \mathbf{g}(\mathbf{Y}_i) = \dfrac{1}{a_{ii}^{IM} \Delta t}(\mathbf{Y}_i - \mathbf{x}_i)$
end

// Calculate the output vector
for $i = 1 \ldots r$ **do**
  //Calculate $\mathbf{u}^{new}$
  $\mathbf{u}_i^{new} = \Delta t \sum_{j=1}^{s} b_{ij}^{IM} \mathbf{G}_j^{new} + \Delta t \sum_{j=1}^{s} b_{ij}^{EX} \mathbf{F}_j^{new} + \sum_{j=1}^{r} v_{ij} \mathbf{u}_j^{old}$
end

Algorithm 1: Algorithm of GLM (Vos [41]).

Implicit evaluation of the stage values $\mathbf{Y}_i$ and stage derivatives $\mathbf{F}_i$ must be provided in the code to complete the algorithm. Initialisation of the multi-step algorithms employs methods of lower order, so that the first step is evaluated by the one-step method.

As an example, the coefficients belonging to the second order "IMEX" scheme for 1D problem form following "Butcher table"

$$\left[\begin{array}{c|c||c} \mathbb{A}^{IM} & \mathbb{A}^{EX} & U \\ \hline \mathbb{B}^{IM} & \mathbb{B}^{EX} & V \end{array}\right] = \left[\begin{array}{cc||cccc} \frac{2}{3} & 0 & \frac{4}{3} & -\frac{1}{3} & \frac{4}{3} & -\frac{2}{3} \\ \hline \frac{2}{3} & 0 & \frac{4}{3} & -\frac{1}{3} & \frac{4}{3} & -\frac{2}{3} \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right] . \tag{2.54}$$

GLM is constructed from the similarities among particular methods and it is extensible for newly occurring schemes.

2.2 Spatial discretization

One of the aims in this work is use of a high order approximation to the elliptic PDEs, which constitute the single steps of the time-stepping algorithm. The high order approach is strongest in approximation of the smooth functions, what is a property expected in various physical processes and is not in contradiction to the observations of the fluid flow phenomena of our interest.

Mathematical analysis discovers various singularities emerging in solutions of the partial differential equations (especially in 2D and 3D). But only particular combinations of the boundary conditions, RHS and computational domain shape admit smooth solutions. Source of singularities is also in the algorithm discretising the solved system in time, a consequence of inaccuracy between the RHS function and boundary conditions. These singularities are not dumped in the high order methods. Therefore, unless the high order methods provide high accuracy and additional tools for analysis of the error, it gets less attention in applications than the lower order methods and up to the authors knowledge, there do not exist commercial software based on high order (spectral) approximations.

The high order approach follows rather the mathematical solution while showing the distance of the mathematical description from the physical reality, since the physical processes do not exhibit most of the mentioned singularities.

Another aspect of the high-order solutions is, that it abandons the strong dependence on design of the computational mesh, what is the case of low order methods. The high order solution then resembles more a function than a piecewise function or a set of function values and it carry also the information about the functions derivatives.

The high order methods also minimize the number of degrees of freedom in the resulting algebraic system.

We will start with formulation of a general framework, the *method of weighted residuals* what will help us to highlight differences among high order/spectral methods and the other commonly used. We will construct the numerical solution u_N and mention estimates controlling its distance from the exact solution in various norms $\|u - u_N\|$. More precisely, methods based on the abstract formulation of the differential equation will be employed to construct projection of the solution $P_N u$ to the space of the finite dimension (sec. 2.2.1). Truncating the space to the finite dimension produces the truncation (projection) error

$$\|u - P_N u\| \ . \tag{2.55}$$

However, the projection $P_N u$ itself can't be computed exactly and the discrete projection $I_N (\neq P_N)$ is performed instead. Therefore the approximation quality is estimated as

$$\|u - P_N u\| \leq \|u - I_N u\| + \|I_N u - P_N u\| \ , \tag{2.56}$$

where $\|u - I_N u\|$, the *discrete truncation error*, will be discussded in (sec. 2.2.4) together with the difference between the continuous and discrete projection, what is the *aliasing error*.

The global estimate

$$\|u - u_N\| \leq \|u - P_N u\| + \|u_N - P_N u\| \tag{2.57}$$

is then established on base of the convergence theory mentioned in Appendix A.

Attention will be paid to the standard elements, intrinsic to the weighted residual framework. These are especially the numerical integration, differentiation or approximation of the curved boundaries of the computational domain.

The section is concluded by various examples.

2.2.1 Method of weighted residuals

Computational methods concerning one common mathematical issue often share some similar features notwithstanding they are developed independently. Discovering the similarity motivates generalised formulation and mathematical structure, which simplifies comparison among particular methods and highlights their pros and cons. Theoretically, software packages based on the common structure would allow computations by multiple methods within one platform.

The *method of weighted residuals*[6] (MWR) unifies methods for approximative solution of PDEs[7]. Moreover it connects fields of mathematical analysis and computational methods, since MWR is finite-dimensional counterpart to the weak formulation.

Let $\mathcal{X}$ denote a functional space, whose definition reflects properties of $u(\vec{x})$, solution to a PDE

$$L\,u = f \text{ in } \Omega \tag{2.58}$$

(accompanied with boundary conditions). Differential operator $L : \mathcal{D}(L) \subset \mathcal{X} \to \mathcal{X}$ is defined on $\mathcal{D}(L)$, subset of functions $b(\vec{x}) \in \mathcal{X}$ satisfying (in some sense) the boundary conditions. In our discussions, L will take mostly forms of

$$\begin{aligned} &L = \nabla^2 && \text{Laplace} \\ &L = \nabla \cdot (\mu(\vec{x})\nabla) \\ &L = -(\nabla \cdot \mu(\vec{x})\nabla) + (\mathbf{v} \cdot \nabla) && \text{steady convection} - \text{diffusion}\,, \end{aligned} \tag{2.59}$$

since these occur as a separate steps in the time-marching algorithms of both the Navier-Stokes and heat equations (sec. 2.1). $f(\vec{x}) \in \mathcal{X}$ is a given function in (2.58).

Space $\mathcal{Y}$, which do not generally coincide with $\mathcal{X}$, is defined for the weak formulation. $\mathcal{Y}$ is used to specify weaker condition, under which the equation is satisfied, resulting in a set of equations

$$\int_\Omega (Lu)\, v \,\mathrm{d}\Omega = \int_\Omega f\, v \,\mathrm{d}\Omega \quad \forall v \in \mathcal{Y}\,. \tag{2.60}$$

This system of conditions tests, whether the weakened form of the PDE is satisfied and coincides with the weak form already presented for the system of Navier-Stokes and energy equations in sec. 1.2.1. Both $\mathcal{X}$ and $\mathcal{Y}$ are generally of the infinite dimension, so the exact solution $u(\vec{x})$ can be found in exact form in special

[6] Some authors use term *Mean weighted residuals* in the same sense (e.g. [1]).

[7] c.f. common concept for the time-marching schemes, the *General Linear Method* (2.1.3) and its implementation in Nektar++ library ([7])

cases only. MWR therefore restricts (2.60) to a finite dimension, introducing $\mathcal{X}_N = \mathrm{span}\{\Phi_p(\vec{x})\}_{p=1}^N \subset \mathcal{X}$ for representation of the approximate solution u_N and $\mathcal{Y}_N = \mathrm{span}\{\Psi_q(\vec{x})\}_{q=1}^N \subset \mathcal{Y}$ for testing. Definition of $\mathcal{X}_N$, $\mathcal{Y}_N$, resp. its basis sets $\{\Phi_p\}_{p=1}^N$, $\{\Psi_q\}_{q=1}^N$, is tied with particular computational method (c.f. overview below). The basic requirement to the basis sets is, that it should form a complete basis in particular space as $N \to \infty$, otherwise a good approximation to the solution is unachievable despite increasing of N.

Testing (2.58) over $\mathcal{Y}_N$ is, thanks to linearity of the space, equivalent to testing over the basis $\{\Psi_q\}_{q=1}^N$

$$\int_\Omega (Lu)\Psi_q \,\mathrm{d}\Omega = \int_\Omega f\,\Psi_q \,\mathrm{d}\Omega \quad \forall q = 1,\ldots,N\,. \tag{2.61}$$

Concurrently, (2.61) may be seen as a system balancing coefficients of the equation projected onto the space $\mathcal{Y}_N$. Denoting the projection as $P_{\mathcal{Y}}^N$ ($P_{\mathcal{Y}}^N : \mathcal{X} \to \mathcal{Y}_N$), we can write

$$P_{\mathcal{Y}}^N(Lu) = P_{\mathcal{Y}}^N f \in \mathcal{Y}_N\,. \tag{2.62}$$

Denoting by $\hat{u}_p^{\mathcal{X}}$ the coefficients of u in $\mathcal{X}_N$, we can substitute its projection ($P_{\mathcal{X}}^N : \mathcal{X} \to \mathcal{X}_N$)

$$P_{\mathcal{X}}^N u(\vec{x}) = \sum_{p=1}^N \hat{u}_p^{\mathcal{X}} \Phi_p(\vec{x})\,, \tag{2.63}$$

into (2.62) and obtain

$$P_{\mathcal{Y}}^N L P_{\mathcal{X}}^N u = P_{\mathcal{Y}}^N f\,. \tag{2.64}$$

Finally, (2.64) defines a matrix-vector system for unknown coefficients $\hat{u}_p^{\mathcal{X}}$

$$\mathbb{L}\hat{\mathbf{u}}_{\mathcal{X}} = \hat{\mathbf{f}}_{\mathcal{Y}}\,, \tag{2.65}$$

where

$$\mathbb{L}_{qp} = \int_\Omega \Psi_q\, L\Phi_p \,\mathrm{d}\Omega\,, \quad \hat{\mathbf{u}}_{\mathcal{X}} = \begin{bmatrix} \hat{u}_1^{\mathcal{X}} \\ \vdots \\ \hat{u}_N^{\mathcal{X}} \end{bmatrix}\,, \quad \hat{\mathbf{f}}_{\mathcal{Y}} = \begin{bmatrix} \hat{f}_1^{\mathcal{Y}} \\ \vdots \\ \hat{f}_N^{\mathcal{Y}} \end{bmatrix}\,. \tag{2.66}$$

Other interpretation of the same system follows, when firstly the solution is approximated by its projection $P_{\mathcal{X}}^N u$ in the original equation

$$L(P_{\mathcal{X}}^N u) \approx f\,. \tag{2.67}$$

This relation is no more an equation, since a non-zero residuum $r \in \mathcal{X}$ is produced

$$0 \neq r = LP_{\mathcal{X}}^N u - f\,. \tag{2.68}$$

$r \in \mathcal{X}$ and especially for nonlinear problems $r \notin \mathcal{X}_N$.
Now, the problem turns to determination of the N unknown coefficients $\hat{u}_p^{\mathcal{X}}$, so the testing in the weak form (2.60) may enforce only N conditions. These are specified through the choice of the N-dimensional space $\mathcal{Y}_N$. $\mathcal{Y}_N$ is defined by its basis $\{\Psi_q\}_{q=1}^N$, what results in the system of equations

$$\int_\Omega r\,\Psi_q \,\mathrm{d}\Omega = 0 \quad \forall q = 1,\ldots,N\,, \tag{2.69}$$

resp.

$$\int_\Omega (L\, P_{\mathcal{X}}^N u)\, \Psi_q \,\mathrm{d}\Omega = \int_\Omega f\, \Psi_q \,\mathrm{d}\Omega \quad \forall q = 1, \ldots, N\,. \tag{2.70}$$

The equation (2.70) may be seen as a set of conditions forcing r to vanish in $\mathcal{Y}_N$, what means

$$P_{\mathcal{Y}}^N r = \sum_{p=1}^{N} \hat{r}_p^{\mathcal{Y}} \Psi_q = 0\,, \tag{2.71}$$

being equivalent to set of conditions

$$\hat{r}_q^{\mathcal{Y}} = 0 \quad \forall q = 1, \ldots, N\,. \tag{2.72}$$

Substituting definition of r (2.68) into (2.72) results in the same system as (2.65).

The integral forms in (2.60) coincide with a "weighted" integral ($\int_\Omega a\, w \,\mathrm{d}\Omega$ is a weighted integral of a function a under the weight w), therefore $\{\Psi_q\}_{q=1}^N$ are denoted as *weight* functions in some literature and here also originates the name Method of *Weighted* Residuals).

If the space $\mathcal{X}$ is such, that the integral forms in (2.60) coincide with an inner product, the whole process gets a geometrical interpretation. The projections $P_N^{\mathcal{X}}$ and $P_{\mathcal{Y}}^N$ are then orthogonal projections and system (2.70) tests orthogonality of r to $\mathcal{Y}_N$.

Errors

Two kinds of error occur in just presented construction. The consequences of reduction the functional spaces to finite dimension and evaluation of the projections coefficients.

Insight to the errors, produced by the restriction to the finite dimension, is given if no approximation is performed, while elements of $\mathcal{X}$ are decomposed according to the projection operators, i.e.

$$u = P_{\mathcal{X}}^N u + (I - P_{\mathcal{X}}^N)u\,. \tag{2.73}$$

$(I - P_{\mathcal{X}}^N)u \in \mathcal{X} \setminus \mathcal{X}_N$ and similarly for $P_{\mathcal{Y}}^N$. Applying such decomposition to r using $P_{\mathcal{Y}}^N$

$$r = P_{\mathcal{Y}}^N r + (I - P_{\mathcal{Y}}^N)r\,,$$

resp. to the original equation (2.58), while substituting u by (2.73), we get

$$\begin{aligned} &P_{\mathcal{Y}}^N L[P_{\mathcal{X}}^N u + (I - P_{\mathcal{X}}^N)u] + (I - P_{\mathcal{Y}}^N)Lu = \\ &= P_{\mathcal{Y}}^N f + (I - P_{\mathcal{Y}}^N)f\,. \end{aligned} \tag{2.74}$$

Comparing elements of the same spaces we find that $(I - P_{\mathcal{Y}}^N)Lu \in \mathcal{Y} \setminus \mathcal{Y}_N$ represents that part of error caused by "insufficient testing" and equals to the

truncation error[8] of the data f in space $\mathcal{Y}_N$. Then, if the operator L is linear, we get (restricting to the elements of $\mathcal{Y}_N$)

$$P_{\mathcal{Y}}^N L P_{\mathcal{X}}^N u + P_{\mathcal{Y}}^N L(I - P_{\mathcal{X}}^N)u = P_{\mathcal{Y}}^N f\,, \tag{2.75}$$

where the second term on the left hand side defines the truncation error, whose bounds follow from particular definition of $\{\Phi_p\}_{p=1}^N$ (c.f. section 2.2.2). $L(I - P_{\mathcal{X}}^N)u$ is exactly the residuum (2.68) in this linear case.

The coefficients of projection (or *forward transform* from physical to functional space) are defined by the integral form

$$\hat{f}_p^{\mathcal{X}} = \frac{(f, \Phi_p)_{\mathcal{X}}}{(\Phi_p, \Phi_p)_{\mathcal{X}}} = \frac{1}{\|\Phi_p\|_{\mathcal{X}}^2} \int_\Omega f\, \Phi_p \,\mathrm{d}\vec{x}\,. \tag{2.76}$$

Integrals in (2.76), however, can't be computed exactly in general, so a quadrature rule must be employed (c.f. "aliasing error" in sec. 2.2.4) and the forward transform provides just an approximation $\tilde{f}_p^{\mathcal{X}}$ (symbol '~') of the coefficients $\hat{f}_p^{\mathcal{X}}$, where generally $\tilde{f}_p^{\mathcal{X}} \neq \hat{f}_p^{\mathcal{X}}$. Computed solution u_N includes both the approximation of the space and possible aliasing, therefore the computation do not provide an exact solution neither of the approximate system (2.70). $u_N \in \mathcal{X}$, but regardless the expectation of MWR $u_N \neq P_{\mathcal{X}}^N u$. $u_N = P_{\mathcal{X}}^N u$ only in the special cases, when the method provides an exact solution.

MWR concept do not impose mutual restriction between spaces $\mathcal{X}_N$ and $\mathcal{Y}_N$, so they can be defined independently.

Every method following under the concept of MWR defines the test functions $\{\Psi_q\}_{q=1}^N$ (basis of $\mathcal{Y}_N$) specifically and results in characteristic condition, which is satisfied by the solution (see table 2.4). Definition of the space spanned over the test (=weight) functions do not restrict the trial basis $\{\Phi_p(\vec{x})\}_{p=1}^N$, but inserts a rule measuring the quality of the approximation in the method-specific sense.

Particular methods in MWR

Particular methods are defined by the test and trial functions in MWR. Details about the trial functions are left to the separate section, so the overview of methods as seen in MWR will follow firstly the differences among the test spaces.

The *point collocation methods* force the residual to zero in a chosen set of (collocation) points $\{\vec{x}_q\}_{q=1}^N \subset \Omega$. In the MWR concept this means testing over a set of Dirac-delta functions shifted to the collocation points

$$\Psi_q(\vec{x}) = \delta(\vec{x} - \vec{x}_q)\,.$$

The set of "test" equations (2.69) then reads

$$(r, \Psi_q(\vec{x})) = \int_\Omega r\, \delta(\vec{x} - \vec{x}_q)\,\mathrm{d}\Omega = r(\vec{x}_q) = 0 \qquad q = 1, \ldots, N\,.$$

[8] The term "truncation error is used in spectral methods since the functions are represented by (infinite) series, which are *truncated* to finite expansion in $\mathcal{X}_N$ (or $\mathcal{Y}_N$)

Table 2.4: Weight/test functions $\Psi_q(\vec{x})$ for various methods in concept of MWR (compare [25], pg. 19 and [42]). $\vec{x}_q \in \Omega$ are chosen points, where the collocation method forces $R(\vec{x}_q) = 0$ and Ω_q are elements of a finite volume mesh $\Omega \approx \cup_q \Omega_q$ approximating Ω as a domain with piecewise linear boundary.

$\Psi_q(\vec{x})$	Method
$\delta(\vec{x} - \vec{x}_q)$	Point collocation
$\dfrac{\partial R}{\partial \hat{u}_q}$	Least-squares
$\chi_q(\vec{x}) \begin{cases} 1, & \vec{x} \in \Omega_q \\ 0, & \vec{x} \notin \Omega_q \end{cases}$	Domain collocation
Φ_q	Galerkin
$\Psi_q (\neq \Phi_q)$	Petrov-Galerkin

The solution is represented in the transform space, but the equation stays in the physical space. This fact substantially simplifies treatment of the non-linear equations, which produce difficulties, when transformed to a linear space.

Taking as the expansion basis a low order interpolants (Lagrange or trigonometric polynomial) through the collocation points, while performing exact integration in the testing phase, the *finite-difference method* or Fourier finite difference method is recovered

$$(L\, u_N)|_{\vec{x}=\vec{x}_q} = f(\vec{x}_q) \qquad q = 1, \ldots, N \ .$$

The process of obtaining the "infinite" order method as a limit of increasing order to finite differences is described in (Hesthaven [18]) and forms a basic motivation for the spectral methods.

Under the term *domain collocation* follow methods, whose test functions are

$$\Psi_q(\vec{x}) = \begin{cases} 1, & \vec{x} \in \Omega_q \\ 0, & \vec{x} \notin \Omega_q \end{cases} \qquad q = 1, \ldots, N$$

and the system of equations is

$$\int_{\Omega_q} (L_\delta\, u_N - f)\, \mathrm{d}\Omega = 0 \qquad q = 1, \ldots, N\,.$$

Generally, depending on the choice of the space $\mathcal{X}_N$ introduction of a discrete operator L_δ may be necessary. This is the case of the finite volume method (FV), which define $\mathcal{X}_N$ as piecewise-constant functions, what results in reformulation of the differential operator, which contains the numerical fluxes.

Other examples following under the MWR concept are the Method of least squares or Method of moments.

The Galerkin formulation unifies the spaces of trial and test functions, $\Phi_p(x) = \Psi_q(x)$ $p, q = 1, \ldots, N$. Concerning the boundary conditions, some authors distinguish *Galerkin* and *Tau* method ([8], [18]). The boundary conditions in the Galerkin approach are satisfied individually by all the trial and test functions. This approach is reasonable only in case of special boundary conditions (e.g. homogeneous), while for more general and time dependent case it becomes untenable. The Tau method is a modification of the Galerkin approach, when the trial functions do not satisfy the boundary conditions. A boundary condition in the Tau method is enforced as collocation at the boundary points. Coincidence of trial and test functions is then limited to $\Phi_p = \Psi_q \quad p = q = 1, \ldots, N-k$, where k is number of the boundary conditions. The problem of construction appropriate functional space is moved to the structure of matrix in the resulting algebraic system, since the method introduce a set of equations to satisfy the conditions. The Tau method can be seen as a special case of the Petrov-Galerkin method, where trial and test functions do not coincide.

Concept of MWR includes methods, which combine approaches of just described main types. This is the case of *spectral element method* (SEM), which takes advantages of both FEM and SM or discontinuous Galerkin method using formulation of fluxes known from FV within the higher than constant order of the trial space.

2.2.2 Trial functions

The set of trial functions $\{\Phi_n\}_{n=1}^N$ do not generally coincide with the test functions $\{\Psi_n\}_{n=1}^N$. However, we will apply method where $\{\Phi_n\}_{n=1}^N = \{\Psi_n\}_{n=1}^N$ in later applications. Therefore we do not distinguish the projection operators to trial and test spaces henceforward and introduce a common symbol $P_N = P_{\mathcal{X}}^N = P_{\mathcal{Y}}^N$ to simplify the notation. Similarly, the coefficients, representing functions in discrete spaces will be simplified to $\hat{f}_n = \hat{f}_p^{\mathcal{X}} \equiv \hat{f}_q^{\mathcal{Y}}$ and $\tilde{f}_n = \tilde{f}_p^{\mathcal{X}} \equiv \tilde{f}_q^{\mathcal{Y}}$. Keeping an exact form of the trial functions unspecified, we distinguish them into two classes

- global over Ω

$$\overline{\mathrm{supp}(\Phi_n)} \approx \Omega \quad \forall n = 1, \ldots, N$$

- localised in Ω

$$\forall n = 1, \ldots, N \; \exists e : \overline{\mathrm{supp}(\Phi_n)} = \Omega_e \subsetneq \Omega \quad \text{where} \quad \Phi_n(x) = 0 \quad \forall x \in \Omega \setminus \Omega_e$$

and

$$\Omega \approx \bigcup_{e=1}^{E} \Omega_e \,.$$

Some methods (spectral/hp FEM)[9] combine both approaches, so that a multiple localised trial functions share a common support, therefore $N \neq E$ generally.

[9] We present the term "spectral/hp FEM" which is used in title of (Karniadakis [**?**]) to show, that wide terminology denotes similar or same methods. The finite element method (FEM) employing high order approximation on every element coincides with the spectral element method and both these methods may exhibit the h-convergence (mesh-refinement) or the p-convergence (increase in order of the approximation), what some authors denote as hp-FEM (e.g. Šolín [38]).

A function f as projected to $\mathcal{X}_N$ is then represented by a finite sum

$$P_N f(x) = \sum_{n=1}^{N} \hat{f}_n \, \Phi_n(x) = \sum_{e=1}^{E} \sum_{m=0}^{M_e} \hat{f}_m^e \, \Phi_m^e(x) \quad x \in \Omega \,, \tag{2.77}$$

where we introduce $\Phi_m^e(x) = \Phi_n(x) \quad \forall m, n : \mathrm{supp}(\Phi_n) = \Omega_e$.

Some methods are formulated such, that it allows M_e to be unique for every sub-domain Ω_e, but we restrict to $M_e = M \; \forall e$ in this work.

Methods using the decomposition of Ω to "elements" Ω_e effectively work on general geometries of the computational domain in higher dimensions, where the global methods ($E = 1$ and $N = M + 1$) are not applicable[10].

We will restrict now to 1D, $\vec{x} \to x$, for sake of simplicity. Since the basis functions in the global method share one common support, $P_N f$ may be seen as an N-element truncation of an expansion series $\sum_{n=1}^{\infty} \hat{f}_n \, \phi_n$

$$P_N f = \sum_{n=1}^{N} \hat{f}_n \, \Phi_n = \sum_{n=1}^{N} \hat{f}_n \, \phi_n \,. \tag{2.78}$$

It is reasonable to use an expansion series on its domain of convergence Ω_{std}. Because the computational domain Ω can not be restricted to coincide with Ω_{std}, we introduce local coordinate $\xi \in \Omega_{std}$ and a mapping $x_\Omega(\xi) : \Omega_{std} \to \Omega$, which allows to expand the function at arbitrary Ω, since $x_\Omega^{-1}(x) \in \Omega_{std}$ and finally

$$P_N f(x) = \sum_{n=1}^{N} \hat{f}_n \phi_n(x_\Omega^{-1}(x)) \,.$$

The mapping may be defined similarly for every Ω_e in case of localised trial functions, $x_e(\xi) : \Omega_{std} \to \Omega_e$, what allows to define an expansion series on every sub-domain Ω_e, while $\Phi_n^e(x) = \phi_n(x_e^{-1}(x))$.

Especially for the global methods, the terms *expansion* and *trial* functions coincide, both denoting the basis $\{\phi_n(x)\}_{n=1}^{N}$ in $\mathcal{X}_N$. The expansion coefficients $\hat{f}_n$ form for increasing n a "spectrum" over the trial basis and the global methods are therefore denoted as *spectral methods* (SM). If a (truncated) expansion is constructed on every Ω_e, the method is known as the *spectral element method* (SEM) and coincides with the finite element (FEM) formulation.

The expansion functions $\{\phi_n\}_{n=0}^{\infty}$ must be chosen such, that it form a complete system in $\mathcal{X}$. Otherwise $\mathcal{X}_N \nrightarrow \mathcal{X}$ as $N \to \infty$ and the existence of particular function wouldn't be ensured in the trial space.

Generally we can think about expansions to the Taylor, Fourier, Chebyshev, Hermite, etc. series, but only some of them provide reasonable convergence properties.

Values of $\hat{f}_n$ are unique for particular expansion. The truncated part of the expansion $\sum_{n=N+1}^{\infty} \hat{f}_n$ define the truncation error. If the coefficients $\hat{f}_n$ decay fast enough as $N \to \infty$, the truncation error becomes finite and the faster the decay is, the smaller truncation error occurs[11].

[10] Applicability of the global methods in 2D and 3D suffer from the absence of expansion basis defined on a standard domain with a hole inside

[11] The reasons, why and how the expansion coefficients decay for particular function expansion is hidden in the complex analysis. For example, the Taylor series has a circular domain of

Sturm-Liouville problem

The Sturm-Liouville problem is defined by

$$L\phi(x) = -\frac{\mathrm{d}}{\mathrm{d}x}\left(p(x)\frac{\mathrm{d}\phi(x)}{\mathrm{d}x}\right) + q(x)\phi(x) = \lambda w(x)\phi(x)\,, \quad x \in [-1,1] \tag{2.79}$$

as a subject to boundary conditions ($a_\pm$, $b_\pm$ are constants)

$$\begin{aligned} a_-\phi(-1) + \beta_-\phi'(-1) = 0, \qquad & a_-^2 + \beta_-^2 \neq 0, \\ a_+\phi(1) + \beta_+\phi'(1) = 0, \qquad & a_+^2 + \beta_+^2 \neq 0\,, \end{aligned} \tag{2.80}$$

where

$$\begin{gathered} p(x) \in \mathcal{C}^1[-1,1];\, p(x) > 0 \quad \forall x \in (-1,1)\,, \\ q(x) \in \mathcal{C}^0[-1,1];\, 0 \leq q < \infty\,, \\ w(x) \in \mathcal{C}^0[-1,1];\, 0 \leq w\,. \end{gathered}$$

Assuming $a_-b_- \leq 0, \quad a_+b_+ \geq 0$ it can be shown, that solutions to the eigenvalue problem form a complete basis in $\mathcal{L}^2[-1,1]$. Since L is self-adjoint, this basis is orthogonal in the weighted $\mathcal{L}^2$ norm $\|\cdot\|_{\mathcal{L}^2_w[-1,1]} = \int_{-1}^{1} |\cdot|^2\, w(x)\mathrm{d}x$, while the eigenvalues λ_n are non-negative and form unbounded sequence with quadratic growth

$$0 \leq \lambda_n \propto n^2 \qquad \text{as} \qquad n \to \infty\,.$$

The expansion coefficients in a series based on a system of eigensolutions ϕ_n to (2.79) satisfy

$$\begin{aligned} &\hat{f}_n = \frac{1}{\gamma_n}\int_{-1}^{1} f(x)\phi_n(x)w(x)\,\mathrm{d}x = \\ &\ldots = \frac{1}{\gamma_n\,\lambda_n}[p(f'\phi_n - f\phi_n')]_{-1}^{1} + \frac{1}{\gamma_n\,\lambda_n}\int_{-1}^{1}(Lf(x))\,\phi_n(x)\,\mathrm{d}x\,, \end{aligned} \tag{2.81}$$

where $\gamma_n = (\phi_n, \phi_n)_{L_w[-1,1]}$. Procedure in (2.81) is valid providing that $Lf, L^2f \in \mathcal{L}^2[-1,1]$. The boundary term in (2.81) vanishes if

$$p(\pm 1) = 0\,, \tag{2.82}$$

or

$$u(1) = u(-1)\,, \quad u'(1) = u'(-1)\,. \tag{2.83}$$

convergence in the complex plane, Fourier series has an infinite strip along the real axis, while the Jacobi polynomials have an ellipse with foci in -1 and 1. The convergence rate of the series is limited by the regularity of approximated function in the domain of convergence (c.f. table 2.5). Convergence of the Taylor series is limited by the distance of the first singularity in the complex plane. On the other hand, the region of convergence may often obviate the singularities relatively near to the real axis and limits of $[-1,1]$ in case of the Jacobi plynomials (Chebyshev, Legendre, ...). The Fourier expansion is dependent solely on the conditions of periodicity. These issues are presented in more detail in (Boyd [1]). This is an illustrative explanation, why the Taylor series is not used in spectral methods. On the other side, both the Fourier and series based on Jacobi polynomials are common in spectral methods, both being eigensolutions to the Sturm-Liouville problem.

Reminding definition (2.79), the condition (2.82) defines the *singular* Sturm-Liouville problem, while $p(\pm 1) \neq 0$ belongs to the *regular* Sturm-Liouville problem. Zeroing the boundary term, we can repeat the whole process (2.81) and conclude that

$$|\hat{u}_n| \simeq C \frac{1}{(\lambda_n)^m} \left\| \left(\frac{L}{w(x)} \right)^m u(x) \right\|_{\mathcal{L}^2_w[-1,1]} . \tag{2.84}$$

This estimate applies asymptotically as $n \to \infty$ and establishes estimete for the coefficients decay for the spectral methods in dependence on the regularity of the expanded function. Exponential (infinite order) decay is recovered for the coefficients of $\mathcal{C}^\infty$ functions. The fast decay of expansion coefficients is the reason, why eigensolutions ϕ_n to the Sturm-Liouville problem are used in construction of the basis in the spectral type methods.

The estimate (2.84) is valid without further limitations to $w(x)$ or $q(x)$, therefore a wide class of orthogonal functions exhibit this property. The Fourier basis is an example of such a set in case of the regular Sturm-Liouville problem, while the Jacobi polynomials belong to the singular version (2.82).

Fourier spectral methods

gained attention after appearance of the fast Fourier transform in 1965. The efficient projection to the transform space extended field of solvable differential equations. Fourier methods, however, are limited to the periodic problems, since this feature is intrinsic to the trigonometric functions, which form the basis in the transform space $\mathcal{X}_N = \mathrm{span}\{e^{inx} | \, |n| \leq (N-1)/2\}$ in this case (i denotes the imaginary unit). The order of convergence for the expansion coefficients $\hat{u}_n$ (as $n \to \infty$) is given by regularity and periodicity of the function and its derivatives.

Let a periodic function $u(x)$ has derivative $u'(x) \in \mathcal{L}^2[0, 2\pi]$. The expansion coefficients are given repeating the procedure (2.81) with $\phi_n = e^{inx}$

$$2\pi \hat{u}_n = \int_0^{2\pi} u(x)\, e^{-inx} \mathrm{d}x = \frac{-1}{i\, n} (u(2\pi) - u(0)) + \frac{1}{i\, n} \int_0^{2\pi} u'(x) e^{-inx} \mathrm{d}x \,, \tag{2.85}$$

what shows that

$$|\hat{u}_n| \propto \frac{1}{n} .$$

The first term on the right hand side of (2.85) vanishes from periodicity. If the function has higher regularity $u^{(m)} \in L^2[0, 2\pi]$ and $u^{(m-1)}$ is periodic, then integration per-parts, applied repeatedly to the last term, results in

$$|\hat{u}_n| \propto \left(\frac{1}{n} \right)^m .$$

For periodic function $u(x) \in \mathcal{C}^\infty[0, 2\pi]$ with periodic derivatives this asymptotic behaviour (as $n \to \infty$) reaches exponential decay.

The periodicity is crucial for the high order decay in the Fourier methods. For illustration, lets take a periodic function with non-periodic derivative

$$u(x) = \sin\left(\frac{x}{2} \right) \quad x \in [0, 2\pi] .$$

Its expansion coefficients on the Fourier basis are

$$\hat{u}_n = \frac{2}{\pi} \frac{1}{(1-4n^2)}\,,$$

exhibiting only quadratic decay in n as a consequence of missing periodicity of the derivative. The truncation error do not decay fast enough in this case and makes the Fourier spectral method expensive and inefficient.

If the Fourier spectral method is applied to approximate solution to a differential equation, the requirement of periodicity reflects in demand of periodic boundary conditions, therefore it is not used in the local type methods, since the solution on every neighbouring domain would be just the periodic extension. However, the Fourier expansion is successfully applied in solutions of differential problems with periodic boundary conditions (e.g. [24]).

Expansions based on Jacobi polynomials

Concerning the singular Sturm-Liouville problem (SSL), a high order coefficient decay (2.84) is recovered without any requirement of periodicity to u. A relevant class of eigenfunctions is obtained setting

$$w_{(\alpha,\beta)}(x) = (1-x)^\alpha (1+x)^\beta \tag{2.86}$$

$$q(x) \equiv 0 \tag{2.87}$$

$$p(x) = (1-x)^{\alpha+1}(1+x)^{\beta+1} \text{ (i.e. } p(\pm 1) = 0). \tag{2.88}$$

Then, the SSL takes form

$$\frac{\mathrm{d}}{\mathrm{d}x}\left[(1-x)^{(1+\alpha)}(1+x)^{(1+\beta)}\frac{\mathrm{d}}{\mathrm{d}x}J_n^{(\alpha,\beta)}(x)\right] = \lambda_n (1-x)^\alpha (1+x)^{(\beta)} J_n^{\alpha,\beta}(x)\,,$$

where $J_n^{\alpha,\beta}(x)$, the eigenfunctions, are the Jacobi polynomials. These are orthogonal under the weighted inner product $(\cdot,\cdot)_w$ (the weight being that defined in (2.86))

$$(J_m^{(\alpha,\beta)}, J_n^{\alpha,\beta})_w = \int_{-1}^{1} J_m^{(\alpha,\beta)}(x)\, J_n^{(\alpha,\beta)}(x)\, w_{(\alpha,\beta)}(x)\,\mathrm{d}x = C_{mn}\,\delta_{mn}\,. \tag{2.89}$$

Taking $\alpha = \beta$ together with normalization defines the *ultraspherical* polynomials $U_n^{(\alpha)}(x)$, whose representatives are the *Chebyshev* ($\alpha = 1/2$) and *Legendre*($\alpha = 0$) polynomials.

Jacobi polynomials play the key role in construction of the spectral methods, since they are used in constructions both of the expansion basis, but also the quadrature formulas with maximal degree of exactness (c.f. Gauss quadrature formulas in sec. 2.2.4). Some of the basic formulas for evaluation of these functions are summarized in Appendix D.

Following theorems provide estimates of the truncation error for ultraspherical polynomials, as formulated for $\Omega_{std} = [-1:1]$ (proofs, e.g. Hesthaven [18]).

Theorem 4. *For any $u(x) \in \mathcal{H}_w^p[-1,1]$, $p \geq 0$, there exists a constant C, independent of N, such that*

$$\|u - P_N u\|_{\mathcal{L}_w^2[-1,1]} \leq C\, N^{-p}\, \|u\|_{\mathcal{H}_w^p[-1,1]}\,. \tag{2.90}$$

Projection operator do not commute with the operator of derivative, however the error exhibits high order decay if the function is regular enough.

Theorem 5. *Let $u(x) \in \mathcal{H}^p_w[-1,1]$, there exists a constant C, independent of N, such that*

$$\left\| P_N \frac{\mathrm{d}u}{\mathrm{d}x} - \frac{\mathrm{d}}{\mathrm{d}x} P_N u \right\|_{\mathcal{H}^q_w[-1,1]} \leq C\, N^{2q-p+3/2}\, \|u\|_{\mathcal{H}^p_w[-1,1]}\ , \tag{2.91}$$

where $1 \leq q \leq p$.

Estimate in Sobolev norm follows consequently

Theorem 6. *For any $u(x) \in \mathcal{H}^p_w[-1,1]$ there exists a constant C, independent of n, such that for $0 \leq q \leq p$*

$$\begin{aligned} &\|u - P_N u\|_{\mathcal{H}^q_w[-1,1]} \leq C\, N^{3q/2-p}\, \|u\|_{\mathcal{H}^p_w[-1,1]} \qquad 0 \leq q \leq 1, \\ &\|u - P_N u\|_{\mathcal{H}^q_w[-1,1]} \leq C\, N^{2q-p-1/2}\, \|u\|_{\mathcal{H}^p_w[-1,1]} \quad q \geq 1\,. \end{aligned} \tag{2.92}$$

These theorems are valid in case of exact evaluation of the expansion coefficients. Its discrete counterpart must consider the aliasing error, what is an issue in sec. 2.2.4, where the discrete version of above theorems are presented.

Smoothness of the expanded function plays a key role in the convergence properties. Functions, which contain some kind of singularity in the convergence domain of the expansion do not achieve exponential decay as shown in table 2.5.

An example of a modal 1D basis on a standard interval $\xi \in [-1:1]$ is

$$\phi_m(\xi) = \begin{cases} \dfrac{1-\xi}{2}, & m = 0 \\ \left(\dfrac{1-\xi}{2}\right)\left(\dfrac{1+\xi}{2}\right) U^{(1)}_{m-1}(\xi), & 0 < m < M \\ \dfrac{1+\xi}{2}, & m = M \end{cases} \tag{2.93}$$

Only the boundary modes ($m = 0, M$) are non-zero at the border points and these provide also the $\mathcal{C}^0$ continuity among neighbouring subdomains Ω_e, needed in the weak formulation of the second order problems. Basis defined in (2.93) is used in later computations and construction of a multi-dimensional basis. Several first modes of this basis are plotted in figure 2.3.

The finite dimensional spaces $\mathcal{X}_N = \mathrm{span}\{\phi_{n-1}\}_{n=1}^N$, where ϕ are defined in (2.93), form a hierarchical structure, since $\mathcal{X}_N \subset \mathcal{X}_{N+1} \quad \forall N = 1, 2, \ldots$. This kind of basis is sometimes noted as *modal*, to distinguish it from the basis described in the next section.

Nodal basis

Another kind of trial basis arises from a set of $(N+1)$ Lagrange characteristic polynomials

$$l_n(x) = \prod_{\substack{m=0 \\ m \neq n}}^{N} \frac{x - \xi_m}{\xi_n - \xi_m} \quad n = 0, \ldots, N \tag{2.94}$$

Type of singularity	Form of singularity	Asymptotic behaviour of expansion coeffs
Simple pole	$1/(x-a)$	$[]p^n$
Double pole	$1/(x-a)^2$	$[]n\,p^n$
Logarithm	$log(x-a)$	$[]n^{-1}\,p^n$
Reciprocal of square root	$1/\sqrt{x-a}$	$[]n^{-1/2}\,p^n$
Cube root	$(x-a)^{1/3}$	$[]p^n/n^{4/3}$
Infinitely differentiable singular at $x=0$	$\exp(-q/\lvert x\rvert)$	$[]\exp(-p\,n^{1/2})$
Branch point within $[-1,1]$	x^y	$[]1/n^{y+1}$
Jump discontinuity	$\mathrm{sign}(x)$	$[]/n$
Continuous function with discontinuous first derivative	$\frac{\mathrm{d}f}{\mathrm{d}x}=\mathrm{sign}(x)$	$[]/n^2$
Infinitely differentiable but singular at endpoint	$\exp(-q/\lvert x+1\rvert)$	$[]\exp(-pn^{2/3})$

Table 2.5: Asymptotic behaviour of the spectral coefficients in approximation of functions with singularities ([1], pg. 60). Empty brackets [] denote slowly varying algebraic factor of n.

associated with quadrature points $\{\xi_n\}_{n=0}^N$. The truncated series in this basis is then the Lagrange interpolation polynomial

$$P_N^{\mathcal{X}} f(x) = \Pi_f^N(x) = \sum_{n=0}^{N} f(\xi_n) l_n(x)\,.$$

Here often arises a misunderstanding of the term "spectral method". Because $\hat{f}_n = f(\xi_n)$, it is misleading to use the term spectrum or expansion coefficients. However, for a smooth function f, we recover the exponential decay as $N\to\infty$ for $\|f - P_N f\|$, similarly to the Fourier or Jacobi-type expansions. This property is sometimes noted as *spectral accuracy*. Equivalence of nodal and Jacobi-type expansions follows from numerical evaluation of the coefficients of the latter one:

Taking points $\{\xi_n\}_{n=0}^N$ and weights $\{w_n\}_{n=0}^N$ from the Gauss-type quadrature generated by $U_N^{(\alpha)}$ (c.f. 2.2.4), we obtain

$$I_N f(x) = \sum_{n=0}^{N} \tilde{f}_n\, U_n^{(\alpha)}(x) = \sum_{n=0}^{N} f(\xi_n) l_n(x) = \Pi_f^N(x) \tag{2.95}$$

since

$$l_n(x) = w_n \sum_{m=0}^{N} \frac{1}{\gamma_m} U_m^{(\alpha)}(x) U_m^{(\alpha)}(\xi_m) \quad n = 0,\ldots,N\,,$$

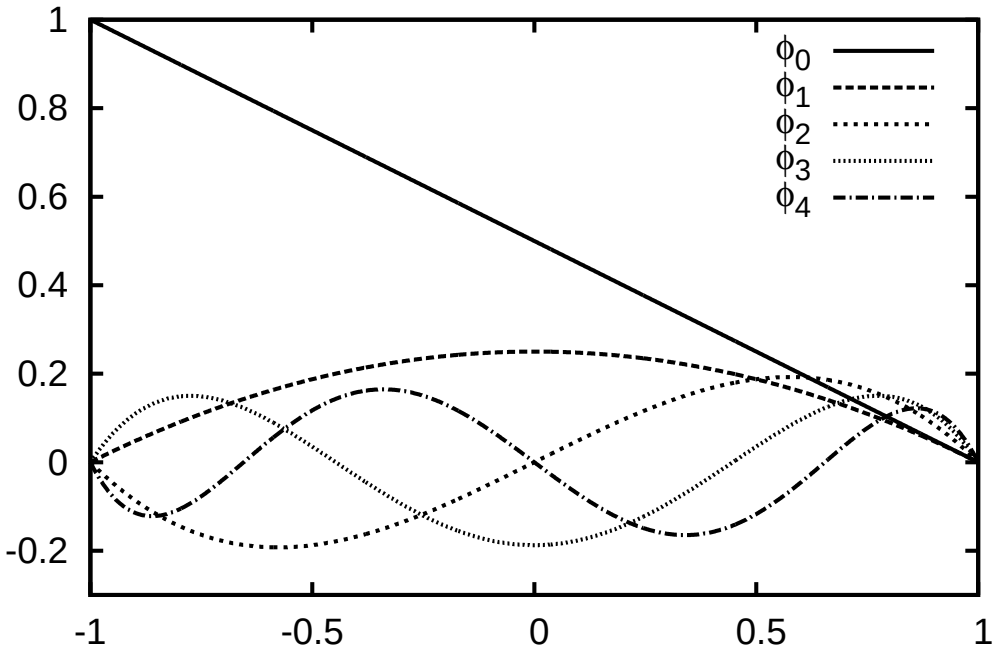


Figure 2.3: Plot of the first 5 functions from 1D basis (2.93) over the standard interval $[-1:1]$. The functions oscillate with increasing order, since all their zeros are in the standard interval, but all of them are bounded by the extremal values of the lowest modes.

where γ_m is discretely, but accurately evaluated inner product $\gamma_m = (U_m^{(\alpha)}, U_m^{(\alpha)})$. (2.95) implies the decay of error similar to discretely evaluated Jacobi expansion (theorem 7).

The nodal basis exhibits so called *δ-property*

$$\phi_m(\xi_n) = \delta_{mn} \quad m, n = 0, \dots N\,. \tag{2.96}$$

The spaces spanned above the nodal basis do not exhibit the hierarchical structure as in case of the modal basis, since all the trial functions are polynomials of order N. Rather it is referred as *nodal* or *collocation*, because the truncated series collocates the function in the nodes $\{\xi_n\}_{n=0}^N$. The nodal basis functions are also referred as discrete δ-functions, because on a discrete set of points it is indistinguishable from the Dirac-δ function (c.f. point collocation methods in sec. 2.2.1).

Definition of this kind of basis is dependent on the associated points. The formulation concerning the boundary condition is simplified substantially, if the points are of the Gauss-Lobatto type, which include the endpoints of the interval. Taking $N+1$ Gauss-Legendre-Lobatto points, nodal functions on the standard interval $\xi \in [-1:1]$ are

$$\phi_n(\xi) = \frac{1-\xi^2}{N(N+1)L_N(\xi_n)(\xi_n - \xi)} \frac{\partial L_N}{\partial \xi} \quad n = 0, \dots, N \tag{2.97}$$

with ξ_i being now the Gauss-Legendre-Lobatto nodes (2.110) and $L_N(\xi) = U_N^{(0)}(\xi)$ the Legendre polynomial of degree N.

The Galerkin approach applied to the nodal basis coincide with the point-collocation, when the numerical integration is performed in the testing phase.

Multidimensional basis

Basis for the quadrilateral standard domain $\Omega_{std} = [-1,1] \times [-1,1]$ is the tensor product of the 1D expansions (2.93) in both coordinates

$$\phi_{mn}(\xi_1, \xi_2) = \phi_m(\xi_1)\phi_n(\xi_2) \quad m = 1, \ldots, M_1\,, \quad n = 1, \ldots, M_2\,. \tag{2.98}$$

The orders of the expansion, M_1, M_2, are not necessarily same for both directions. The trial functions on the standard triangle are derived also from the tensor product of 1D expansions. The formulation taken from (Karniadakis [25]) uses collapsed coordinates (η_1, η_2), which define the transform between the standard quadrilateral (ξ_1, ξ_2) and standard triangle (η_1, η_2)

$$\begin{aligned} \xi_1 &= \frac{(1+\eta_1)(1-\eta_2)}{2} - 1\,, \\ \xi_2 &= \eta_2 \end{aligned} \tag{2.99}$$

resp.

$$\begin{aligned} \eta_1 &= \frac{2(1+\xi_1)}{1-\xi_2} - 1\,, \\ \eta_2 &= \xi_2\,. \end{aligned} \tag{2.100}$$

The standard triangle is then given as $\mathcal{T} = \{(\eta_1, \eta_2) | -1 \leq \eta_1, \eta_2 \leq 1\}$ and associated orthogonal basis

$$\phi_{mn}(\xi_1, \xi_2) = \phi_m(\eta_1)\psi_{mn}(\eta_2)\,, \tag{2.101}$$

where

$$\psi_{mn}(\xi) = \begin{cases} \phi_m(\xi)\,, & m = 0, 0 \leq n \leq M_2\,, \\ \left(\dfrac{1-\xi}{2}\right)^{m+1}\,, & 1 \leq m < M_1,\, n = 0\,, \\ \left(\dfrac{1-\xi}{2}\right)^{m+1} \dfrac{1+\xi}{2} J_{n-1}^{2m+1,1}(\xi), & 1 \leq m < M_1,\, 1 \leq n < M_2\,, \\ \phi_n(\xi), & m = M_1,\, 0 \leq n \leq M_2\,. \end{cases} \tag{2.102}$$

2.2.3 Differentiation

The approximation by high order expansion includes an approximation also to the function derivatives, which may be directly evaluated in any physical point in Ω. For high order approximation, it allows the second or higher derivatives in the computational schemes.

Differentiation is applied in evaluation of elements of the Jacobi matrix (2.124) from the weighted residual formulation on the deformed elements, but also in the explicit steps of our scheme (2.28), (2.41) and (2.45), where the second order derivatives are evaluated. However, the accuracy of approximation decays with order of the derivative due to the finite order of the basis and the finite computer precision. The numerically evaluated differentiation is then one of the most dangerous operations, since the round-off error, caused by representation of the expansion coefficients in finite precision arithmetic may become significant.

Having an N-th order polynomial approximation of a function f on the standard interval

$$f_N(\xi) = \sum_{n=0}^{N} \tilde{f}_n \, \phi_n(\xi) \quad \xi \in [-1, 1]$$

or a set of its values in physical points $\{\xi_n\}_{n=0}^N$, we recognize

1. the *collocation differentiation*, which takes the set of physical function values $\{f(\xi_n)\}_{n=0}^N$, constructs the interpolation polynomial (l_n are the Lagrange characteristic polynomials (2.94))

   $$\Pi_f^N(\xi) = \sum_{n=0}^{N} f(\xi_n) \, l_n(\xi)$$

   and differentiates this interpolation

   $$\frac{\mathrm{d}f}{\mathrm{d}\xi} = \sum_{n=0}^{N} f(\xi_n) \, \frac{\mathrm{d}l_n}{\mathrm{d}\xi} \, .$$

   If only values at a discrete points are of interest, the above formula define the differentiation matrix

   $$D_{mn} = \left. \frac{\mathrm{d}l_n(\xi)}{\mathrm{d}\xi} \right|_{\xi=\xi_m}$$

   Values of these matrix entries for Gauss type quadrature points can be found in most of the books concerning spectral methods ([25], [8], [18], ...). The derivatives in the nodal points are then

   $$\left. \frac{\mathrm{d}f(\xi)}{\mathrm{d}\xi} \right|_{\xi=\xi_m} = \sum_{n=0}^{N} D_{mn} f(\xi_n) \, .$$

2. the *differentiation in the transformed space*, which takes the function expansion as primal. If derivatives of the expansion functions are known, we obtain

   $$\frac{\mathrm{d}f}{\mathrm{d}\xi} = \sum_{n=0}^{N} \tilde{f}_n \, \frac{\mathrm{d}\phi_n}{\mathrm{d}\xi} \, . \tag{2.103}$$

   Evaluation of the derivative in the discrete points of the physical space can be again represented by a matrix vector product

   $$\left. \frac{\mathrm{d}f(\xi)}{\mathrm{d}\xi} \right|_{\xi=\xi_m} = \sum_{n=0}^{N} D_{mn} \tilde{f}_n \, ,$$

   where

   $$D_{mn} = \left. \frac{\mathrm{d}\phi_n}{\mathrm{d}\xi} \right|_{\xi=\xi_m} \, .$$

   E.g. evaluation of the basis derivatives are not complicated in case of the basis (2.93), since values of $\left(J_n^{(\alpha,\beta)} \right)'$ satisfy the recurrence formula (47).

Both approaches coincide in the case, when $f \in \mathcal{X}_N$.[12]

Because of the finite precision effects in differentiation, it is advantageous to formulate the problems in spaces concerning derivatives as low as possible. Beside the smoothness of the solution the accuracy suffer from evaluation of the higher order derivatives.

Figure 2.4 illustrates the methods and mentioned accuracy lowering on a concrete function.

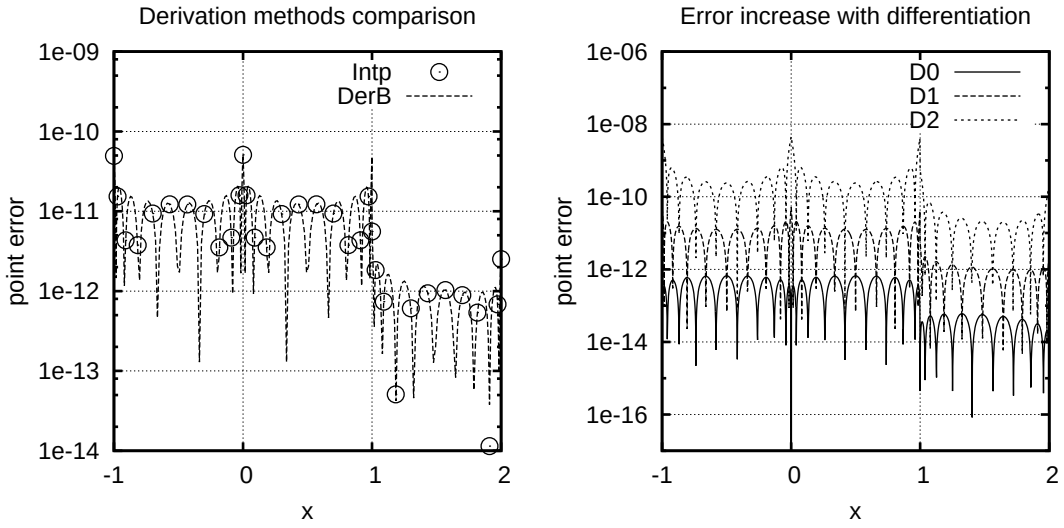


Figure 2.4: LEFT: Error in derivative of $cos(x)$ comparing the method using interpolation, *Intp*, and basis-derivation, *DerB*. RIGHT: Lowering of the approximation accuracy with differentiation-$D0 \sim u = cos(x)$, $D1 \sim \mathrm{d}u/\mathrm{d}x$, $D2 \sim \mathrm{d}^2u/\mathrm{d}x^2$. Both plots refer to the spectral element method, when domain $\Omega = [-1 : 2]$ is divided equidistantly to three elements and the local basis consists of $M = 10$ modes (polynomial order 9).

2.2.4 Integration

Both the matrix elements and data, as defined in MWR (2.70), have integral forms, which can't be evaluated analytically in general.

Already mentioned test and trial functions are based on the Jacobi polynomials or trigonometric series (sec. 2.2.2). The former are not known in a closed form [13] and the latter often lead to the elliptic integrals. However, highly accurate quadrature formulas successfully replace the continuous integration in these cases. Particularly, for the trigonometric integrands the *fast Fourier transform* (FFT) is available and the Gauss type quadratures applies to polynomials. Accuracy of mentioned formulas is limited only by the number of available discrete points and computer precision.

The Gauss quadrature is of special interest in this work, since the trigonometric expansion applies only to periodic problems. We will briefly summarize

[12] If $f \in \mathcal{X}_N$, then $\hat{f}_n = \tilde{f}_n$.

[13] Values of the Jacobi polynomials in discrete points are available through the recurrence formulas (D)

its properties[14].

Gauss integration

A quadrature formula based on Q quadrature nodes $\{\xi_i\}_{i=0}^{Q-1}$ takes form of a weighted sum

$$\int_{-1}^{1} f(x)\mathrm{d}x = \sum_{i=0}^{Q-1} w_i\, f(\xi_i) + E(f)\,, \tag{2.104}$$

where w_i are the quadrature weights, $\{f(\xi_i)\}_{i=0}^{Q-1}$ values of integrand in ξ_i and $E(f)$ an error term. Equidistant grid of nodes, which is suitable for integration of trigonometric integrands, leads to erroneous behaviour if applied to evaluation of higher order polynomials. An alternative is provided in the Gauss quadrature formulas, which define specific distribution of $\{\xi_i\}_{i=0}^{Q-1}$. These are defined on the *standard* interval $[-1, 1]$. Integration over arbitrary interval is achieved by appropriate mapping.

Three types of the Gauss quadratures are distinguished accordingly to inclusion of intervals end-points $\xi_0 = -1$ and $\xi_{Q-1} = 1$.

1. *Gauss* quadrature, avoids the endpoints of the computational interval and is accurate for polynomial integrands up to order $2Q - 1$

2. *Gauss-Radau* quadrature involves one of the endpoints -1 or 1 and is accurate for polynomials of order up to $2Q - 2$

3. *Gauss-Lobatto* quadrature involves both the points ± 1 and is accurate up to order $2Q - 3$

More precisely, the formulas provide computation of a weighted integral

$$\int_{-1}^{1} w(\xi)u(\xi)\,\mathrm{d}\xi\,, \tag{2.105}$$

where the weight $w(\xi) = (1-\xi)^\alpha(1+\xi)^\beta$ is specified by the kind of Jacobi polynomial $J_n^{(\alpha,\beta)}$ used in definition of the quadrature nodes $\{\xi_i\}_{i=0}^{Q-1}$. Presence of the weight may complicate formulation of the algorithm, so the Gauss points and weights based on the Legendre polynomials ($\alpha = \beta = 0$) are used preferably.

Denoting by $\{\xi_{i,Q}^{(\alpha,\beta)}\}_{i=0}^{Q-1}$ the Q zeros of the Jacobi polynomial $J_Q^{(\alpha,\beta)}(\xi)$, the quadrature points and weights for the Legendre type formulas are

1. *Gauss-Legendre*

$$\xi_i = \xi_{i,Q}^{(0,0)}, \quad i = 0, \ldots, Q-1\,, \tag{2.106}$$

$$w_i^{0,0} = \frac{2}{1-(\xi_i)^2}\left[\frac{\mathrm{d}}{\mathrm{d}\xi}(J_Q^{(0,0)})|_{\xi=\xi_i}\right]^{-2}, \quad i = 0, \ldots, Q-1 \tag{2.107}$$

[14]For detail description of Gauss type quadrature we refer reader to [34]

2. *Gauss-Legendre-Radau* (including left boundary point $\xi_0 = -1$)

$$\xi_i = \begin{cases} -1, & \\ \xi^{0,1}_{i-1,Q-1}, & i = 1, \ldots, Q-1 \end{cases}, \tag{2.108}$$

$$w_i^{0,0} = \frac{1-\xi_i}{Q^2[J^{(0,0)}_{Q-1}(\xi_i)]^2}, \quad i = 0, \ldots, Q-1 \tag{2.109}$$

3. *Gauss-Legendre-Lobatto*

$$\xi_i = \begin{cases} -1, & i = 0, \\ \xi^{1,1}_{i-1,Q-2}, & i = 1, \ldots, Q-2, \\ 1, & i = Q-1, \end{cases} \tag{2.110}$$

$$w_i^{0,0} = \frac{2}{Q(Q-1)[J^{(0,0)}_{Q-1}(\xi_i)]^2}, \quad i = 0, \ldots, Q-1 \tag{2.111}$$

Formulas employing various values of α and β are used in higher dimensions, reflecting the structure of multidimensional basis. Particular definitions of ξ_i and w_i for more general Gauss quadratures can be found in most of the textbooks concerning the spectral methods (e.g. [25], [18],...; some reference values are available in [38]).

In fact, all the Gauss rules employ Lagrange-type polynomial interpolation of the integrand through the quadrature points. Their weights w_i are defined as integrals over Lagrange characteristic polynomials $l_i(x)$ (2.94) belonging to nodes $\{\xi_i\}_{i=0}^{Q-1}$

$$w_i = \int_{-1}^{1} l_i(x)\mathrm{d}x \quad i = 0, \ldots, Q-1\,. \tag{2.112}$$

Quality of the Gauss integration is given by the interpolation polynomial, which is defined by distribution of $\{\xi_i\}_{i=0}^{Q-1}$. It follows from the theory of electrostatic points and optimisation of the Lebesgue constant ([18]), that none of the above mentioned point distributions provides optimal interpolation and integration.
Thanks to the accuracy of these formulas, orthogonality relations as (2.89) are preserved to the truncated expansions based on Jacobi polynomials, when sufficient number of quadrature points is available. Concretely taking $i, j < Q - 3/2$ for Gauss-Legendre-Lobatto points and weights, we get the *discrete orthogonality relation*

$$\sum_{i=0}^{Q-1} w_q\, J_p^{(\alpha,\beta)}(\xi_i) J_q^{(\alpha,\beta)}(\xi_i)\, w_{(\alpha,\beta)}(\xi_i) = C_{pq}\delta_{pq}\,. \tag{2.113}$$

Accurate integration of high order polynomials allows exact evaluation of elements in the system matrix (2.66) without limitation to order of the generating basis.

The number of quadrature points is primarily optimised for exact evaluation of mass matrix elements. But special attention must be paid to forward integration of non-linear functions emerging in data (RHS).

Aliasing error

Aliasing error is intrinsic to discretely evaluated coefficients in projection (2.63). It emerges if the number of quadrature points Q is optimised for accurate projection of functions from space $\mathcal{X}_N$, but the projected function is an element of higher order space $\mathcal{X}_H$, where $H > N$. Without loss of generality, the phenomenon will be shown using expansion to Legendre polynomials $\{\phi_n\}_{n=1}^{N} = \{U_m^{(0)}\}_{m=0}^{M}$ and use of Q point Gauss-Legendre-Lobatto quadrature (GLL(Q)). The discrete space is indexed $n = 1, \ldots, N$ in (2.77), so restricting to a single element ($E = 1$), $N = M + 1$ and M is the highest polynomial order occurring in the expansion basis.

It would seem, that the coefficient, belonging to the highest order mode, ϕ_M, from the transformation of arbitrary function f to the space $\mathcal{X}_N$, requires maximally $Q = M + 2$ points for GLL to be accurate. Unfortunately the number of quadrature points needed is given by the space, where f is located before the transform. If $f \in \mathcal{X}_H$, $H > N$, then there exists $\bar{M}$ such, that $\tilde{f}_m$, where $\bar{M} < m \leq M$ is not exact, if GLL($M + 2$) is used. These coefficients include the aliasing error, which can be precisely described.

Let $f(x)$ is a function with expansion

$$f(x) = \sum_{m=0}^{\infty} \hat{f}_m \, \phi_m(x) \,, \tag{2.114}$$

where $\hat{f}_m$ are known exact projection coefficients and $\phi_m(x) = U_m^{(0)}(x)$. Evaluation of an n-th coefficient $\tilde{f}_n$ ($0 \leq n \leq M$) using the GLL(Q), $Q = M + 2$, states

$$\tilde{f}_n \; \|\phi_n\|_{\mathcal{X}}^2 = \sum_{q=0}^{Q-1} f(\xi_q) \, \phi_n(\xi_q) \, w_q \,. \tag{2.115}$$

Substitution of the infinite and accurate expansion (2.114) to (2.115) results in

$$\begin{aligned} \tilde{f}_n \; \|\phi_n\|_{\mathcal{X}}^2 \; &= \sum_{q=0}^{Q-1} \left(\sum_{m=0}^{\infty} \hat{f}_m \phi_m(\xi_q) \right) \phi_n(\xi_q) w_q = \\ &= \sum_{m=0}^{M} \hat{f}_m \sum_{q=0}^{Q-1} \phi_m(\xi_q) \phi_n(\xi_q) w_q + \sum_{m=M+1}^{\infty} \hat{f}_m \left(\sum_{q=0}^{Q-1} \phi_m(\xi_q) \phi_n(\xi_q) w_q \right) . \end{aligned} \tag{2.116}$$

The first term contains the discrete orthogonality relation (2.113), while the second represents the aliasing error A_N of the N-th coefficient

$$\tilde{f}_N = \hat{f}_N + A_N \,. \tag{2.117}$$

Aliasing error highlights the fact, that discrete orthogonality among basis functions, is limited only to that subset, for which is the numerical integration accurate and is lost for the remainder. Overall error for the whole transformation is then

$$A_N^f = \sum_{n=0}^{M} \left(\sum_{m>M}^{\infty} (\phi_n, \, \phi_m) \, \hat{f}_m \right) = \sum_{m>M}^{\infty} (I_N \phi_m) \hat{f}_m \,. \tag{2.118}$$

Existence of this error evoked discussions about applicability of the spectral methods in its early development, since it was not known, if the coefficients may be computed with sufficient accuracy.

The aliasing error occurs rather in the testing step of the problem in MWR, because the number of the grid points is usually optimised to accurate evaluation of the integrals in elements of the mass matrix, while the function on RHS is from an unknown space.

Applying numerical integration, we do not have projection $P_{\mathcal{Y}}^N$, but its discrete counterpart $I_{\mathcal{Y}}^N$ connected with the error. It follows from orthogonality of the polynomial basis, that

$$\|f - I_N f\|^2_{L^2_w[-1,1]} = \|f - P_N f\|^2_{L^2_w[-1,1]} + \left\|A_N^f\right\|^2_{L^2_w[-1,1]} . \tag{2.119}$$

and the overall estimate (2.57) then becomes

$$\|u - u_N\| \leq \left\|u - I_{\mathcal{Y}}^N u\right\| + \left\|I_{\mathcal{Y}}^N u - P_{\mathcal{Y}}^N u\right\| + \left\|u_N - P_{\mathcal{Y}}^N u\right\| . \tag{2.120}$$

It can be shown, that both $\left\|u - I_{\mathcal{Y}}^N u\right\|$ and $\left\|I_{\mathcal{Y}}^N u - P_{\mathcal{Y}}^N u\right\|$ are of the same order as $\left\|u - P_{\mathcal{Y}}^N u\right\|$, c.f. Theorem 4. This fact is summarized in the following theorem (Hesthaven [18]), which is valid for spectral methods with basis constructed on ultraspherical polynomials.

Theorem 7. *(Discrete truncation error estimate)*
For $u \in H_w^p[-1,1]$ where $p > 1/2 \max(1, 1+\alpha)$, there exists a constant, C, which depends on α and p but not on N, such that

$$\|u - I_N u\|_{L^2_w[-1,1]} \leq C N^{-p} \|u\|_{H^p_w[-1,1]} , \tag{2.121}$$

where $I_N u$ is constructed using ultraspherical polynomials, $U_n^{(\alpha)}(x)$, with $|\alpha| \leq 1$. This holds for Gauss and Gauss-Lobatto based interpolations.

More specific results for higher norms and particular Gauss-type quadrature points are available (see e.g. [18], [8], [9]).

Results in the above theorem refer to asymptotic behaviour as $N \to \infty$ and it should be traced carefully for limited N. The aliasing error is not present if $f(x) \in \mathcal{X}_N$ (then $\hat{f}_n = \tilde{f}_n = 0 \quad \forall n > N$) or its expansion coefficients decay is such, that it covers the whole range of the computer precision, then the aliasing error coincides with the error given by the finite arithmetic.

When the approximation space can't be constructed rich enough, the technique of aliasing removal should be considered. This is the case of approximation of the convective term $\mathbf{v} \cdot \nabla \mathbf{v}$, representing a quadratic nonlinearity in the Navier-Stokes equations.

Aliasing removal in polynomial spectral methods consists in increasing the number of quadrature points Q. It is so called the "3/2 rule", which dealiases projection of a function $f \in \mathcal{X}_{2N-1}$ onto the space $\mathcal{X}_N \equiv \mathcal{X}_{M+1}$.[15] The highest order integrand of the transformation is then of order $3M$, so $Q = \frac{3}{2}(M+1)$ is needed for exact integration using GLL formula.

Following example shows the aliasing error for the basis (2.93):

[15] Maximal polynomial order in space $\mathcal{X}_{2N-1} \equiv \mathcal{X}_{2M+1}$ is double the order of $\mathcal{X}_N$ since $N = M + 1$, where M is the maximal order of the expansion.

- Input: converged coefficients $\{\hat{f}_m\}_{m=0}^{M}$ of a function $f \in \mathcal{X}_N$ and coefficients of its square $\{\widehat{(f^2)}_m\}_{m=0}^{M}$
- Evaluate f in physical space both in Q quadrature points ($Q = M+2$) and $\tilde{Q}$ quadrature points ($\tilde{Q} = 3(M+1)/2$)
- Square the function in the physical space in both sets of points
- Evaluate the discrete inner products $(f^2, \phi_m)_Q$, $(f^2, \phi_m)_{\tilde{Q}}$ $m = 0, \ldots, M$. The difference shows the aliasing error, see fig. 2.5.
- Finalise the projection problem $\mathbb{I}u = f^2$ and compare obtained coefficients with $\{\widehat{(f^2)}_m\}$ from the input, see fig. 2.6.

The above example is illustrated in figures 2.5 and 2.6

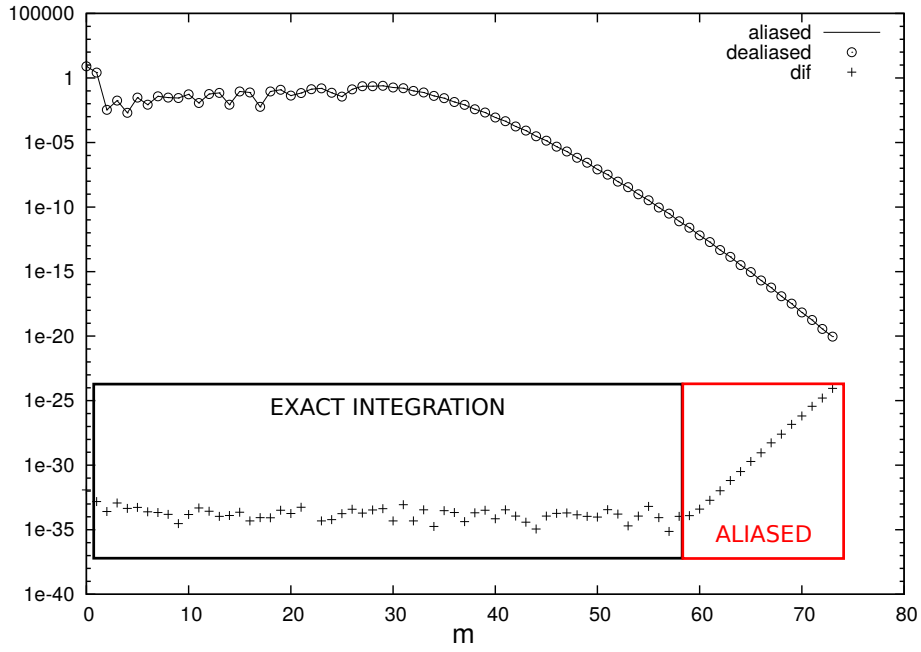


Figure 2.5: Values of the RHS integrals $(\int_\Omega f\phi_m \mathrm{d}x \quad m = 0, \ldots, M)$ evaluated using GLL$(M+2)$, *aliased*, and GLL$(3(M+1)/2)$, *dealiased*. We work on $\Omega = [-1:30]$ and $f = \cos^2(x)$, so the expansion with $M = 74$ produces the truncation error, since the coeffs $\tilde{f}_m$ decay to machine precision level for $H \approx 91$. Index $\bar{M}$, such that aliasing error occurs in $\tilde{f}_m \quad m > \bar{M}$, is possible to calculate ($\bar{M} \approx 2M + 2 - H$), but H is usually unknown. Values denoted as *dif* are differences between aliased and dealiased coefficients, $\bar{M} = 59$ in this case. The aliasing error grows (exponentially) in interval $m \in [60, 74]$, but it do not exceed the value of the highest expansion coefficient.

This simple example shows, that the aliasing error decays in similar manner as the truncation error, so it is not important in case of converged spectrum. However it strongly influence the transformation if the spectral coefficients are not converged. In such a cases using the dealiasing technique is recommended.

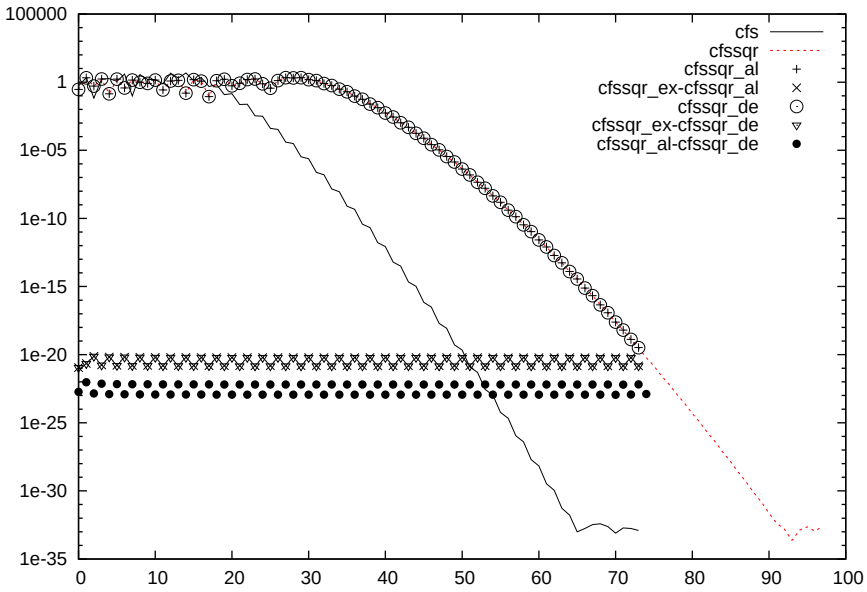


Figure 2.6: Coefficients cfs belong to function $\cos(x)$, $x \in [-1, 30]$, $cfssqr$ are coefficients of $\cos^2(x)$, $x \in [-1, 30]$. Ending "al" denotes aliased, "de" denotes dealiased and "ex" denotes exact. The differences of exact-aliased, and exact-dealiased are nearly the same, but difference of aliased-dealiased shows the difference. It is the truncation error, what dominates in this example. The aliasing error do not develop as in fig. 2.5, because the coefficients belong to the solution of projection problem $\mathbb{M}u = f$, where inversion $\mathbb{M}^{-1}$ is full and distributes the aliasing error included in the coefficients of RHS f over all the expansion coefficients.

Integration in higher dimensions

Gauss and Gauss-Radau rules do not include some of the boundary points. This fact can be very profitably used if there are geometrical singularities in the computational domain (integration on triangle). Integration on quadrilaterals in higher spatial dimensions is given as tensor product of 1D rules, in 2D we have

$$\int_{[-1:1]\times[-1:1]} f(x,y)\mathrm{d}x\mathrm{d}y = \sum_{p,q=0}^{Q_1,Q_2} w_{pq} f(x_p, y_q)\,,$$

where $w_{pq} = w_p w_q$. Integration on triangles brings complications and the quadrature rules are not, up to the authors knowledge, closed, especially for high order integrands. But spectral methods are applicable also when using triangular shape of the standard element. The trick is in transformation of standard quadrilateral to triangle. Similarly to construction of the basis on the triangular standard element, we denote η_1, η_2 local coordinates on the standard quadrilateral and ξ_1, ξ_2 coordinates on the standard triangle, the transformation follows

$$\begin{aligned} \eta_1 &= 2\frac{1+\xi_1}{1-\xi_2} - 1 \\ \eta_2 &= \xi_2 \end{aligned} \tag{2.122}$$

Transformation (2.122) has singularity at $\xi_2 = 1$ which excludes use of the Gauss-Lobatto type of quadrature in the η_2 resp. ξ_2 direction. Solution is in enforcement

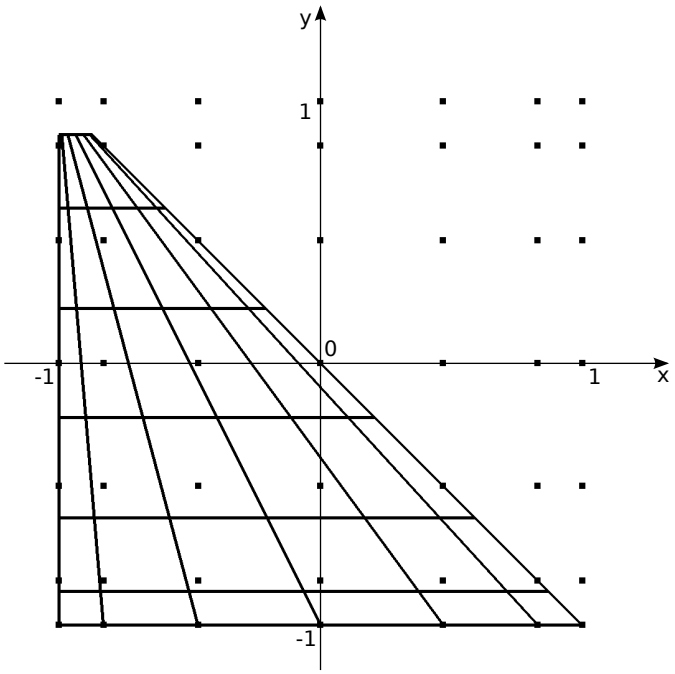


Figure 2.7: Gauss-Lobatto grid (dots) on standard quadrilateral $[-1,1]\times[-1,1]$ and standard triangle (grid connected by lines) as transformed quadrilateral (2.122) with Gauss-Lobatto points in x-direction and Gauss-Radau points in y-direction.

of the Gauss-Radau rule defined on $[-1:1)$ avoiding the point in the singularity of the transform. This approach results in quadratures on triangles, having no other limitation than computer precision. Distribution of the quadrature points used on the quadrilateral and triangular standard element is shown in figure 2.7.

Deformed domains

The subdomains Ω_e are transformed standard domains Ω_{std} in the multi-domain methods. This applies if the subdomain include a curved boundary of the computational domainfor triangular domain. As the Fourier expansion is defined on $[0,2\pi]$ or the Jacobi polynomials on $[-1,1]$, the standard (quadrilateral in 2D) domain is defined as $\Omega_{std}=[0,2\pi]^D$ or $\Omega_{std}=[-1,1]^D$, where D denotes spatial dimension of the problem. The mapping may be described as a change of coordinates between the physical $(x_1,\ldots,x_D)$ and the local $(\xi_1,\ldots,\xi_D)$ system

$$\mathbf{x}^e(\xi_1,\ldots,\xi_D):\Omega_{std}\to\Omega_e\,,$$

where the expansion basis is defined on the Ω_{std}. Especially in the case of high order approximations, where a large part of domain boundary is covered by a single subdomain, the mapping must be capable of describing the boundary geometry. The mapping is represented in the polynomial space built on the standard element. This *isoparametric transformation* (in 2D) reads

$$x_i^e(\xi_1,\xi_2)=\sum_{m=1}^{M_1^e}\sum_{n=1}^{M_2^e}x_{mn}^e\phi_{mn}(\xi_1,\xi_2)\quad i=1,2\,, \tag{2.123}$$

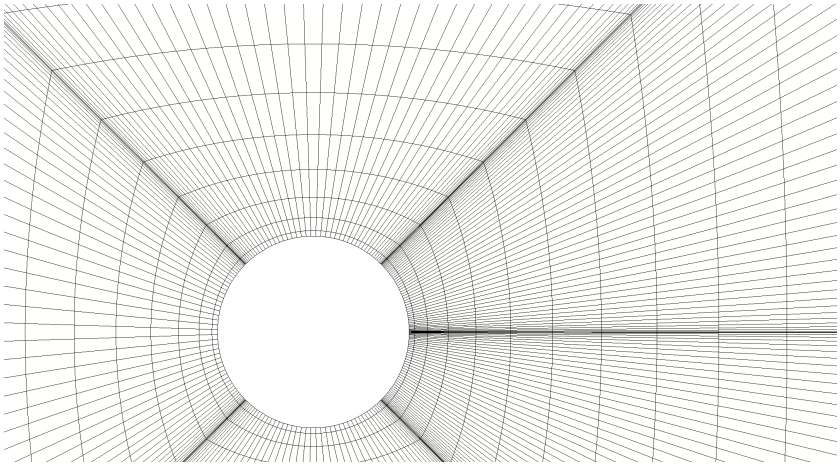


Figure 2.8: A detail of curved boundary in a high order approximation. Lines are connecting the Gauss-Legendre-Lobatto integration points.

where $\phi_{mn}(\xi_1, \xi_2)$ are the 2D expansion functions, e.g. (2.98) or (2.101). Then, $f(\mathbf{x}) = f(\mathbf{x}^e(\xi))$ and integration becomes

$$\int_{\Omega_e} f \mathrm{d}\mathbf{x} = \int_{\Omega_{std}} f(\mathbf{x}^e(\xi)) |\mathbb{J}_e| \mathrm{d}\boldsymbol{\xi} \,,$$

where

$$\mathbb{J}_e = \begin{pmatrix} \dfrac{\partial x_1^e}{\partial \xi_1} \dfrac{\partial x_2^e}{\partial \xi_1} \\ \dfrac{\partial x_1^e}{\partial \xi_2} \dfrac{\partial x_2^e}{\partial \xi_2} \end{pmatrix} \tag{2.124}$$

is the Jacobi matrix. The coefficients of isoparametric transformation (2.123) behave same as projection of a function to the trial space and decay with increasing index relatively to its smoothness and complexity. Values of testing integrals

$$\int_{\Omega_e} 1 \, \phi_{mn}(\xi) |\mathbb{J}_e| \mathrm{d}\boldsymbol{\xi} \quad m = 1, \ldots, M_1^e \, n = 1, \ldots, M_2^e$$

for a particular situation are illustrated in figure 2.29. In comparison to elements without curved boundaries, convergence of "projection" integrals

$$\int_{\Omega_e} f \, \phi_m n |\mathbb{J}_e|$$

as m and n increases, differs.

The mapping influences also the matrix elements in the algebraic system, so that it affects orthogonality among the basis functions as illustrated in fig. 2.30.

2.2.5 Spectral approach-demonstrations

In this section, aspects of high order computations will be examined in practice. We will concern Galerkin type method, $\mathcal{X}_N = \mathcal{Y}_N = \mathrm{span}\{\phi_n\}_{n=1}^N$, where ϕ_n are

given by (2.93). The forms in the MWR formulation will be always evaluated using the Gauss-Legendre-Lobatto integration.

For a number of smooth functions, we will observe fast decays of spectral coefficients as predicted in (2.84). Substantial difference in convergence between

- p-convergence: increasing the order of expansion on a fixed mesh (or whole computational domain)
- h-convergence: sub-dividing the computational domain at fixed order local expansion.

will be shown on numerically evaluated error $\|u - u_N\|_\infty$ for known functions u.

Decay of a sequence $\{\hat{f}_n\}_{n=0}^{\infty}$ may be classified as follows

Definition 3. *(Orders of convergence, see also [1])*

1. *The Algebraic index of convergence is the largest k for which*

$$\hat{f}_n \approx O(1/n^k) \qquad n \gg 1 \tag{2.125}$$

2. *If the convergence is faster than $1/n^k$ for any k, than the series has infinite order or exponential convergence*

$$\hat{f}_n \approx O(\exp(-m\, n^r)) \qquad n \gg 1 \tag{2.126}$$

   for $m = \text{const.}$ and $r = \text{const.} > 0$. For exponential convergence three rates are defined

   (a) *subgeometric, if $r < 1$ in 2.126*

   (b) *geometric, if $r = 1$ in 2.126*

   (c) *supergeometric, if $\hat{f}_n \approx O(\exp(-(n/j)\log(n)))$*

These types of convergences are easily distinguishable in linear-log and log-log plots, figure 2.9.

Character of the subgeometric convergence changes between the linear-log and log-log graphs, when in the first it is hardly distinguishable from the algebraic type and in second it mimics shape of the geometric and supergeometric convergence.

The coefficient spectra $\{\hat{f}_n\}_{n=0}^{\infty}$ or values of $\|u - I_N u\|$ for $N \to \infty$ form sequences, whose decay may be classified useing above definition.

Estimate (2.84) and theorem 7 states, that both the coefficient values and truncation error reach infinite asymptotic decay for smooth functions. Unless the situation for finite series of coefficients or error values can be hardly predicted the examples below show, that the asymptotic decay occurs for finite expansions.

The level, to which the coefficients are converged, is an important tool for estimate of the truncation error. This can be formulated as follows

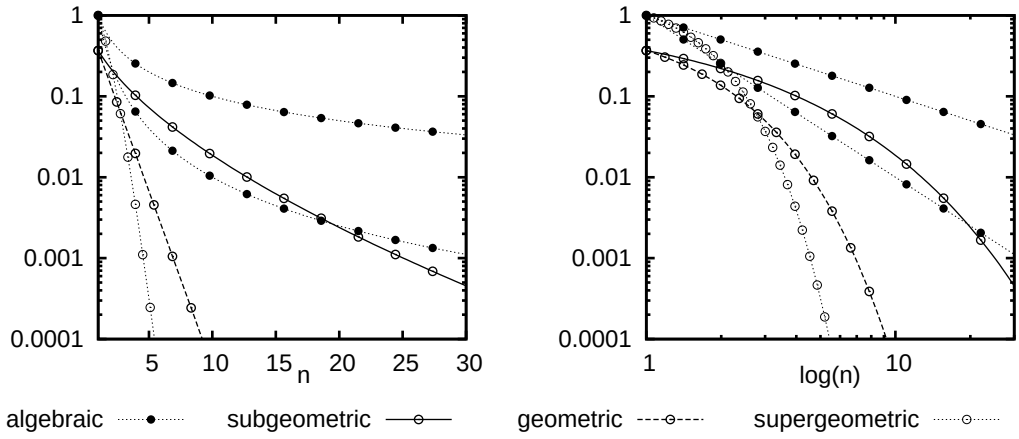


Figure 2.9: Illustration of convergence orders in a linear-log and log-log graph. All empty circles belong to the exponential decays. These graphs clarify usage of notion "infinite order" for the exponential types of convergence, since as $n \to \infty$ the negative slope of the curve is unbounded.

Theorem 8. *(Last coefficient error estimate; [1])*
The truncation error is the same order of magnitude as the last coefficient retained in the truncation for series with geometric convergence (Def. 3). Since the truncation error is a quantity, we can only estimate anyway (in the absence of a known, exact solution), we can loosely speak of the last retained coefficient as being the truncation error, that is:

$$\|u(x) - u_N(x)\| \sim O(|\hat{u}_N|) \tag{2.127}$$

If the series has algebraic convergence index, k, i.e., if $\hat{u}_n \sim O(1/n^k)$ for large n, then

$$\|u(x) - u_N(x)\| \sim O(N|\hat{u}_N|) \tag{2.128}$$

In the above theorem, the index N in the coefficient $\hat{u}_N$ refers to the highest order mode of the basis, what is $M-1$ in basis (2.93).

In the following examples, we will restrict to solutions of the Helmholtz equation

$$\begin{aligned} \text{1D}: &\quad \frac{\partial^2 u}{\partial x^2} - \lambda u = f \\ \text{2D}: &\quad \nabla^2 u - \lambda u = f\,, \end{aligned} \tag{2.129}$$

and its limit, Poisson equation ($\lambda = 0$), which state all the spatial steps in the computational algorithm for the incompressible heated fluid motion, cf. (2.32),(2.33) and (2.42).

In FV or FEM, the computational domain is decomposed to mesh of subdomains $\Omega_e : \Omega \approx \bigcup_e^E \Omega_e,\ E \gg 1$ and the order of the expansion is fixed on every Ω_e [16]. Only a small number of trial functions share a common support. If the

[16] Formally, FV takes only the first term of a polynomial expansion $P_e = 0\ \forall e = 1, \ldots, E$, since the approximation is constant on every Ω_e. Similarly in FEM only a few first terms are used, linear ($M = 1$) or quadratic ($M = 2$) expansions on Ω_e is mostly used.

mesh is globally characterized by parameter $h = \sup_e\{\mathrm{diam}(\Omega_e)\}$ convergence is achieved when refining the mesh ($h \to 0$; h-convergence).

Concerning a smooth problem, we compare solution to 1D Helmholtz equation as achieved by increasing number of elements in Ω (FEM) and increasing order of the basis (SM). Results are shown in figure 2.10.

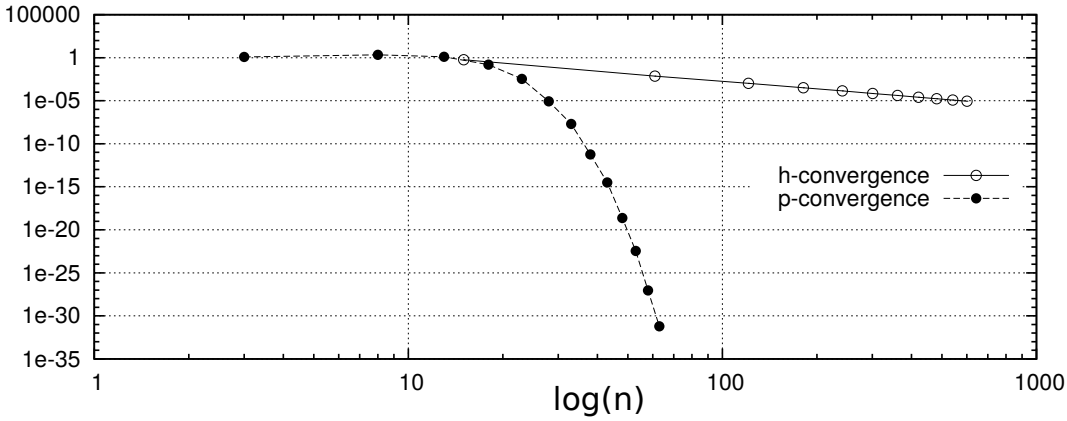


Figure 2.10: Numerically evaluated $\|u - u_N\|_{\mathcal{L}^\infty}$ as dependent on number of degrees of freedom, representing the solution $u = \cos(x)$ in equation $\dfrac{\partial^2 u}{\partial x^2} - 1000u = f$. "h-convergence" represents solution with fixed order 2 ($M = 3$), while increasing the number of elements, "p-convergence" is computed on the whole domain (one element) and increasing the order of approximation ($E = 1$). Problem was implemented in quad representation of real numbers. Both axis have logarithmic scale.

Both plots show the strength of spectral method as it is capable to calculate the solutions up to precision given by the (quad) computer precision in reasonable number of overall DOFs.

Matrix structures

Every coefficient is strongly influenced by values of the other coefficients in the spectral method, what can be deduced from the structure of matrix inverse fig. 2.12. The values of the matrix elements decay slowly with distance from the matrix diagonal in comparison to the low order method (fig.2.11), where the values decay exponentially. Expansion coefficients in first order method coincide with the function values, so the fast decay of coefficients in the matrix inverse implies, that function values in distant parts of the computational domain do not influence each other significantly.

The high order/spectral methods highlight the mathematical properties of the solutions. Improper boundary condition or singularity affects whole solution stronger than in low order methods. Therefore interdependence of spectral expansions may be seen as disadvantageous in some cases, because computational algorithm introduces various inaccuracies in boundary conditions especially and spectral methods do not dump these solution properties in contrast to lower order methods.

The three figures (fig. 2.11-2.13) compare the matrix structures belonging to a 1D problem as generated by the finite element method, spectral method and the spectral element method, while the same number of degrees of freedom is kept in the algebraic system. Only absolute values of the matrix elements are shown in these plots, since the scale in z direction is logarithmic.

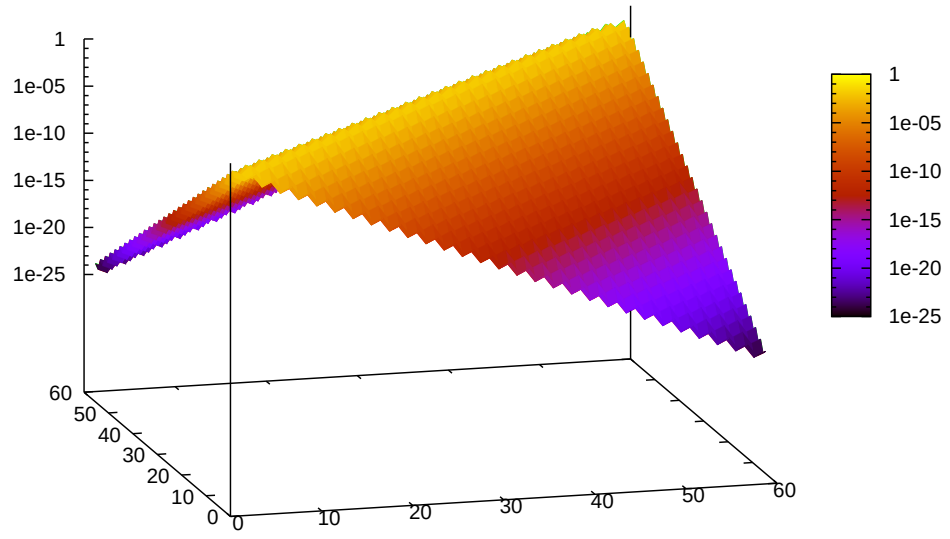


Figure 2.11: Structure of matrix inverse in case of piecewise linear ($M = 1\ \forall e$) approximation and decomposition of the domain to $E = 60$ elements.

1D Examples

Demonstrations in this section are mostly computed in quad computer precision (16 bytes), what can be seen as impractical in comparison with physical reality and accuracies achieved in physical experiments. It shows rather the mathematical aspects in the solved problems, because the precision used, refer to measuring of a hair thickness on scale of astronomical distances.

All the 1D calculations were performed by the Fortran code written by author, which may be switched easily to any of the single, double or quad precisions. The precision plays crucial role in all the following computations, since the converging coefficients fulfil the whole range given by the significand of the floating point number representation. Taking the expansion basis in form (2.93), the boundary modes carry the values of the function in the boundary points. Therefore the range scaled by coefficients converged to the chosen computer precision has the upper limit in the function values, while the lower is given by length of the significand in the computer representation of the floating point number. This range is ± 7 decimal places for single precision and the coefficients has values in range $[v * 10^{-7}, v]$. Similarly ± 16 decimal places are available for the double

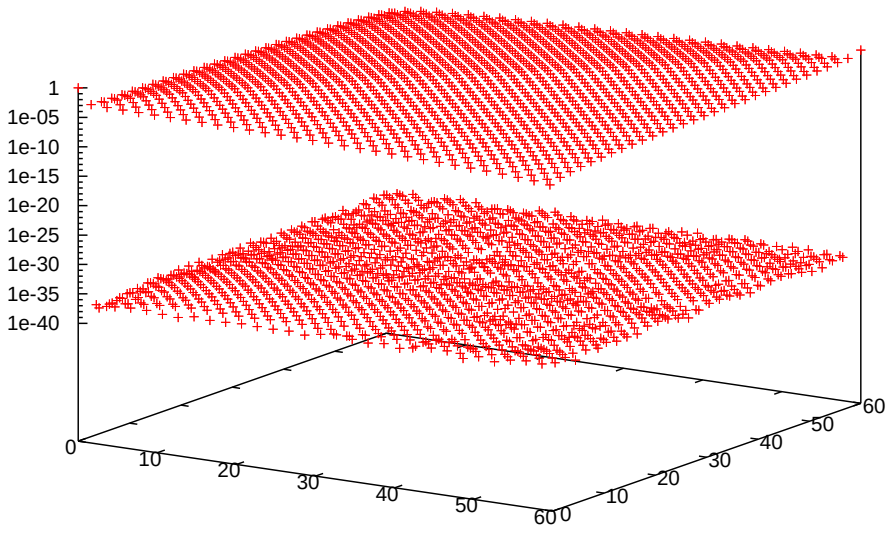


Figure 2.12: Inverse of the matrix in case of the spectral method ($M = 60$, $E = 1$)

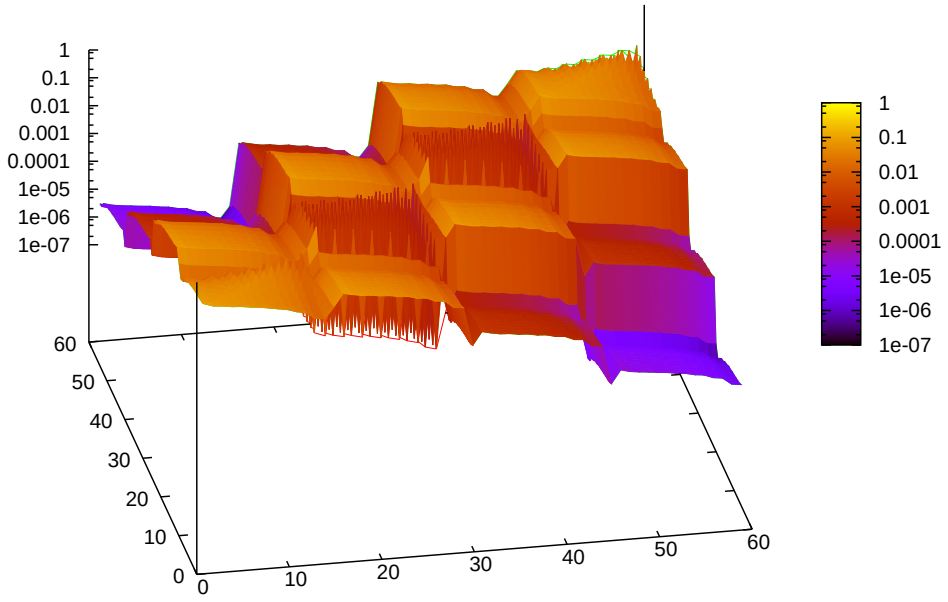


Figure 2.13: Structure of matrix inverse to the discretised 1D Helmholtz equation (2.129); spectral elements with $M = 16$, $E = 4$.

precision and ± 33 for the quad precision. We chose the quad precision in this section to highlight the spectral effects in figures.

The finite arithmetic limits also the maximal number of modes M, which can

be used in the basis. The computation crashes, if M is higher limits

- *single precision*: $M < 20$
- *double precision*: $M < 97$
- *quad precision*: $M > 500$

found experimentally using our code. Finding limit in the quad precision seems to be unreasonable presently, since the calculations with $M \sim 500$ are computationally expensive and usable only in special cases.

Geometry itself is not a source of singularities in 1D problems. The convergence in spectra is fully dependent on the regularity of the approximated functions in this case.

Taking a smooth function, the number of modes needed to resolve the function to machine precision, is given only by complexity of the function and length of the computational interval Ω. This number of modes is specific for a particular function and we will denote it *spectral length* for our purposes.

The RHS, f, and the Dirichlet boundary conditions are always set accordingly to the chosen exact function in (2.129). Influence of the value of λ is discussed as the last example.

Convergence limits given by finite arithmetics

The effect of various levels of the machine precision for the problem (2.129) is shown in figure 2.14 and also a pre-asymptotic behaviour when no convergence occurs. Expecting the coefficients with an exponential decay, the rate of exponential decay can be recognized among the coefficients, figure 2.14. The area of convergence is better distinguishable from the plot of the rate of the decay.

Spectral length vs. dimensions of the computational domain

The spectrum reflects the dimensions of the computational domain. Taking as an exact solution $u = \alpha \cos(\beta\, x)$, we solve the problem (2.129) on three intervals

1. $\Omega = [-1, 0]$
2. $\Omega = [-1, 6]$
3. $\Omega = [-1, 24]$

observing that, shorter the interval is, lower number of modes to suppress error to machine precision is needed (fig. 2.15). On the other side, M needed for resolution to the computer precision is not proportional to the length of the interval.

This fact motivates the decomposition of Ω to subdomains, where the spectral method is applied locally. Coefficients of the problem decomposed to 7 spectral elements is shown in fig. 2.16.

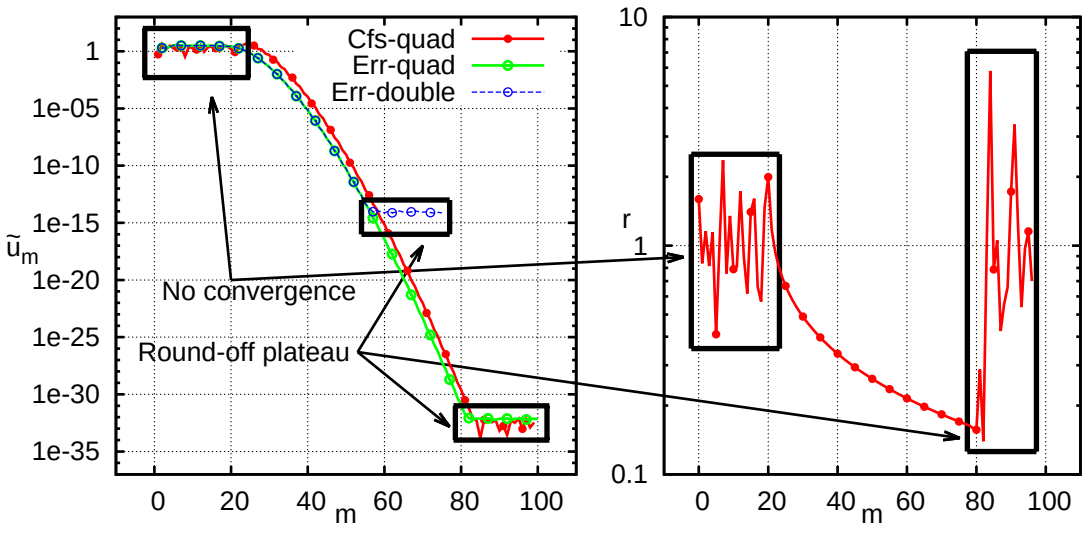


Figure 2.14: LEFT: Asymptotic exponential decay of expansion coefficients ($\tilde{u}_m$ $m = 0, \ldots, M$; $E = 1$; $M = N + 1$), *cfs-quad*, and numerically evaluated L^∞ norm for smooth functions. Comparison of computations in double *Err-double* and quad-precision *Err-quad*. Solutions are fully converged through whole extent of the particular number representation, what is seen in existence of the *round-off plateau*. RIGHT: Rate of exponential convergence, r, calculated from the sequence $\tilde{u}_m$, which is expected to satisfy $\tilde{u}_{m+1} = r\,\tilde{u}_m$. $r = \sqrt{\tilde{u}_{m+1}/\tilde{u}_{m-1}}$ to suppress effects caused by (anti)symmetry of the function over the interval.

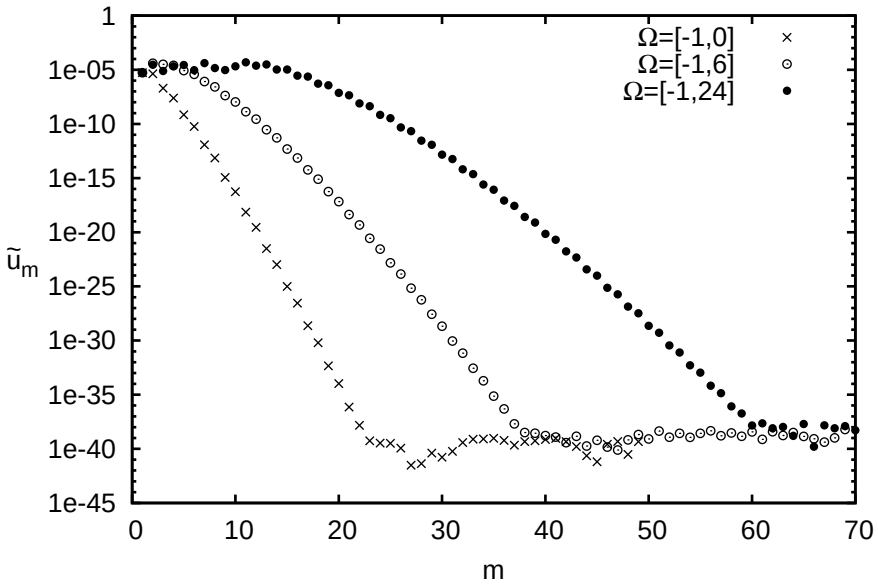


Figure 2.15: Comparison of spectra to the function $u = 10^{-5}\cos(x)$ on various intervals. Note, that the round-off plateau is on level 10^{-40}, since the maximal values of the function in every of the intervals is 10^{-5}.

Nonuniform quality of approximation over the computational interval

Quality of the approximation differs over the computational interval in dependence to the distribution of quadrature points. Taking

$$u(x) = e^{-1000x^2}\,, \tag{2.130}$$

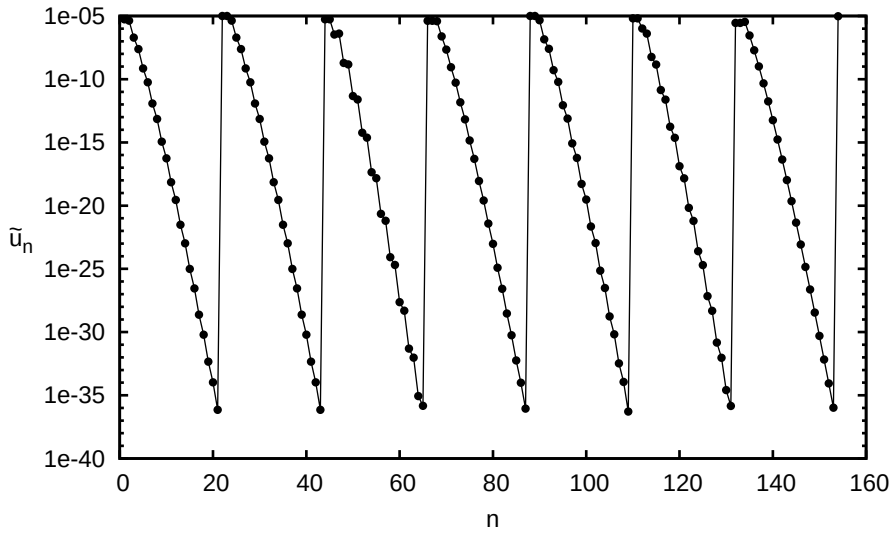


Figure 2.16: Function $u = 10^{-5}\cos(x)$ represented on 7 spectral elements of order 21 ($E = 7$, $M = 22$). Convergence of spectra to machine precision is achieved on all the elements. The total number degrees on freedom in the system, $N = 154$, needed for the computer precision accuracy is much higher than in case of the spectral method when approaching the same accuracy (single element, $N = M \sim 38$).

which change rapidly around $x = 0$, we observe, that quality of approximation slightly differs depending on the position of area of the rapid changes over the computational interval if the approximation is not at machine precision. Maximal numerical error on grid of 500 equidistant points in Ω was

1. 10^{-14} for $\Omega = [-0.2, 0]$ and
2. 10^{-12} for $\Omega = [-0.1, 0.1]$,

since $M = 40$ were used while full convergence demand $M > 80$. This phenomenon is explained in ([1]) as an impact of variable distribution of quadrature points over Ω. Especially in time dependent problems combined with low order approximation, this phenomenon results in variable quality of spatial approximation. However, this is not the case, when the spectral coefficients fall to the machine precision, so that the approximation is indistinguishable from the original function.

If the approximated function is symmetrical in Ω, every second coefficient falls to the machine precision as shows the figure 2.17, in the case of function (2.130). The reason is, that the basis functions (2.93, figure 2.3) are only symmetric or antisymmetric (except the boundary modes) and only symmetrical modes assert in this case.

Approximation using polynomial of order 500

The Helmholtz equation admits arbitrary value as a boundary condition. If the RHS and this boundary condition are not compatible, the "boundary layer" emerges in the solution, having gradient dependent on value of λ (c.f. 2.1.2).

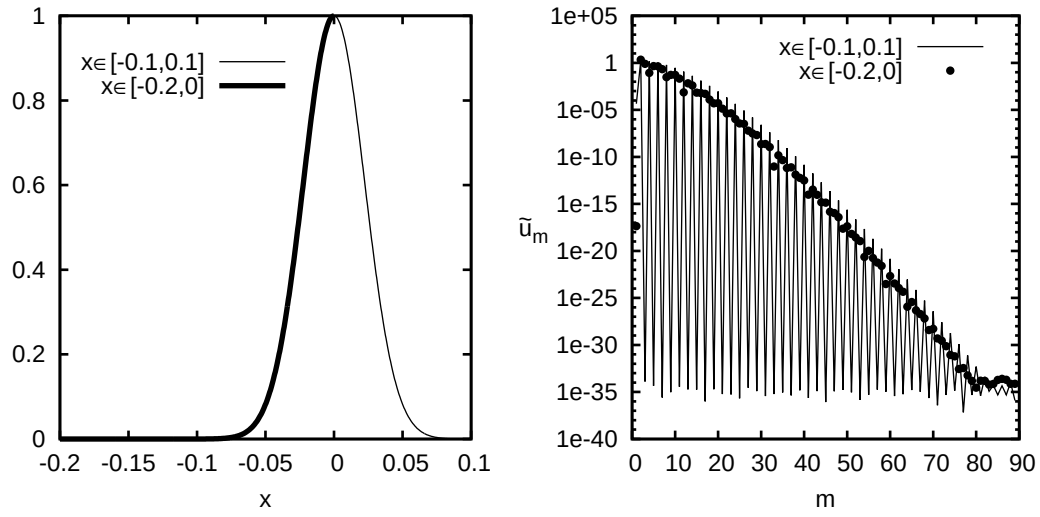


Figure 2.17: Effect of symmetry seen on jigsaw shape of spectra (right) of the function (left), if the function is symmetric over Ω or not.

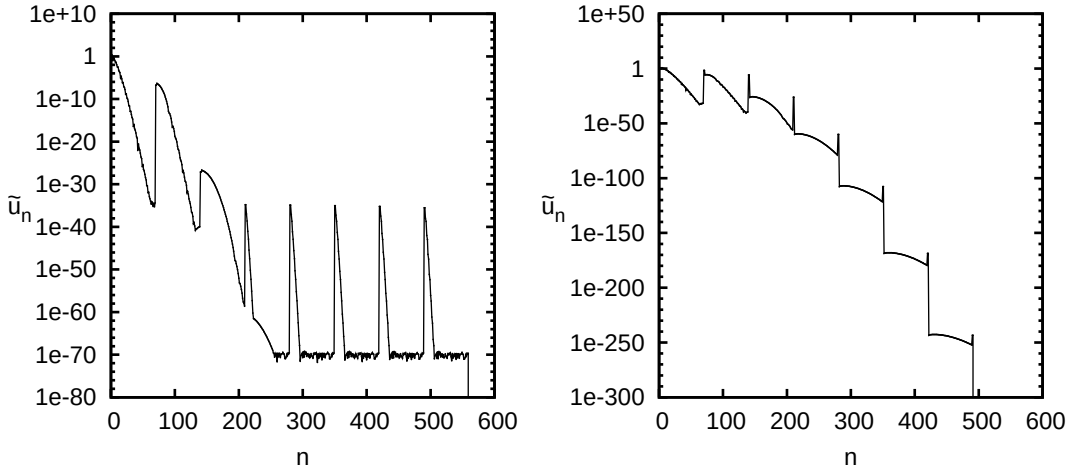


Figure 2.18: Function (2.130) as the exact solution to (2.129) for $x \in [0, 1]$. Ω is divided to 8 elements, $E = 8$,$M = 70$. Left: spectrum of the solution. Right: projection coefficients of function on the RHS. n denotes index of global modes $n = 1, \ldots, N = 553$

Simplest case with homogeneous RHS is shown in fig. 2.19, where the solution is the solution to the homogeneous problem, which is characterized by exponential solution. However, the solution is smooth and exponential decay occurs, notwithstanding it decays very slowly.

Nonhomogeneous problem is shown in fig. 2.20, where the RHS spectra decays rapidly. If the boundary condition would be set such, that the solution of the homogeneous problem would be zero, the spectra of the solution would resemble the decay of RHS. This is not the case of the solution presented and the spectra decays slower.

Construction of infinitely smooth problem, exhibiting the spectral accuracy,

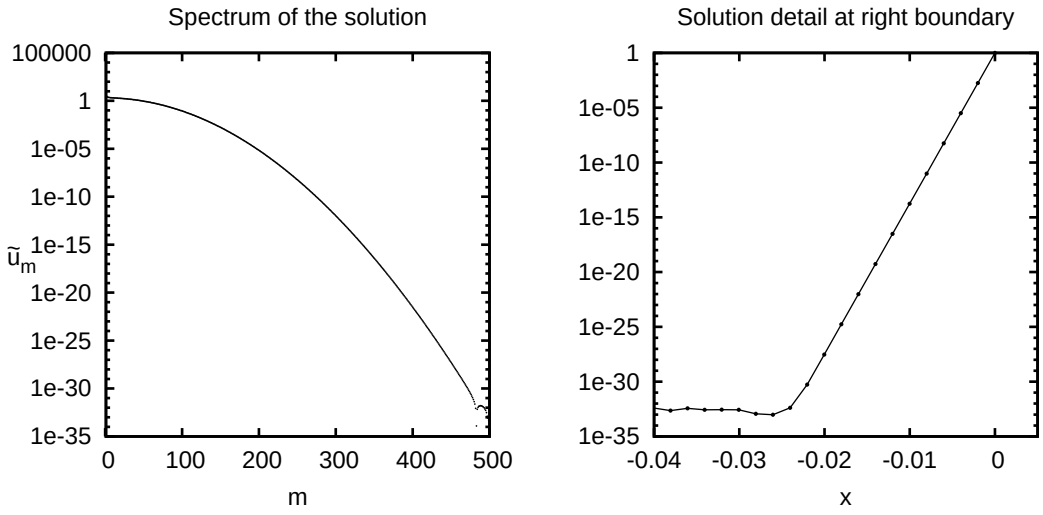


Figure 2.19: Solution to the Helmholtz equation with $\lambda = 10^7$, $f = 0$, $x \in [-1, 0]$ and $u(-1) = 0$, $u(0) = 1$. Extremely steep gradient, covering 33 orders occurs in the vicinity of the right boundary. Linear shape of the function in the log-scale plot suggest its exponential origin. Coefficients are converged to quad precision near the value 500! So the approximation to the solution is exact in view of finite arithmetic.

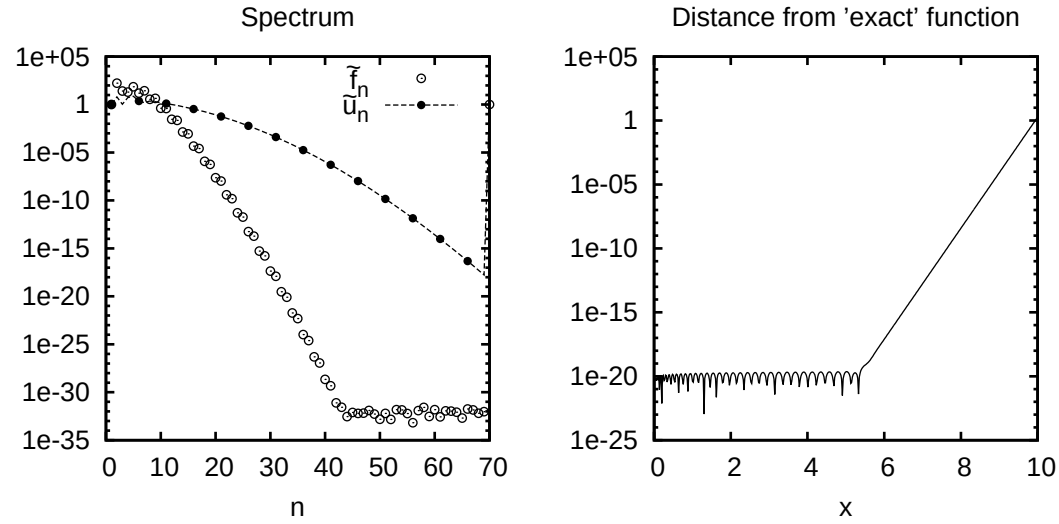


Figure 2.20: Solution to Helmholtz equation $\lambda = 100$, $x \in [0, 10]$, $f = (\partial^2/\partial x^2 - \lambda)\cos(x)$. Only left boundary condition is set to suit the solution $\cos(x)$ "expected" in the setting of RHS. Dirichlet condition on the right boundary is chosen to be $u(10) = 1 \neq \cos(10)$. In contrast to the spectra of the RHS, the solution spectra is not converged to the machine precision. Solution u recedes from $\cos(x)$ exponentially as approaching the prescribed boundary condition.

is not difficult in one-dimension, but in higher dimensions the smooth solutions for differential equations are seldom, especially as a consequence of complications in construction of a smooth computational domain.

2D Examples

Singularities in the 1D computations were given by the functions on the RHS, so construction of smooth problem was not difficult. This is not the case in higher dimensions, where the smooth solutions are seldom, since singularities are generated also from the shape of the computational domain itself. This can be shown on example of the Laplace equation $\nabla^2 u = 0$ on square domain with Dirichlet boundary conditions provided. It has been proven (Grisvard [15]), that the solution u in polar coordinates (r, Θ) behaves as

$$u \sim C\, r^2 \ln(r) \sin(2\Theta) \tag{2.131}$$

at the corners, where the Dirichlet condition is prescribed.

The third order derivatives of u at the corner are infinite, what results in loss of spectral accuracy. The problem is illustrated in figure 2.21. Occasional zero values among the spectral coefficients are substituted by "machine" zero (10^{-16}) and only absolute values of the coefficients are used to improve plotting in logarithmic scale of the following graphs. M_x and M_y are numbers of modes in x and y directions of the tensorial basis (2.98).

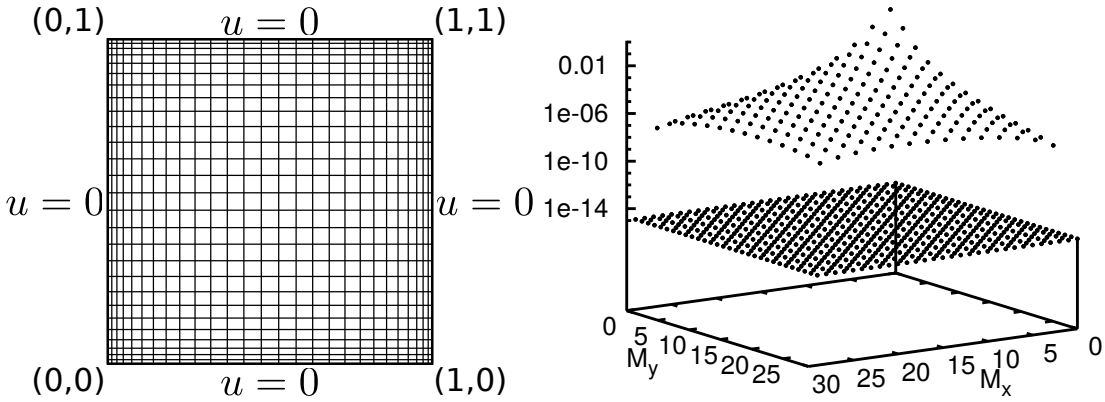


Figure 2.21: Expansion coefficients $\hat{u}_{xy}$ of solution to $\nabla^2 u = 1\,,\ \Omega = [0,1]^2$. Algebraic convergence is achieved in this case, since Dirichlet conditions $u = 0$ are prescribed on all the boundaries.

Mixed boundary conditions on domain with non-smooth boundary

We can avoid the corner singularity when Dirichlet boundary condition suits the known solution values on boundary. In special case of a constant RHS, a combination of the Dirichlet and Neumann type conditions is another case producing a smooth solution, as illustrated in fig. 2.22.

Domain with smooth boundaries

If the domain boundary is free of singularities, the coefficients also exhibit exponential decay, as shown in figure 2.23.

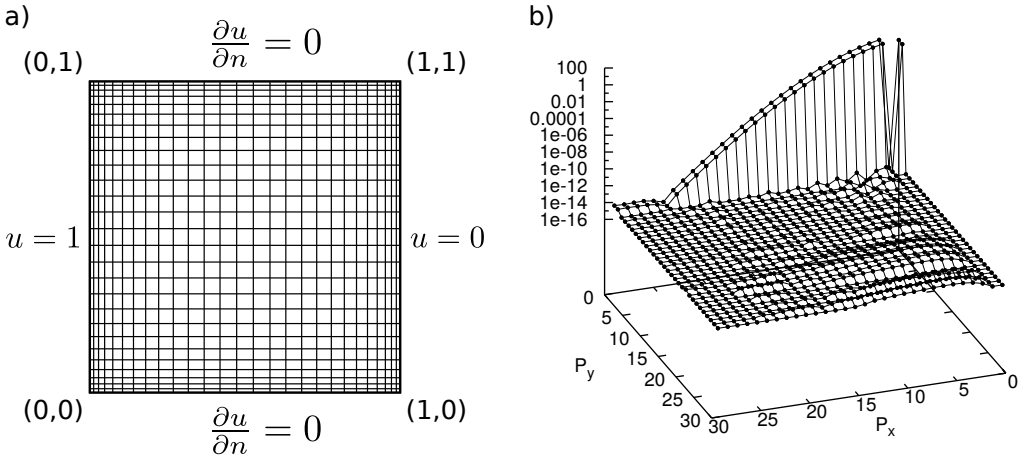


Figure 2.22: Exponential convergence of expansion coefficients occurs for combination of the Dirichlet $u = 0$ (boundaries with $x = 0$ and $x = 1$) and Neumann $\frac{\partial u}{\partial n} = 0$ condition (on boundaries $y = 0$ and $y = 1$) in spectral solution to $\nabla^2 u - 100u = 0$ on single domain $\Omega = [0, 1]^2$.

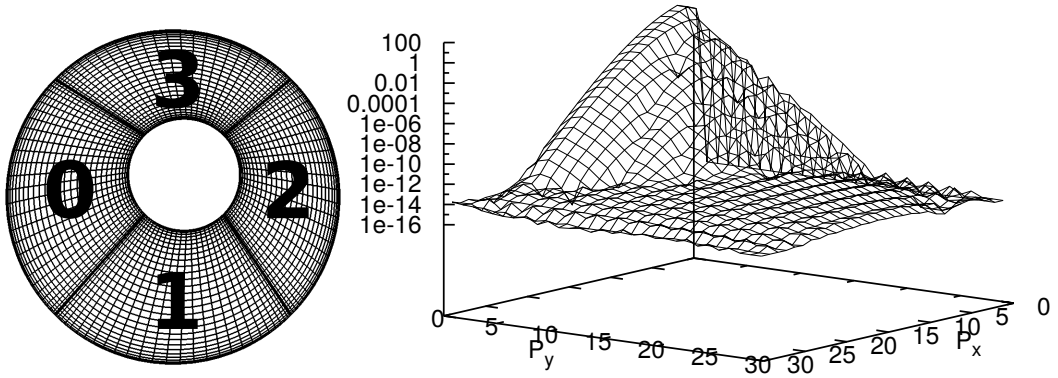


Figure 2.23: Visualization of the spectral coefficients $\hat{u}_{xy}$ of the solution u to the Helmholtz equation $(\nabla^2 - 100)u = 0$ with constant Dirichlet boundary conditions (zero on inner circle 1 on the outer circle). Domain (*left*) has smooth boundaries and the solution coefficients exhibit spectral convergence in both x and y direction. P_x and P_y here denote the indices in x,y direction. Coefficients in *right* belong to the element "2" (*left*).

Smoothness in finite arithmetic

Exponential decay in the spectra may be achieved for the Poisson problem also on the domain with sharp corners in case, when the approximated function and its derivatives up to a sufficiently high order vanish at the corner, as shown in figure 2.24 for increasing order of boundary-vanishing derivatives. The solution with high order vanishing derivatives accepts both the homogeneous Dirichlet or Neumann conditions and the results are indistinguishable.

Smoothness of the domain boundary

Similarly to the 1D case described in 2.1.2, the equation admits solution to the problem, where the boundary conditions are not compatible to the RHS. However,

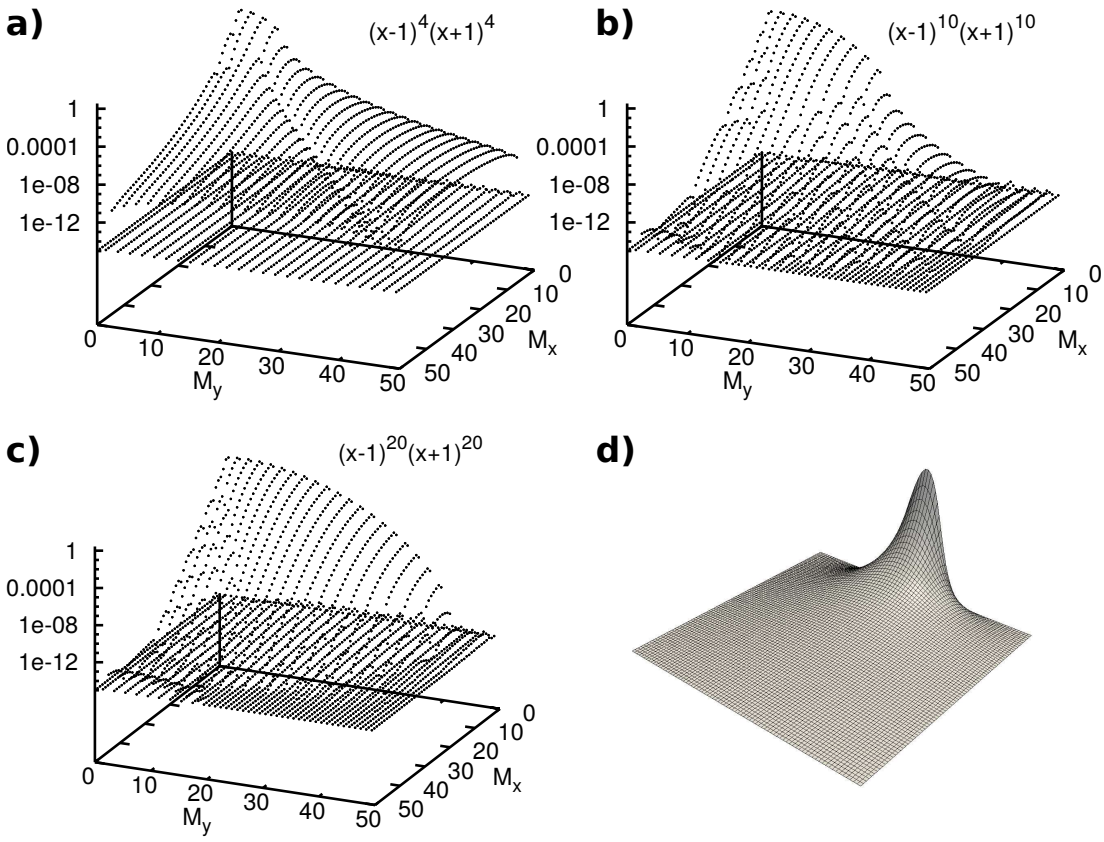


Figure 2.24: Poisson equation solved for the Dirichlet boundary conditions on standard square $[-1:1]^2$. 2D spectra for three cases, when the Dirichlet values on three of the four boundaries are $u|_{\partial\Omega-\Gamma}=0$ and $u|_\Gamma=(x-1)^2(x+1)^2$ at "**a)**", $u|_\Gamma=(x-1)^{10}(x+1)^{10}$ at "**b)**" and $u|_\Gamma=(x-1)^{20}(x+1)^{20}$ iat "**c)**. Solution of the third case is at "**d)**" position for illustration. The higher the derivative vanishing at the boundary is, the shorter the slowly converging part of the spectra occurs.

in such a case any singularity of the boundary geometry reflects in the solution, resp. influences the decay of the coefficients in the spectra. If the computational domain is smooth, it do not introduce singularities to solutions of (2.129) and exponential decay may occur, as illustrated in fig. 2.23. The exponentially decaying spectra occurs also for the non-smooth domains boundary, if the smooth function is on the RHS and the boundary conditions are compatible with the RHS.

The solution spectrum reflects singularities present in a higher derivatives of the boundary curve, as illustrated on spectra of the solution (see fig. 2.25)

$$\begin{aligned}\nabla^2 u &= f\\ f &= 0\\ u|_{\Gamma_I} &= 0\\ u|_{\Gamma_O} &= 0\,.\end{aligned} \tag{2.132}$$

The domain boundary do not have continuous second derivatives in this case, since it consists of concatenations of the linear and circular arcs. The exponential convergence is then limited to a few first coefficients and then it continues

with a slow decay reflecting the singularity (fig. 2.25 b)). Regardless the slow convergence, the inaccuracy in the solution is hardly detectable from its values, since the coefficients are converged to values $\sim 10^{-7}$ (fig. 2.25 c)). The smoother the boundary is, the weaker singularity occurs in the solution and longer part of the spectra decay exponentially, c.f. fig. 2.24. Finally, spectra of solution $u = cos(x)$ to

$$\begin{aligned} \nabla^2 u &= f \\ f &= -cos(x) \\ u|_{\Gamma_I} &= cos(x) = u|_{\Gamma_O}\,, \end{aligned} \tag{2.133}$$

where BC and RHS are compatible, is shown in fig. 2.25. This solution exhibits exponential decay of coefficients, while computed on a domain with a non-smooth boundary.

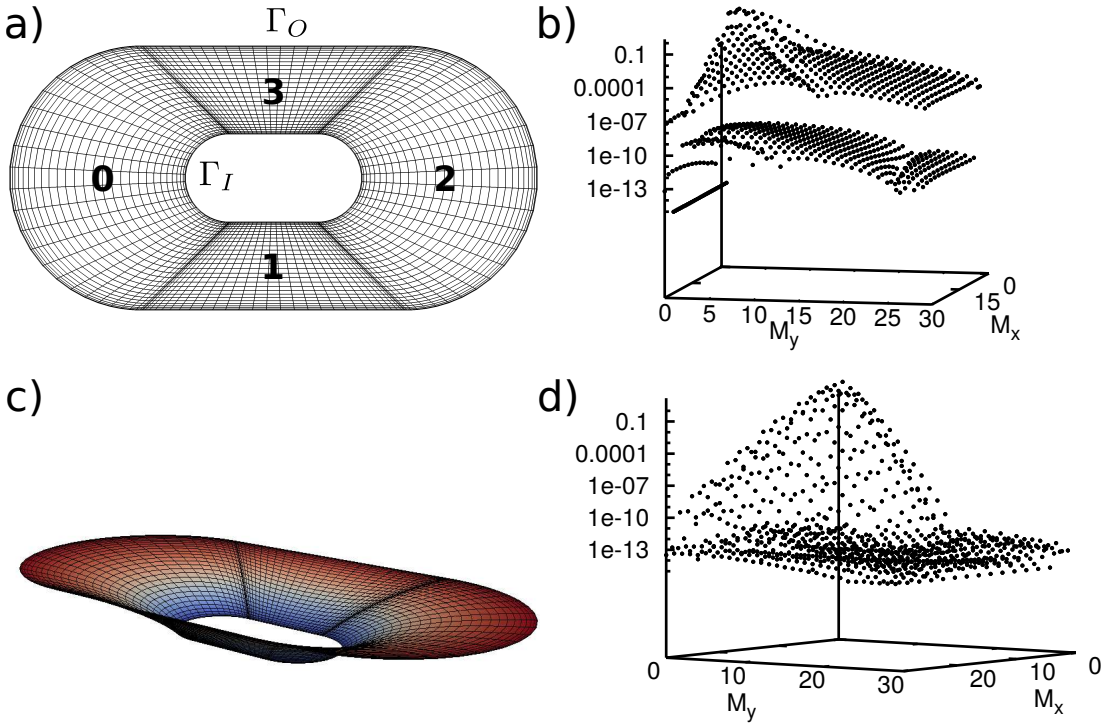


Figure 2.25: a) Domain consisting of straight-sided and circularly deformed elements. b) Coefficient spectra of the solution to problem (2.132) on element no. 0. c) Visualisation of the solution to problem (2.132). d) Coefficients of solution to problem (2.133) on element no. 0.

Insufficient number of points describing curved boundary

If the computational domain has a curved boundary $\partial\Omega_c$, Q quadrature points lying on $\partial\Omega_c$ has to be specified to define the curve. Here meet the problems of accurate approximation of the boundary curve and sufficient quadrature for evaluation of the integrand over the deformed domain/element. But a shape of the domain and the solved problems are two independent aspects of the computation.

As already noted in sec. 2.2.4, we use an isoparametric description of the deformed domains, so the boundary is approximated as same as a function in $(D-1)$-

dimensional basis. Having two neighbouring elements on a smooth boundary, the approximation is not smooth in given (machine) precision, if the expansion coefficients of the curve approximations are not converged to this precision. A problem expected to be smooth in the continuous analysis then adopts limited regularity and the spectra of solution do not decay exponentially in whole range (see figure 2.26).

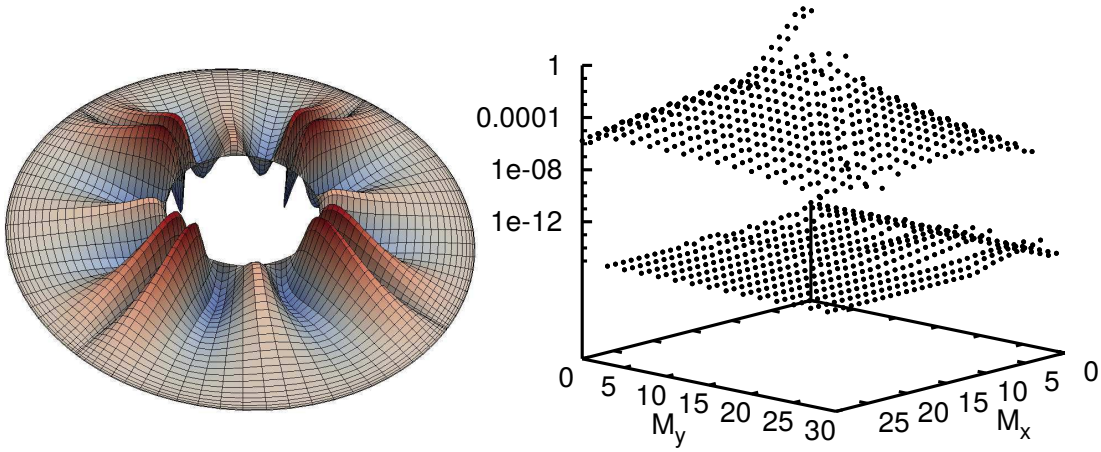


Figure 2.26: *left:* Difference $(u_e - u)$ between solution with exactly (u_e) and insufficiently (u) approximated curved boundary (circular domain with circular hole). Values multiplied by 10^4. *right:* Coefficients in element no.2, convergence stops at level $\sim 10^{-5}$, what corresponds with amplitude of the oscillations of error in the left figure.

The number of points Q_c needed for accurate approximation of a curve may differ from the number of points adjusted for polynomial order of the basis Q_b. If the $Q_b > Q_c$, interpolation polynomial of order $Q_c - 1$ is constructed and Q_b quadrature points are evaluated accurately and the curve is approximated to machine precision. However, if the polynomial order of the basis is set such, that $Q_b < Q_c$, the curve is not approximated with the same level of accuracy. This situation is improved by refinement of the boundary elements, but the convergence to the accurate approximation of the curve is of algebraic type and may result in exceeding increase of computational complexity.

The interpolation polynomials oscillate on equidistant grids, this is the reason, why the high order approximations to the curve should not be given by equidistant points. A specialised tool, programme "JP_XmlEdit", was written as a generator of curved (elliptical or arbitrary polynomial) boundaries approximated by points distributed as Guass-Legendre-Lobatto over the curve length.

Spectral length on deformed domains

Both the testing/projection integrals

$$\int_{\Omega_{std}} f(\mathbf{x}(\boldsymbol{\xi}))\, \phi_{mn}\, |J_e| \mathrm{d}\Omega \quad m, n = 1, \ldots, M\ ,\ e = 1, \ldots, E \tag{2.134}$$

and matrix elements (e.g. for Mass matrix $\int_{\Omega_{std}} \phi_{pq}\,\phi_{mn}\,|J_e|\mathrm{d}\Omega \quad p,q,m,n = 1,\ldots,N\,,\, e = 1,\ldots,E$) include the Jacobian in the integrand. Taking $f = \cos(2y)$, we compute coefficients of its projection (2.134) (left plots in fig. 2.27). In the projection integrals, Jacobian change order of the integrand in a priory unknown manner, so it should be checked, if the space is sufficient to keep the truncation error under the tolerance. If the coefficients are not converged for smooth function with exponential coefficient decay, value of the last mode refers to the truncation error (Theorem- 8). Then, concerning the simplest problem, forward transform of f

$$\mathbb{I}u = f\,, \tag{2.135}$$

the truncation error in f refers to the "testing error", as mentioned in sec. 2.2.1. The result is shown in the right-column plots in figure 2.27.

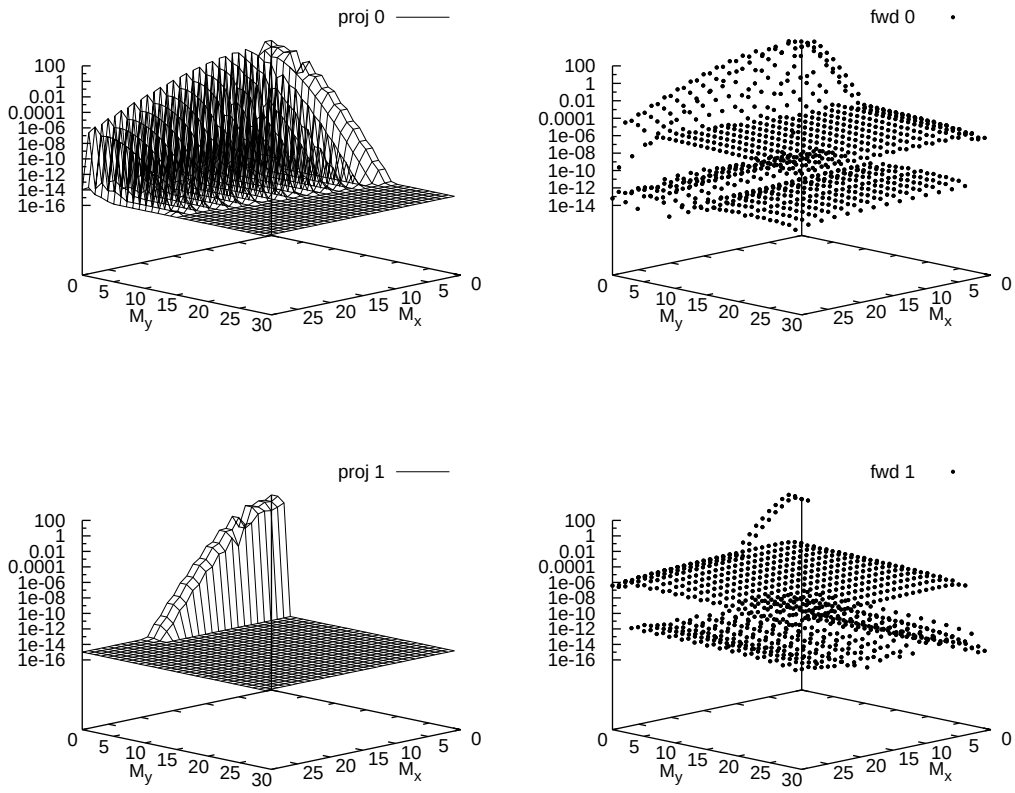


Figure 2.27: (Domain and mesh shown in fig. 2.25) Testing error (truncation error of the RHS), *proj*, sets the level, under which the coefficients of the solution can not decay. Coefficients of the result of the forward transform are shown in the right column (*fwd*), reflecting the fact, that the truncation/testing error spreads over all the coefficients, but also into neighbouring elements, since the boundary modes are shared.

It is not only the projection, what is influenced by the deformation. In elements of the system matrix, the non-constant Jacobian influences the orthogonality property of the basis, but also order of the integrand.

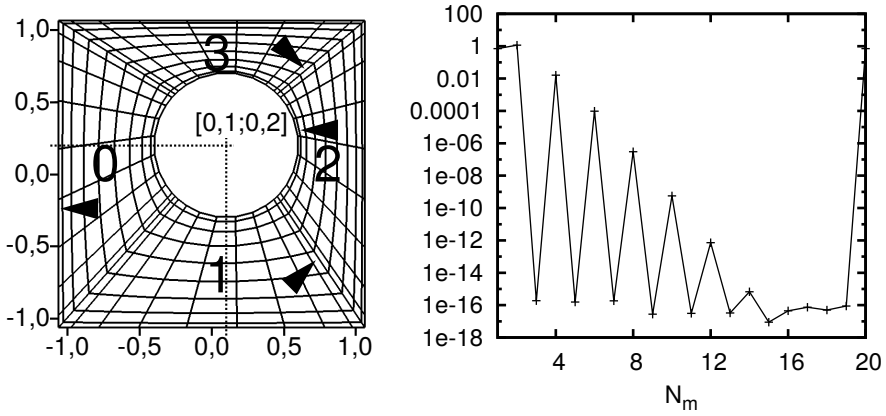


Figure 2.28: Square domain with eccentric circular hole. Mesh consists of 4 elements and the grid belongs to GLL(10). Local x-coordinate is pointed by the arrow. Estimate of N_m, resp. number of grid points $N_m + 1$, needed for accurate approximation of the boundary curve follows from 1D-Projection problem. $N_m = 14$ is needed for one quarter of circle shape in double precision.Therefore the curve in every element should be described at least by 15 GLL points.

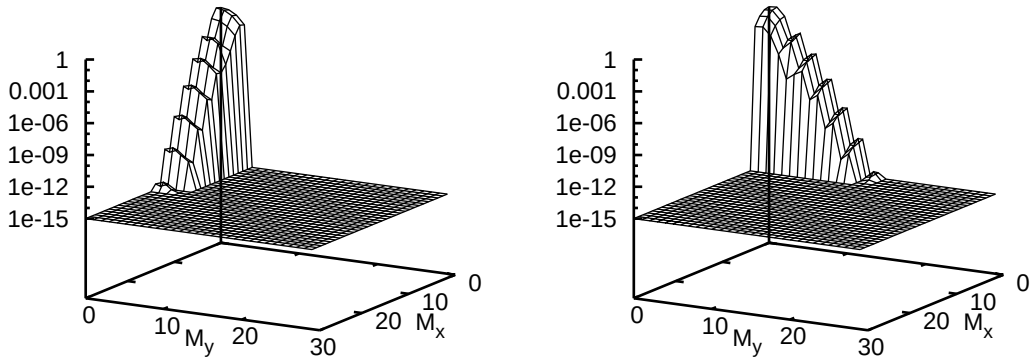


Figure 2.29: Projections $\int_{\Omega_{std}} f\,\phi_{pq}\,\mathrm{d}\Omega \quad p,q = 1,\ldots,30$ of function representing the Jacobian, $f = |J_e|$, belonging to elements 0 and 1 in mesh described in fig. 2.28. Curved boundary is in local-x direction for the element 0 and local-y for element 1. $N_m = 15$ is needed only for approximation of the Jacobian.

Basis functions, those mutually orthogonal in the standard element, loose some part of orthogonality in the deformed elements, as illustrated on comparison of fig. 2.30 and fig. 2.31. It is result of a non-constant Jacobian in the integral definitions of the matrix elements. E.g. the mass matrix elements in 2D are defined

$$\begin{aligned} &\int_{\Omega_e} \Phi_{pq}\,\Phi_{mn}\mathrm{d}\Omega = \int_{-1}^{1}\int_{-1}^{1} \phi_p(\xi_1)\phi_m(\xi_1)\,\phi_q(\xi_2)\phi_n(\xi_2)\,|J_e(\xi_1,\xi_2)|\mathrm{d}\xi_1\mathrm{d}\xi_2 \\ &\quad \forall p,q,m,n = 1,\ldots,N_m\,,\; e = 1,\ldots,E. \end{aligned} \tag{2.136}$$

Increasing complexity of the function describing the Jacobian increases polyno-

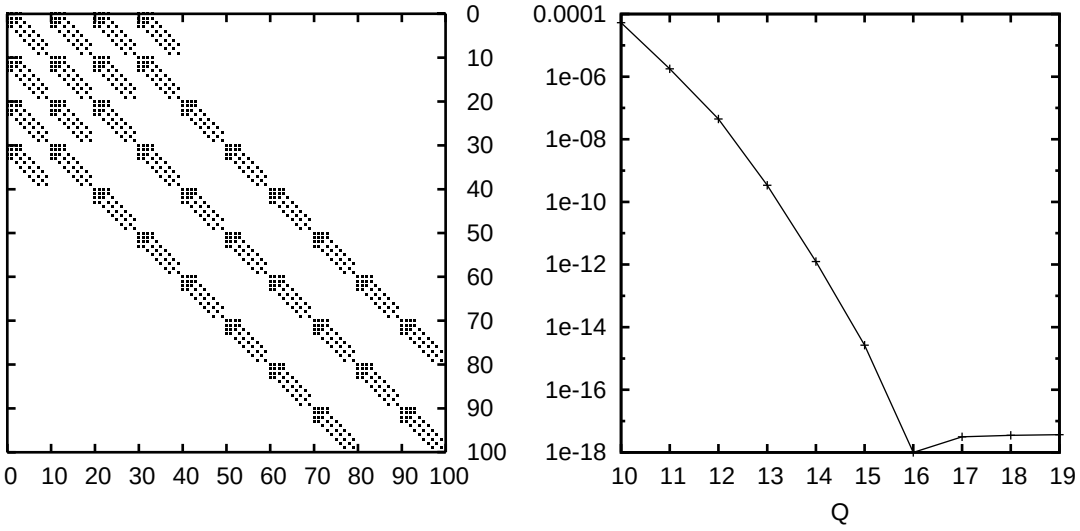


Figure 2.30: **left:** Structure of the mass matrix generated on 2D "Lobatto" basis and $M = 10$ in standard quadrilateral element. **right:** Convergence with increasing number of points Q, defining the curve. Values represent the difference from the most accurate value of the last Mass matrix element ($M_{100\,100}$) in case of Jacobian as described in fig.2.28 and 2.29.

mial order of integrand in (2.136). However, if the quadrature is sufficient only for the undeformed domain, the inaccuracy arising from the Jacobian occur still only in a limited number of the matrix elements. It is strongest in matrix element referring the highest-order modes and decreases for lower orders in the integrand, resp. for lower indices in the matrix structure (single-element matrices are discussed, not the matrices of the full multi-element system).

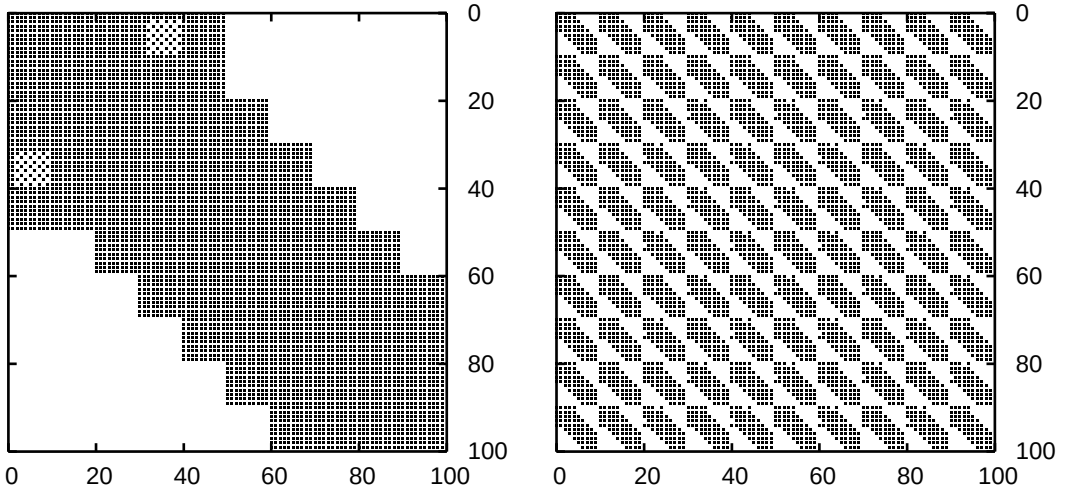


Figure 2.31: Structures of the mass matrices belonging to 2D "Lobatto" basis with $M = 10$ in both spatial directions as influenced by Jacobians shown in fig. 2.29. Matrix elements $> 1e-15$ are represented by black dots. Left diagram belongs to element 0, the right to element 1 in mesh from fig. 2.28.

3. Flow around a (heated) cylinder: numerical results

Our aim is to perform direct numerical simulation of non-stationary flow around heated cylinder. The scheme, introduced in previous chapter, is formulated in primitive variables and discretizes the equations presented in Chapter 1. The scheme allows to use a high order approximation in space and we will investigate this non-standard approach in our computations. No parameters, e.g. that suggested in model of outflow boundary condition in sec. 2.1.2, are used to fit the computational results to suit better the known results from the physical experiment.

Computations of both the isothermal and temperature dependent flows will be shown on the problem of flow around a (heated) cylinder. Flow around a cylinder is one of the classical problems, used as a test case in development of numerical methods, since it is well defined and suited for comparison. We will begin with calculation of the isothermal flow to validate quality of our scheme. The new results concern the Strouhal number obtained from the numerical simulation and its dependence on temperature and the Reynolds number. The results are compared with experimental data (Wang [43]) and (Vít [40]). Second benefit stays in investigation of the angle of flow separation, which is compared with the result (Wu [46]) and gives a preliminary estimate of its temperature dependence, for which an experimental data are not available up to now.

All the results presented in this chapter were based on the Nektar++ library [7] (version 3.3), which is an open-source program package allowing computations by the spectral element method, resp. with an unlimited approximation order in the finite element concept. The Nektar++ library was fundamentally modified on several parts.

3.1 Used software and software packages

The spectral methods and spectral element methods do not get such attention as lower order methods, especially in engineering and up to the authors knowledge, there do not exist commercial software based on these methods. Software used in this work is freely available, open source or newly written. The only commercial software (Comsol, Fluent), was used only for comparison.

Figure 3.1 presents the computational software and its place in the computational process.

Spectral method, especially in higher dimension, can't avoid breaking the computational domain Ω into elements Ω_e ($\Omega = \cup_e \Omega_e$) if a hole is present in the computational domain. In our study the circular hole is representation of the cylinder. This results in need of some mesh generation software. Gmsh is a program capable of generating mesh of various shapes (triangular, quadrilateral,...) and up to three dimensions. It allows work in graphical mode or through script files and beside meshing it can be used for visualization of the results. All the meshes, used for computations in this work, were prepared in Gmsh. An extra advantage is Gmsh's export format, which suits also as input for Fluent.

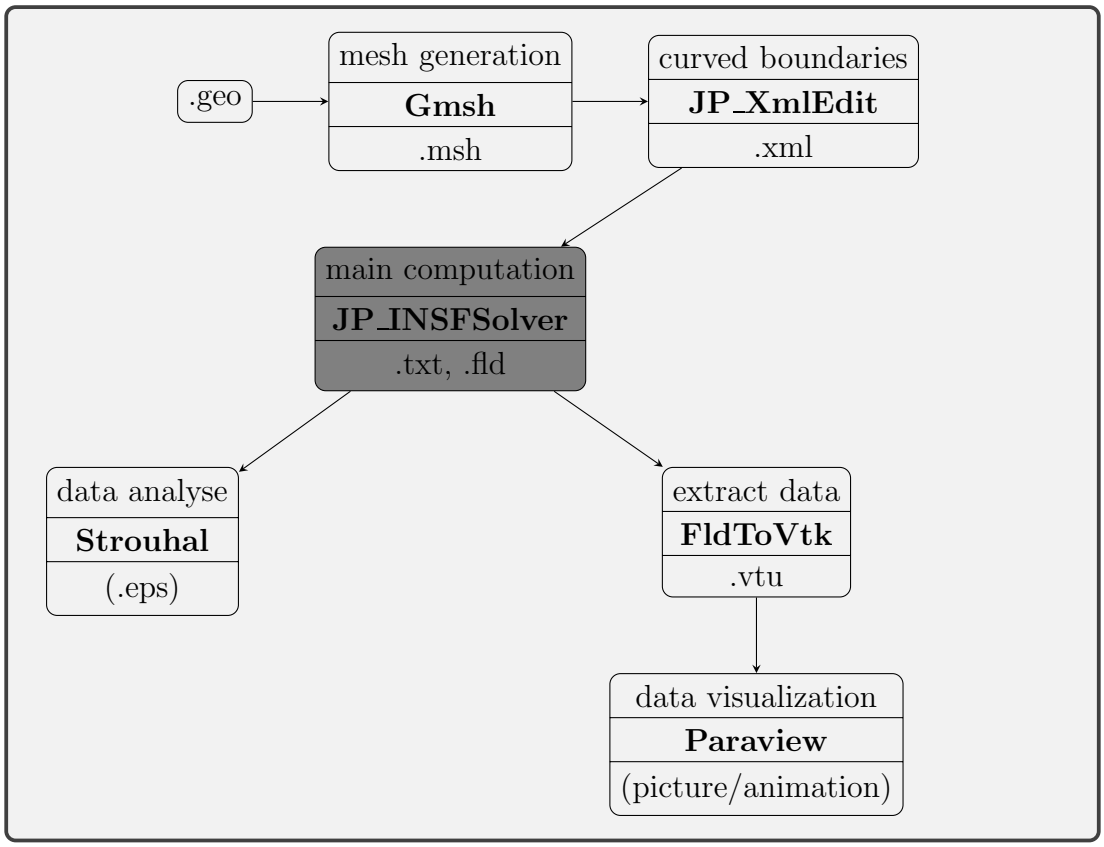


Figure 3.1: Sceme of the computational process. The blocks contain basic feature of the programme, its name and type of output.

Due to large dimensions of elements in spectral element meshes, the mesh must be deformed such that the elements well approximate curved boundaries of the computational domain. For more, the points describing the curve should have the same distribution as the Gauss-type points needed for numerical integration, otherwise interpolation is performed inside the Nektar++, what may cause dramatically different shape of the boundary curve than expected. Output from the Gmsh contains only the vertex points of the mesh. Therefore a programme for generation of Gauss points distribution lying on elliptical or arbitrary polynomial curve was developed. This program performs following process to generate the curve-describing points

- Calculate whole length of the curve using various number of quadrature points to ensure convergence and choose the sufficient number for accurate evaluation.
- Perform cycle starting at one of the boundary points to find coordinates of the internal points such, that the lengths of the curve between the points refer to distribution of chosen points type (GLL, equidistant,...).
- Last point calculated should coincide with the endpoint, vertex of the element. Tolerance for this check was set to $1e-14$ in the double precision.

Nektar++ is the heart of computations presented in this work. In the code is implemented both the C^0 continuous form of spectral elements and the Discontinuous Galerkin method. Theory and structure of the code follows (Karniadakis [25]). The C++ language and the object structures simplify orientation in the code especially with use of automatically generated documentation by program Doxygen. The package includes various utility programs for pre- and post-processing. Also some complete solvers[1] are delivered with the package. It was the Incompressible Navier-Stokes solver, which was enhanced for computations of temperature dependent flows with variable coefficients in this work. The fact, that the developed code is capable of parallel computations, should be emphasized. During work on this thesis, new releases of the Nektar++ package emerged, what confirms fast growth in interest to the high-order methods in last years.

There are two outputs from the newly developed solver program

- .txt files containing instant values of the drag and lift coefficients, and values of recognized separation angles.
- packed data of field values (.fld file).

Programs performing transform of .fld file to visualising programs as Gmsh (FldToGmsh), Paraview (FldToVtk) or Tecplot (FldToTecplot) are provided in the Nektar++ package. The utility FldToVtk was enhanced by optional adding data field of the divergence, vorticity, magnitude, etc. but also output of the expansion coefficient[2], which is an important tool in analysis for the spectral methods.

Program *Strouhal* was developed for analysis of the Strouhal number (St, c.f. table 1.1) from the periodical evolution of the lift or drag coefficients. It automatically recognizes the part of data, which belongs to fully developed flow field and calculates the final value from this part of the data, while providing the average value completed by the value of dispersion (c.f. sec. 3.3.1).

Paraview and gnuplot represent the last step of data visualisation. Both freely available software (http://www.paraview.org/, http://www.gnuplot.info/).

3.2 Aspects of the computations

Computational domain and mesh

We construct numerical model of a flow in wind tunnel and water tank. Setting of the physical experiments minimizes influence of the tunnel/tank walls, so that the measuring area is far enough from them. Due to the lack of knowledge of the inflow profile and need of reasonable computational efficiency we constraint

[1] Advection-diffusion solver, Acoustic Perturbation Equation Solver, Cardiac Electrohysiology Solver, Compressible Flow Solver, Incompressible Navier-Stokes Solver, Pulse Wave Solver, Shallow Water Solver

[2] The element-wise output creates a file of ordered coefficient values, c.f. figures in sec. 2.2.5.

the dimensions of the computational domain and try to prescribe boundary conditions, which resemble the farfield flow without influence of the object inside. These boundary conditions are necessarily artificial, since a disturbance caused by the object in all the velocity and pressure field decays with the distance so slowly, that in the machine precision level it can be observed so far, that the computational demands would be too large. The boundary conditions prescribe the balanced flow and the distance of the cylinder to the boundary should be corresponding. Unfortunately, the larger the domain is, the weaker singularity is present thanks to finite precision, but worser the spatial approximation is. Oppositely, the closer the outer domain boundary is to the object in the flow, the better spatial approximation is obtained, but stronger singularity emerge in the corners.

The influence of the boundaries is not known a priori and has to be tested. (Williamson [44]) presents a visualization of wake vortex structures existing in distance $350L$ downstream behind the cylinder (L is diameter of the cylinder, which coincides with the characteristic length in definition of Re). Most attention is usually paid to the outflow boundary, but thanks to the spectral methods we recognize, that source of singularities in the approximate mathematical solution occurs also at connection of inflow and side boundary conditions. However, this is not problem of the computational method, since this feature follows from properties of solutions to the differential equations.

Mesh

Decomposition of the *computational domain* Ω to sub-domains Ω_e

$$\Omega = \bigcup_{e=1}^{N_e} \Omega_e$$

is an artificial operation required by some of the numerical methods (e.g. FEM, FV) and its origin is not in the formulation of solved equations. But particular decomposition may strongly affect the result of the computation. More precisely, existence of the inter-elemental borders makes the solution function to be function with only piecewise features, in contrast to global expectations from mathematical analysis of the original equations.

An important part of the solutions construction is then design of the mesh in low order approximations. However, particular mesh is rather result of experience of a designer. As a quality test of the mesh is used its refinement, but in view of the algebraic convergence in the h-convergence process (fig. 2.10) it indicate rather the exhaustion of the method, than vicinity to the mathematical solution. More advanced approach seems to be automatic/dynamic construction of the mesh, as a defined reaction to a solution property. This approach is implemented in some computational packages (Hermes [19]).
We will present results mostly from computations on two meshes. The first, with only necessary elemental decomposition to capture the geometry of the problem, while keeping very high polynomial space on every element (fig. 3.2). The unit cylinder diameter in the mesh was chosen for simplicity and the spatial dimensions of the computational domain were: 20 units upstream, 60 downstream and 20

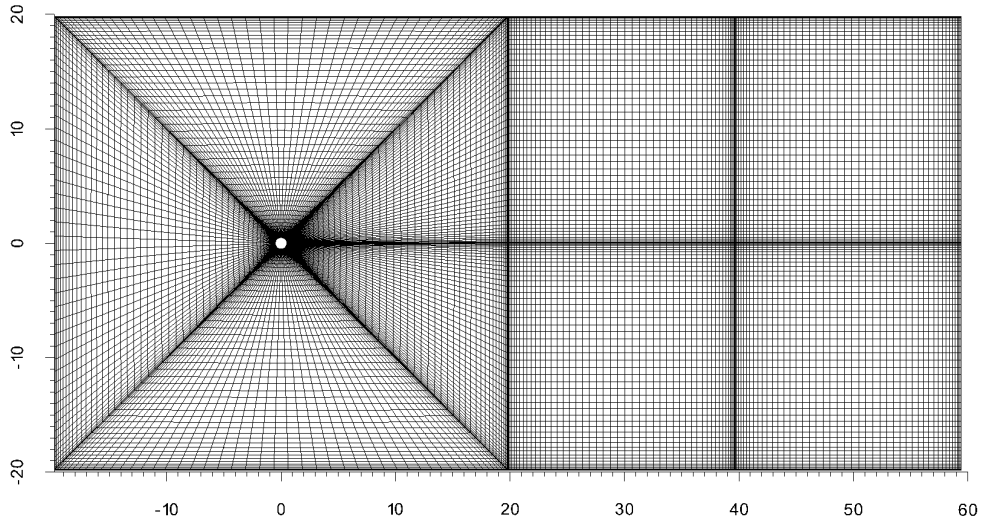


Figure 3.2: Mesh with grid points. Used for computations of flow around heated cylinder. Grid points are connected by lines. Grid refers to polynomial order 49. Grid points are concentrated toward the element boundaries, since the Gauss quadrature is employed.

above and under the cylinder. We divided the computational domain to small number of elements ($E = 9$) and used the rich expansion basis, having polynomial orders up to $M = 49$ in each coordinate variable (2500 degrees of freedom per single element in the algebraic system).

The second mesh is less extravagant, but still consisting of relatively large (triangular) elements (fig. 3.3).

Spectral convergences in flow computations

Examples of convergence of spectral coefficients were shown in previous chapter already. We remind, that values in the spectra indicate the quality of function approximation. If the coefficients of the approximated function are not fully converged, we can observe oscillations in adequately enlarged visualisation. In case of full convergence, no spurious oscillations are present until resolution, which coincide with the machine precision.

If the expansion is rich enough, we can observe asymptotic behaviour of the coefficients, which reflects regularity of the solution as mentioned in table 2.5. If the exponential decay is not observed, the truncation error or singularity may be expected in the solution. Enhancement of dimension of the expansion space improves the former case and as suggested in previous chapter, increase in polynomial order becomes stronger in high accuracies, than sub-dividing the domain to elements.

The computational algorithm for the Navier-Stokes-Energy system (sec. 2.1.1) is a subject of solution to Poisson and Helmholtz equations in the spatial solver at each time level. Solutions to these equations are strongly dependent on con-

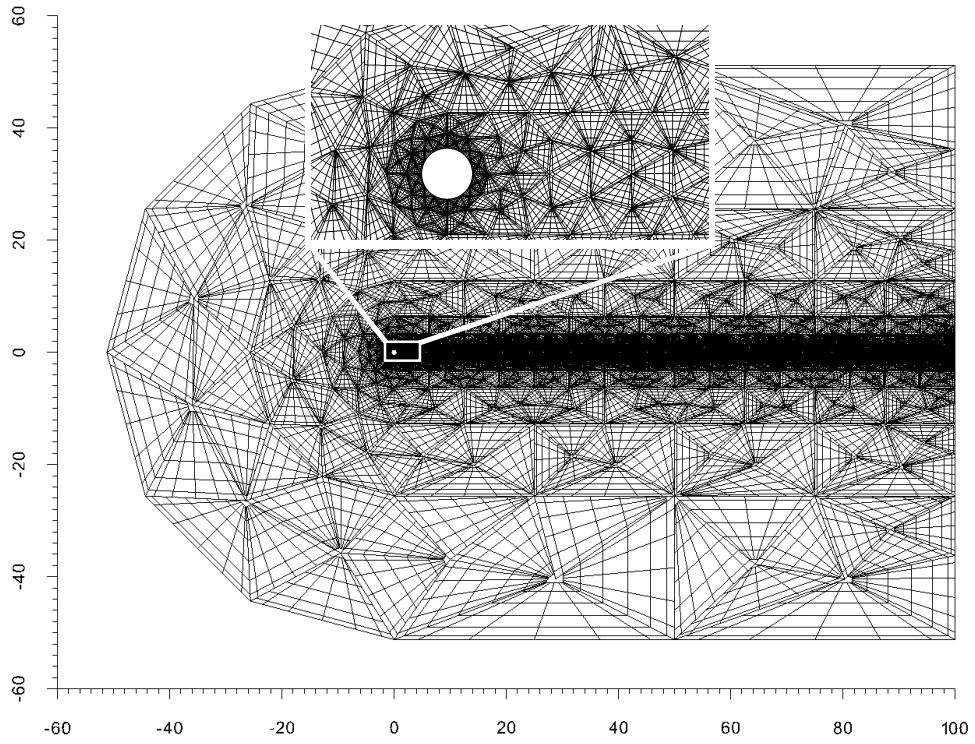


Figure 3.3: Triangular mesh with detail in the cylinder area and grid for basis of polynomial order 6. Dimensions of the whole domain are 100 downstream, 50 upstream and 100 in cross flow direction, while the cylinder diameter is 1.

sistence of BC and RHS as illustrated in sec. 2.1.2. The inconsistence of BC and RHS in these steps results in "boundary layers" and singularities in the spatial solution. The "boundary layer" in this sense is result of inaccuracy of the time discretisation and do not coincide with physical reality of the model. In view of the spectral decays, the boundary layer arising in the mentioned algorithm is computationally extremely expensive, as illustrated on the 1D example (figure 2.19). For more, this boundary layer introduces singularity into the solution if the boundary of the domain is not smooth. Spurious oscillations in the solution of flow around cylinder on the outer boundary of the domain, where constant Dirichlet conditions were prescribed is shown in figure 3.4.

Decay in spectra is achievable in solutions to the Navier-Stokes system as illustrated in figure 3.5. However, the computational demands are very high. The computational domain is restricted and the vicinity of the artificial outer boundary conditions to the cylinder makes the solved problem physically less realistic. Note, that coefficients of the spectra are not fully converged and in some of the elements the convergence is very slow.

3.3 Results for flow around cylinder

3.3.1 Strouhal number analysis

In this paragraph, we will describe evaluation of the forces acting to the body as done in our code and the tool "Strouhal", written for analysis of obtained data.

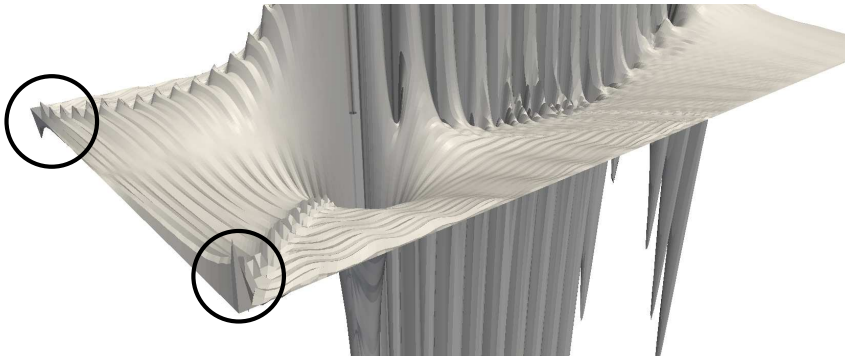


Figure 3.4: Oscillations in "y" component of the velocity, solution to the (isothermal) flow around cylinder, when constant Dirichlet conditions are prescribed on two boundaries, which intersect. Note, that the values are multiplied by 1000, so the precision of the computation is usually sufficient.

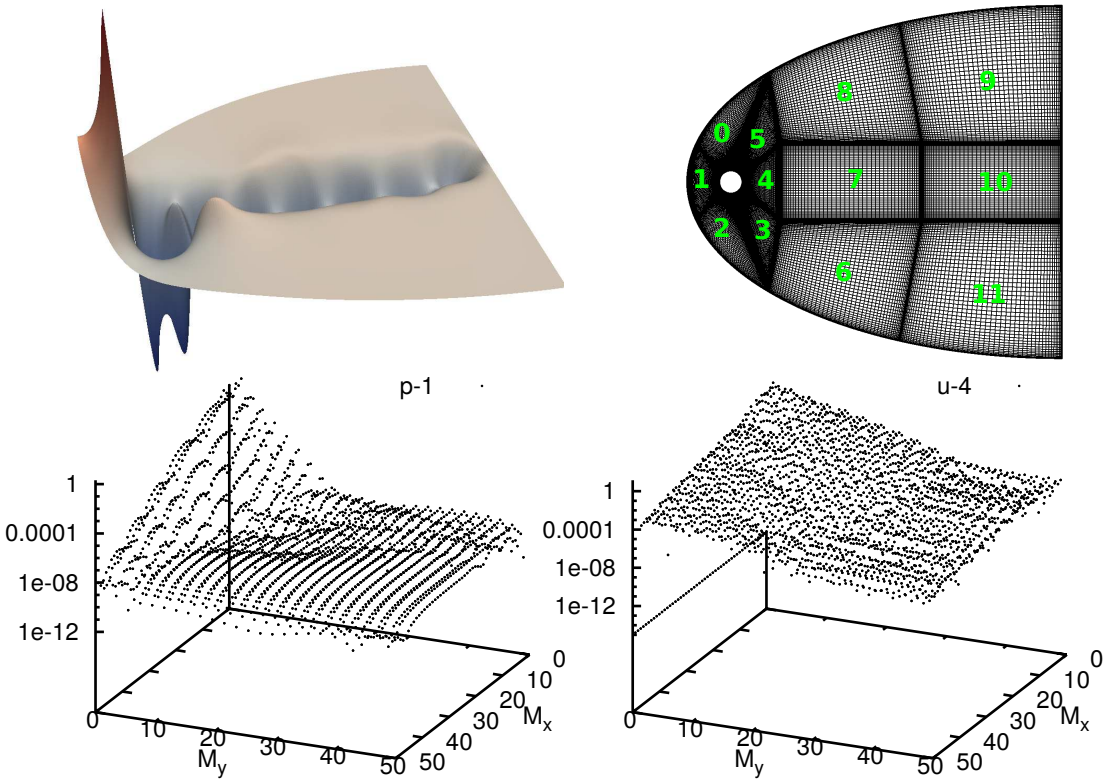


Figure 3.5: *top-left*: Pressure field in isothermal flow around cylinder in domain with half-ellipse shape, values in z-direction multiplied 10x. *top-right*:Computational mesh with highlighted numbering of elements. The cylinder with an unit diameter is in focus of an ellipse with major semi-axis length 17, minor 8 and eccentricity 15. The upstream length is the only 2. *bottom*:Coefficient spectra of pressure in element no. 1 (*left*) and x-component of velocity in element no.4 (*right*).

Data for this analysis will be taken from models of laminar vortex shedding in flow around a cylinder, where von Kármán vortex street occurs. The Strouhal

number is dimensionless expression related to the (characteristic) frequency of vortex shedding f in this case. The cylinder diameter was chosen to be the characteristic length L and the reference flow velocity $|\mathbf{v}_\infty|$ is given by inflow boundary condition. Characteristic frequency f may be determined from oscillations of drag F_D or lift F_L force, defined as vector components of total surface force $\mathbf{F}$ acting on the cylinder. Supposing that the main stream has direction of x-coordinate we arrive to (Schafer [37])

$$\mathbf{F} = \begin{pmatrix} F_D \\ F_L \end{pmatrix} = \begin{pmatrix} \int_C \left(\nu \frac{\partial v_t}{\partial n} n_y - p\, n_x \right) \mathrm{d}S \\ \\ -\int_C \left(\nu \frac{\partial v_t}{\partial n} n_x + p\, n_y \right) \mathrm{d}S \end{pmatrix} \tag{3.1}$$

where $\mathbf{n} = (n_x, n_y)^{\mathrm{T}}$ is a normal vector to the cylinder surface C, pointing into the domain Ω. Tangential vector to the cylinder surface is denoted by $\mathbf{t} = (n_y, -n_x)$ and respective tangential velocity is $v_t = \mathbf{v} \cdot \mathbf{t}$. The term $\dfrac{\partial v_t}{\partial n}$ then refers to the shear stress.
The components of the force may be seen also directly in terms of the stress vector $\mathbb{T}\,\mathbf{n}$

$$\mathbf{F} = \int_C (\mathbb{T}\,\mathbf{n}) \mathrm{d}S = \begin{pmatrix} \int_C (2\nu \frac{\partial v_1}{\partial x} - p) n_x + \nu (\frac{\partial v_1}{\partial y} + \frac{\partial v_2}{\partial x}) n_y \mathrm{d}S \\ \\ \int_C (2\nu \frac{\partial v_2}{\partial y} - p) n_y + \nu (\frac{\partial v_1}{\partial y} + \frac{\partial v_2}{\partial x}) n_x \mathrm{d}S \end{pmatrix} \tag{3.2}$$

Both previous representations are corresponding and may be also computed using volume integration. This approach is recommended in (John [20]) for reason of accuracy and lower sensitivity of approximation of the cylinders shape[3].

Drag C_D and lift C_L coefficients are then normalized values of F_D, F_L

$$C_D = \frac{2F_D}{|\mathbf{v}_\infty|^2 L}\ ,\ C_L = \frac{2F_L}{|\mathbf{v}_\infty|^2 L} \tag{3.4}$$

Output of these values is done in every time step of the computation and results in a series of temporal data. The timestep Δt used in the computational scheme is short enough to resolve the frequency with enough accuracy.
All the formulas are valid also in the case of temperature dependent flow, while simply substituting the temperature dependent viscosity into definition of the stress tensor. Formula (3.2) is implemented in our code.

Figures 3.6, 3.8 contain an example data of drag and lift forces, output from the main code for isothermal flow on the triangular mesh and Re = 120.

[3]Surface forces given by volumetric integrals are:

$$\mathbf{F} = \begin{pmatrix} \int_\Omega [\nu \nabla \mathbf{v} : \nabla \mathbf{v}_d + (\mathbf{v} \cdot \nabla)\mathbf{v} \cdot \mathbf{v}_d - p(\nabla \cdot \mathbf{v}_d)] \mathrm{d}x\mathrm{d}y \\ \\ \int_\Omega [\nu \nabla \mathbf{v} : \nabla \mathbf{v}_l + (\mathbf{v} \cdot \nabla)\mathbf{v} \cdot \mathbf{v}_l - p(\nabla \cdot \mathbf{v}_l)] \mathrm{d}x\mathrm{d}y \end{pmatrix} \tag{3.3}$$

where $\mathbf{v}_l, \mathbf{v}_d \in (H^1(\Omega))^2$ with $(\mathbf{v}_d)|_S = (1,0)^T$ and $(\mathbf{v}_l)|_S = (0,1)^T$ vanish on $\partial\Omega \setminus S$.

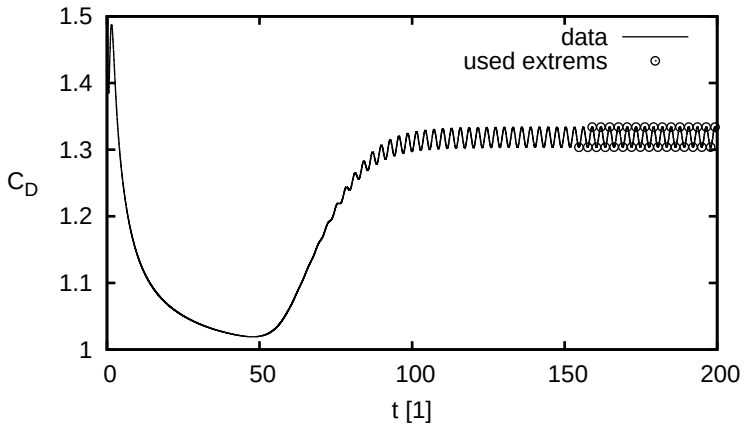


Figure 3.6: Time evolution of the drag coefficient in flow around cylinder. Variable t denotes the dimensionless time, (1.29).

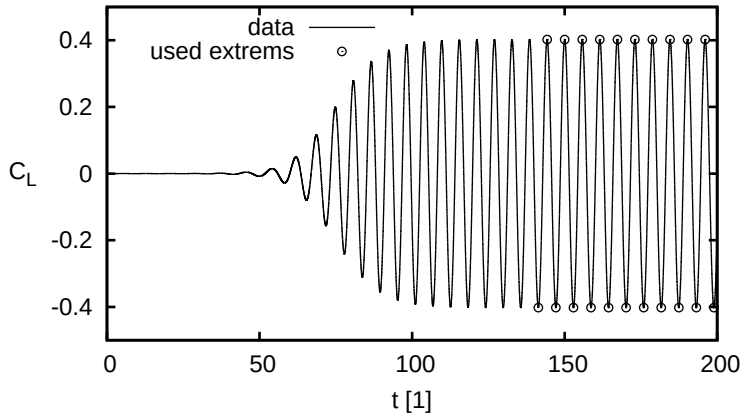


Figure 3.7: Time evolution of the lift coefficient in flow around cylinder.

Two methods of obtaining the frequency from the data, the Fourier analysis and an direct detection of periods between the extremes, were compared. The programme "Strouhal" was written for the latter method, algorithm of the Fourier analysis was found as less accurate[4]

The programme "Strouhal" firstly reads all the data and detects a local extremes. Temporal distances between relevant extremes define periods in the oscil-

[4]Data after manual cut of the initial disturbances may be analysed through following MATLAB (Octave) script using Fourier method

```
L=length(input);
NFFT=2^(nextpow2(L)+6);
Y=fft(input(:,2),NFFT)/L;
f=1/((input(2,1)-input(1,1))*2)*linspace(0,1,NFFT/2);
fce=2*abs(Y(1:NFFT/2));
'['C,I']'=max(fce);
plot(f,fce)
title(['Maximum:  '  num2str(f(I)) ',' num2str(C) ' = > St= ' num2str(f(I)) '
df=' num2str((f(I+1)-f(I-1))/2)]);
axis ([0 0.5]);
```

In the previous script the *input* are data in columns, *NFFT* is number of points (frequencies) in the transformed space (higher number of points results in shorter frequency step).

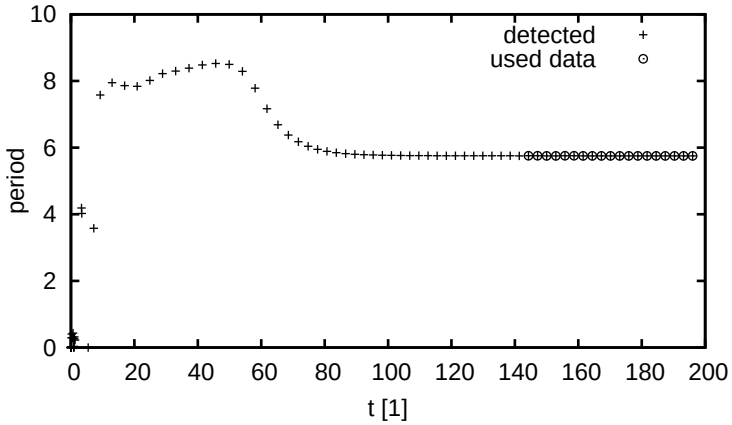


Figure 3.8: Time evolution of the time periods between local extremes in the lift coefficient data for Re = 120 and the triangular mesh. *detected* denotes the detected periods and *used data* denote the periods used in St computation. The values stabilizes with increasing time.

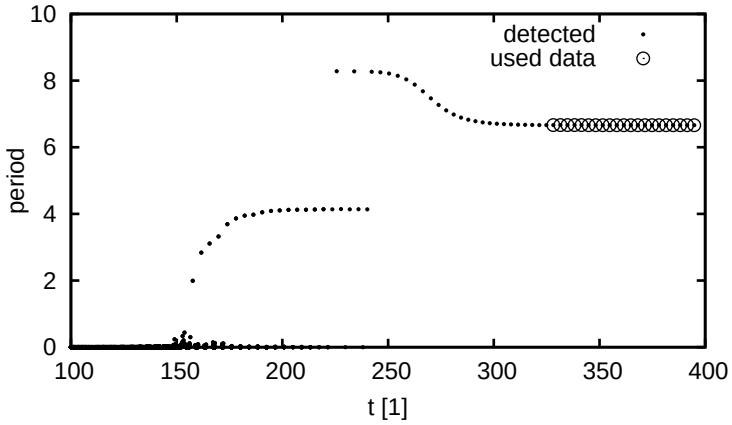


Figure 3.9: Time evolution of the oscillation periods for $Re = 73.8$ in case of isothermal incompressible flow as calculated on the "9-element" mesh. Notation is the same as in Figure 3.8. In comparison to the triangular mesh, the oscillations at frequency of the vortex shedding begin at a higher time. Difference is also in the jump onset of this frequency, which we observed at lower Re and only on the "9-element" mesh.

	velocity		pressure	
Boundary	BC-type	value	BC-type	value
Inlet	Dirichlet	$\mathbf{v} = (1, 0)$	Neumann	$\partial p/\partial \mathbf{n} = 0$
Sides	Dirichlet	$\mathbf{v} = (1, 0)$	Neumann	$\partial p/\partial \mathbf{n} = 0$
Outflow	Neumann	$\partial \mathbf{v}/\partial \mathbf{n} = (0, 0)$	Dirichlet	$p = 0$
Cylinder	Dirichlet	$\mathbf{v} = (0, 0)$	Neumann	HOPBC

Table 3.1: Boundary conditions for velocity and pressure. HOPBC refers to the *High Order Pressure BC* defined in (2.43). Inlet and sides forms a single boundary in case of the triangular mesh.

lating data, see figure 3.8. Having temporal series of periods $\{p_i\}$, the algorithm goes through it oppositely, from the largest time, and comparing every change of dispersion

$$\sigma^2 = \sum_{i=1}^{N} \frac{(\bar{p} - p_i)^2}{N},$$

when new period to computed average is added, it searches, where to stop reading the data.

This process is done in cycles, taking a tolerance of the dispersion larger and larger, until the number of involved periods reach or exceed the number of vortices (set by user) in the wake behind the cylinder in the computational domain. This is done to detect possibly not fully developed wake.

Periods in the stabilised shedding regime are very accurately constant with precision much better than needed for the final comparison with data from physical experiment. In contrast to the Fourier analysis, this result is independent on particular cut-off of the initial data.

Having the final average value of period length and its dispersion, final frequency is just its reciprocal value $f = 1/\bar{p}$.

3.3.2 Critical Reynolds number

Between laminar and turbulent flow regimes there exists a range, where parallel vortex shedding occurs in the flow around a cylinder. This range may be specified as

$$\mathrm{Re}_c \leq \mathrm{Re} \leq 180\,, \tag{3.5}$$

where Re_c is the *critical Reynolds number* separating the stationary (if $\mathrm{Re} \leq \mathrm{Re}_c$) and nonstationary ($\mathrm{Re} \geq \mathrm{Re}_c$) flow. Theoretical study on the value of the Re_c was presented in (Fedorchenko [10]), where value of $\mathrm{Re}_c = 47.5$ was derived.

We performed a set of computations to find Re_c through the flow simulation. The observed quantity was a time evolution of the lift (or drag) coefficient. The expectation was, that the initial disturbance, which arises from incompatibility of the initial condition, will be suppressed in time, if the flow regime is stable, resp. $\mathrm{Re} < \mathrm{Re}_c$. The above critical regime would be characterized by development in value of this coefficient and its oscillatory behaviour.

Prescribed boundary conditions are summarized in table 3.1.

The final results are presented in figures 3.10 and 3.11. The first figure shows strong dumping of the initial inaccuracy, since the flow regime was far under the Re_c. From the second figure it may be deduced, that the dumping weakens towards the Re_c and oppositely the disturbances are developed weakly, if Re is a little above the Re_c.

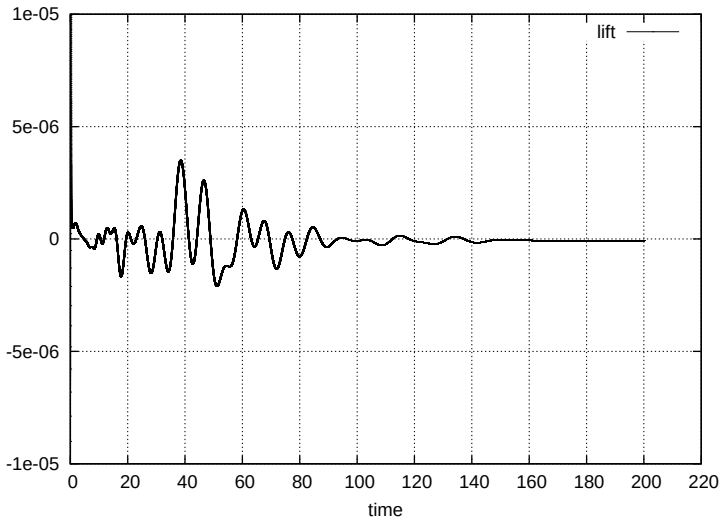


Figure 3.10: Dumping of initial disturbance if computation starts incompatible initial condition ($Re = 20$).

We conclude, that the Re_c found from these computations was in interval $\mathrm{Re}_c \in [46; 47]$, what is in good agreement with the expected value. The difference coincides with the later described observation in the calculated Strouhal number for higher Re, which is overvalued in comparison to the empirical results for low Re and converge to the value with increasing accuracy of the simulation.

It is noticeable also in the figures, that the lift coefficient develops periodically also in these low amplitudes, when the vortex shedding is not observed. The tool for analysis of the Strouhal number was used on data of $\mathrm{Re} = 47$, with the result $\mathrm{St} = 0.117$ being in good agreement with (3.7), which gives $\mathrm{St} = 0.1188$ for $\mathrm{Re} = 47.5$. This suggests, that the oscillatory behaviour with period resembling the period of vortex shedding is present in the flow earlier, than the vortex shedding is developed.

3.3.3 Strouhal-Reynolds relationship

The parallel vortex shedding occurs in the interval

$$\mathrm{Re}_c \leq \mathrm{Re} \leq 180 \tag{3.6}$$

in the flow around a cylinder. The Van Kármán vortex street develops in this regime, which is also denoted as *transition to turbulence*. The upper bound of mentioned interval is chosen to be safely under the value, above which oblique shedding occurs. Therefore for $\mathrm{Re} < 180$ a 2D computational model may be sufficient approximation.

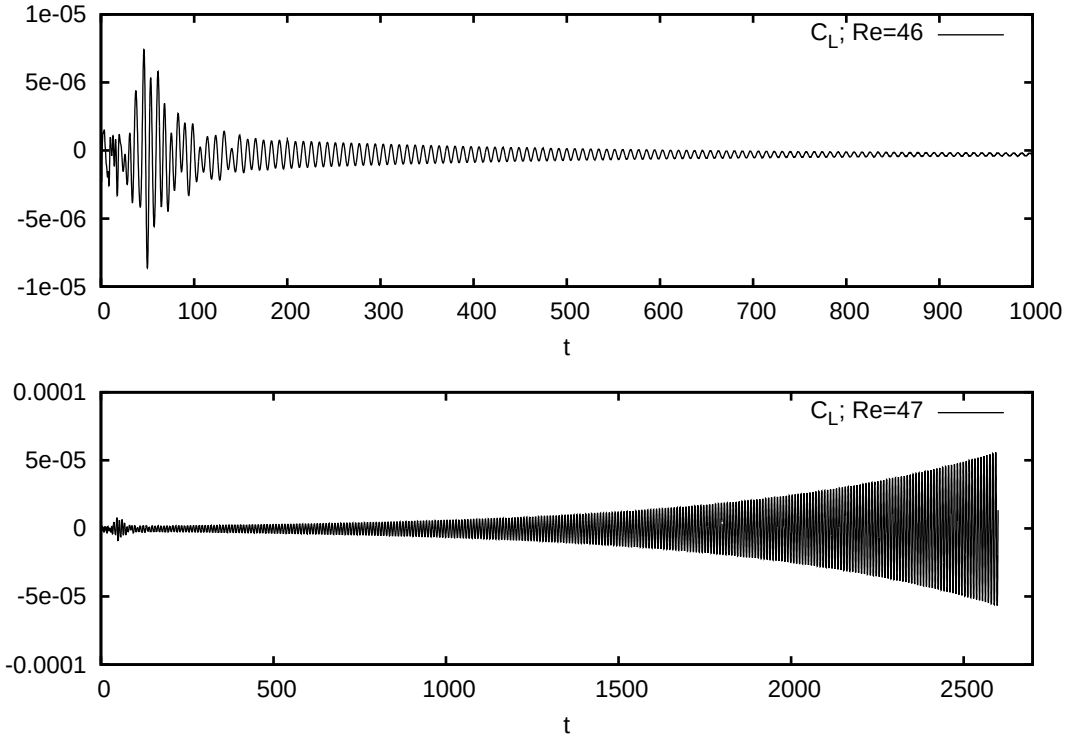


Figure 3.11: Evolution of lift coefficient in time for sub-critical Reynolds number $Re = 46$ (top) and above-critical Reynolds number $Re = 47$. An initial disturbance is suppressed in the former, while slowly (note the difference in time range) increases in the latter.

This flow regime was already extensively studied both experimentally and in simulations. From the experimental data there were derived various relations of Strouhal and Reynolds numbers. (Williamson [44]) provided a formula, an expansion in powers of $\mathrm{Re}^{1/2}$, concerning a wide range of Reynolds numbers ($50 \leq \mathrm{Re} \leq 1.4 \times 10^5$). This formula takes the form of

$$\mathrm{St}(\mathrm{Re}) = 0.2665 - \frac{1.0175}{\sqrt{\mathrm{Re}}} \tag{3.7}$$

in mentioned range of parallel shedding.

Since the flow develops slowly from the constant initial conditions, a long time computation exceeding 200000 time steps was needed (the time step was set always to $\Delta t = 0.001$). The computation was performed for a whole set of the Reynolds numbers to approximate the dependence St(Re) in whole range (3.6). Whole set of computations was performed on both the 9-element and triangular mesh. Computations on the 9-element mesh used one step IMEX scheme and those on the triangular mesh employed 3-step IMEX scheme, c.f. section 2.1.

Prescribed boundary conditions coincide with those in table 3.1.

The final values from computations are listed in table 3.2 and plotted in figure 3.12.

Re	St_e	St_t	St_s
60	0,1351	0,1356	0,1376
62,9	0,1382	0,1387	0,1406
73,8	0,1481		0,1501
80	0,1527	0,1527	0,1545
85,8	0,1567	0,1567	
90,4	0,1595		0,1613
100	0,1647		0,1663
101,4	0,1655	0,1655	0,1671
120	0,1736	0,1738	0,1749
123,2	0,1748	0,1750	0,1764
146,3	0,1824	0,1829	0,1838
160	0,1861	0,1868	0,1872
163,3	0,1869		0,1884
180	0,1907	0,1917	

Table 3.2: Calculated values of the Strouhal number for various Re. St_e is value from empirical formula (3.7), St_t values belonging to computation on the triangular mesh and St_s on the 9-element mesh. The average value of the relative difference to the empirical formula is 0.2% for the triangular mesh and 1% for the "9-element" mesh.

3.3.4 Strouhal-Reynolds-Prandtl relationship

There is no difference in flow characteristics among various materials in isothermal case if the same Reynolds number is prescribed. However, if the temperature change occurs in the flow, we can observe fundamental difference in the response.

In our studies, working fluids were water and air. As mentioned in sec.1.1.1, viscosity of air increases with temperature, while decreases for water. In a virtual situation, when the temperature of the fluid would change uniformly, the flow would behave as set to a different Reynolds number. Increase in viscosity causes decrease of the Reynolds number, therefore the heated air would behave as in flow of lower Re and uniformly heated water would behave as with higher Re. However, the heat exchange is given only through the contact with the cylinders wall, so the "change of the Reynolds number" is only local, what results in a more complex change of flow structures.

The temperature distribution itself influences the heat propagation through the fluid, since the thermal conductivity is also temperature dependent. Again, an uniform change in temperature causes a change in the Prandtl number. But in the case, when only viscosity would reflect the temperature change, the Prandtl number would not be correct up to the farfield values. Therefore we consider the temperature dependent thermal conductivity too.

The $St - Re$ relationship for the heated cylinder in flow of water and air was investigated experimentally in (Vít [40]) and a theoretical analysis was done in

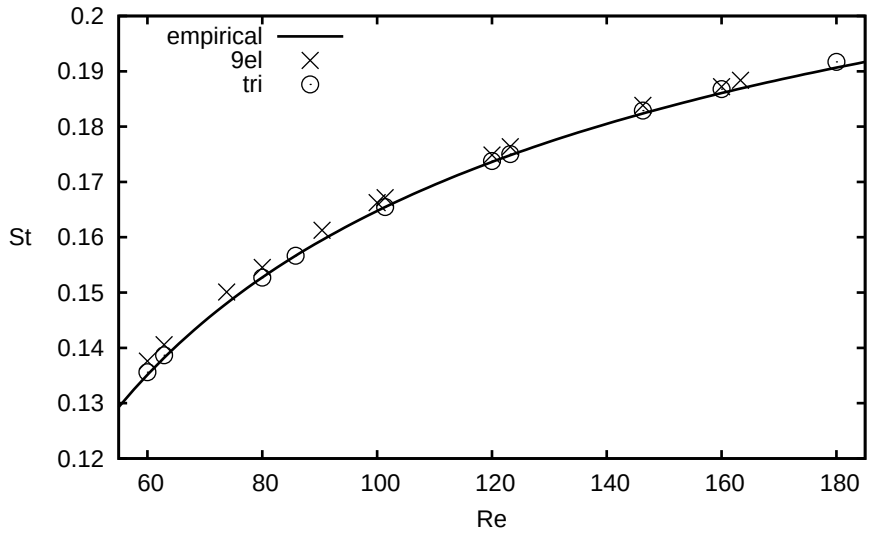


Figure 3.12: Strouhal number dependent on the Reynolds as a result of computations both on the 9-element mesh "9el" (fig. 3.2) and triangular mesh "tri" (fig. 3.3) with the empirical formula (3.7).

(Maršík [27]), resulting in the empirical formula

$$\mathrm{St}(\mathrm{Re},\mathrm{Pr},T^*) = 0.2665 - \frac{1.0175(T^*)^{\frac{\omega}{2}}}{\sqrt{\mathrm{Re}}}\left[1+\frac{0.227(1-T^*)}{\mathrm{Pr}^{1/3}(2T^*-1)}\right]^{\frac{1}{2}}, \qquad (3.8)$$

where ω is the exponent from the power-law approximation to the temperature dependence of dynamical viscosity (1.24). Above formula well approximates the experimental data and it was used for comparison with results from our computations.

Ratio of the inlet temperature T_∞ and the temperature of the cylinder wall T_c define a characteristic, dimensionless temperature T^*. Results for $T^* = 1.1$, $T^* = 1.5$ and $T^* = 1.8$ are available for flow of the air and $T^* = 1.0034$, $T^* = 1.0048$ and $T^* = 1.0096$ for flow of the water, all as a result from the experimental studies of (Vít [40]). A set of computations covering whole range of all the mentioned temperature ratios was performed for particular values of the Reynolds numbers for both the air and water.

Because the 9-element mesh represents non-standard computation, all the results were done primarily on the triangular mesh. Setting of the boundary and initial conditions for velocity $\mathbf{v}$ and pressure p was same as in the isothermal case, only the pressure-Neumann condition took the form with the variable viscosity (2.45) instead of (2.43).

The boundary conditions for temperature field are summarized in table 3.3.

Evaluation of the surface forces, resp. the drag and lift coefficients, followed (3.2) with temperature dependent viscosity. The data analysis was done using the programme "Strouhal" as in the previous sections.

The resulting data for both air and water are summarized in tables 3.4, 3.5, 3.6, 3.7 and graphs of $St-Re$ dependencies, fig. 3.13 and fig. 3.15. The continuous curve is given by the formula (3.8).

Boundary	BC type	value
Inlet	Dirichlet	$T = 1$
Sides	Dirichlet	$T = 1$
Outflow	Neumann	$\partial T/\partial \mathbf{n} = 0$
Cylinder	Dirichlet	$T = T^*$

Table 3.3: Settings of temperature boundary conditions. Note, that for the triangular mesh, inlet and sides form a single border of the computational domain.

Air $T^* = 1.1$				
Re	9-el.	tri.	emp.	exp.
55.2	0.13	0.1279	0.1259	0.1267
123.2	0.1748	0.1735	0.1724	0.1727
146	0.1824		0.18	0.1804
163.3	0.1871	0.1863	0.1848	0.1845

Table 3.4: Strouhal numbers from simulation on the "9-element" mesh, *9-el.*, "triangular" mesh, *tri.*, calculated from (3.8), *emp.*, and those obtained experimentally (Vít et al. [40]), *exp.* in case of the cylinders wall temperature $T_W = 1.1T_\infty$.

Air: $T^* = 1.5$				
Re	9-el.	tri.	emp.	exp.
62.9	0.1303	0.1282	0.1211	0.12
100.7	0.1586		0.1516	0.152
123.2	0.1691	0.1675	0.1626	0.1631
146.3	0.1775		0.1712	0.1724
163.8	0.1821	0.1812	0.1763	0.1776

Table 3.5: Strouhal numbers from simulation on the "9-element" mesh, *9-el.*, "triangular" mesh, *tri.*, calculated from (3.8), *emp.*, and those obtained experimentally (Vít et al. [40]), *exp.* in case of the cylinders wall temperature $T_W = 1.5T_\infty$.

3.3.5 Angle of separation

An other quantity, for which the experimental results were available in literature, is the separation angle on cylinder. As defined in [46], the separation point in the boundary layer is on the surface at the place, where the shear stress is zero. The separation angle is then the angle between the frontal stagnation point and the separation point. The shear stress on the cylinder surface is computed as the tangential velocity gradient in the radial direction

$$\frac{\partial \mathbf{v}}{\partial \mathbf{n}} \cdot \mathbf{t}\,. \tag{3.9}$$

Water: $T^* = 1.0034$				
Re	9-el.	tri.	emp.	exp.
54.5	0.1319	0.1298	0.1303	0.1313
70.4	0.1477	0.1458	0.1467	0.1481
90.4	0.1616	0.1598	0.1608	0.1618

Table 3.6: Strouhal numbers from simulation on the "9-element" mesh, *9-el.*, "triangular" mesh, *tri.*, calculated from (3.8), *emp.*, and those obtained experimentally (Vít et al. [40]), *exp.* in case of the cylinders wall temperature $T_W = 1.0034T_\infty$.

Water: $T^* = 1.0096$				
Re	9-el.	tri.	emp.	exp.
56.3	0.1348	0.1328	0.1354	0.1369
73.8	0.1511	0.1492	0.1520	0.1538
85.8	0.1595	0.1576	0.1603	0.1613
90.4	0.1622	0.1604	0.1631	–

Table 3.7: Strouhal numbers from simulation on the "9-element" mesh, *9-el.*, "triangular" mesh, *tri.*, calculated from (3.8), *emp.*, and those obtained experimentally (Vít et al. [40]), *exp.* in case of the cylinders wall temperature $T_W = 1.0096T_\infty$.

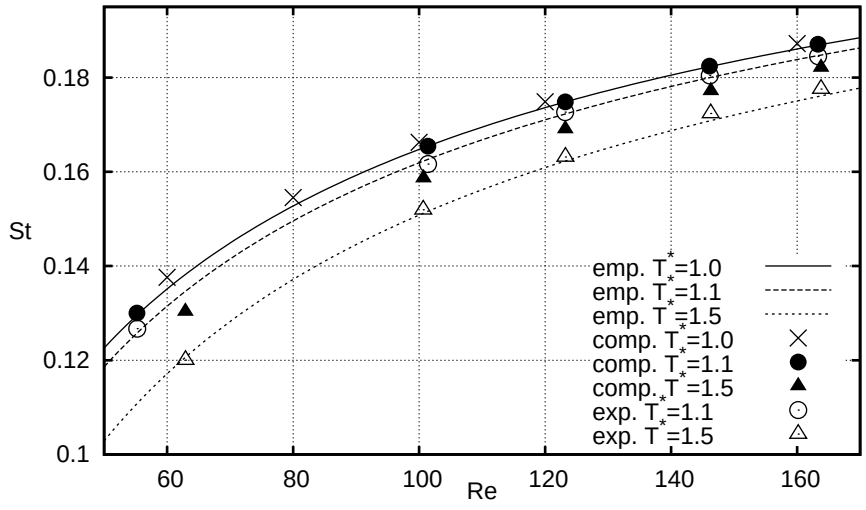


Figure 3.13: Strouhal-Reynolds number relation for flow of air at various ratios $T^* = T_W/T_\infty$ of the temperature of the cylinder wall T_W and inlet flow temperature T_∞. Data sets belong to (3.8), *emp.*, computation on the "9-element" mesh, *comp.* and experiment (Vít [40]), *exp.*

If the neighbouring discrete values, calculated in the quadrature points, have opposite signs, we use the linear interpolation to find the position of zero value. The accuracy of the position determination then varies through every element, since the distribution of the Gauss-Legendre-Lobatto quadrature points is not

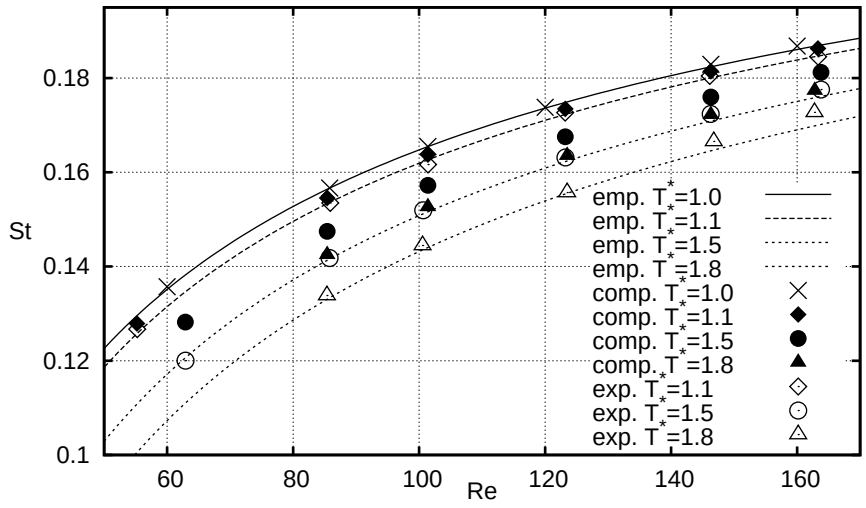


Figure 3.14: Strouhal-Reynolds number relation for various temperatures and various ratios $T^* = T_W/T_\infty$ of the temperature of the cylinder wall T_W and inlet flow temperature T_∞. Data sets belong to (3.8), *emp.*, computation on the "triangular" mesh, *comp.* and experiment (Vít [40]), *exp.*

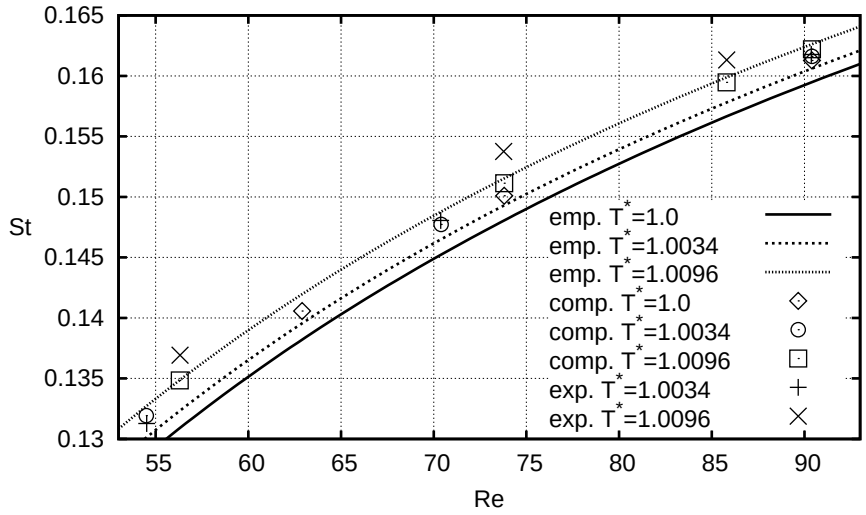


Figure 3.15: Resulting $St - Re$ dependence for various temperatures in flow of water, as calculated on "9-element" mesh (tab. 3.6 and 3.7).

uniform over the element.

The separation angle Θ is measured on the arc defined by frontal stagnation point, cylinder center and the separation point (figure 3.3.5).

We recognized four positions of the zero shear stress on the cylinder surface in the calculated data. These correspond to frontal and backward stagnation points and points of separation on the "top" and "bottom" side of the cylinder. The positions of all the four points oscillate in time with the frequency of vortex shedding. Resulting angles, as dependent on the Reynolds number and temperature, are plotted in figure 3.3.5. Presented results were computed on the "triangular mesh" with number of modes $M = 7$ in both coordinate directions of the standard element.

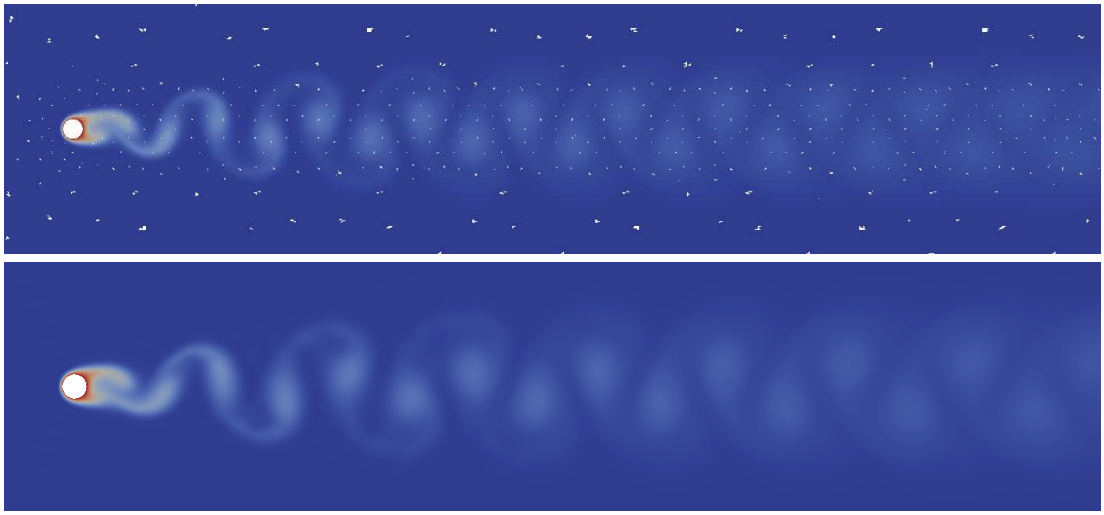


Figure 3.16: Visualisation of the temperature fields calculated at triangular mesh (fig. 3.3) with polynomial order 6, *top*, and 9-element mesh (fig. 3.2) with polynomial order 49, *bottom*.

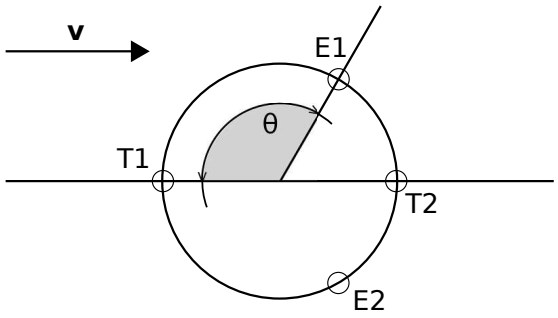


Figure 3.17: The separation angle Θ is measured from time-averaged position of the upstream zero shear stress point. Approximative positions on frontal (T1), backward (T2) stagnation points and two separation points (E1, E2) is drawn.

The amplitude of the oscillations increases with increasing Re in the simulation, as shown in the Figure 3.3.5. The values plotted are summarized in the Table 3.8.

The frontal stagnation point very accurately preserves mean value of $\Theta = 0$. The amplitude of the oscillations of the separation angle is approximately twice the amplitude of oscillations of the frontal stagnation point. Time-averaged positions of the separation points (E1 and E2 in figure 3.3.5) appear symmetrically ($\Theta_{S1} \simeq 360 - \Theta_{S2}$). A difference between angles evaluated on the "triangular" mesh and "9-element" mesh was observed. This deviation in results is up to 2 angular degrees and suggests a need of improvement of the present evaluation method, whose accuracy depends on local distribution of quadrature points, which is non-uniform over the elements and is significant especially in the high-order case.

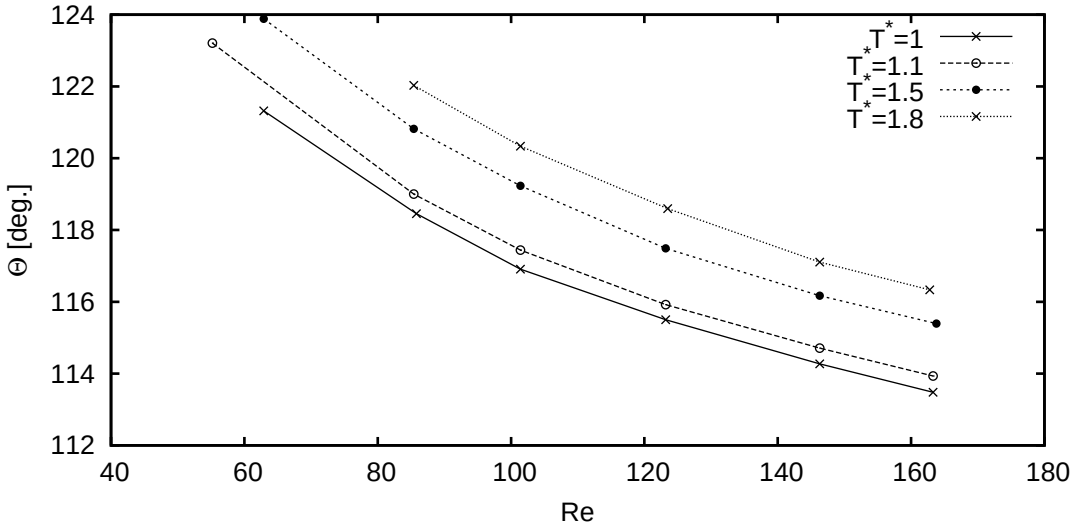


Figure 3.18: The dependence of the separation angle on Re for various temperature ratios $T^* = T_W/T_\infty$. The values of separation angles are the mean values in time. For the maximal and minimal values observed we refer to figure 3.3.5.

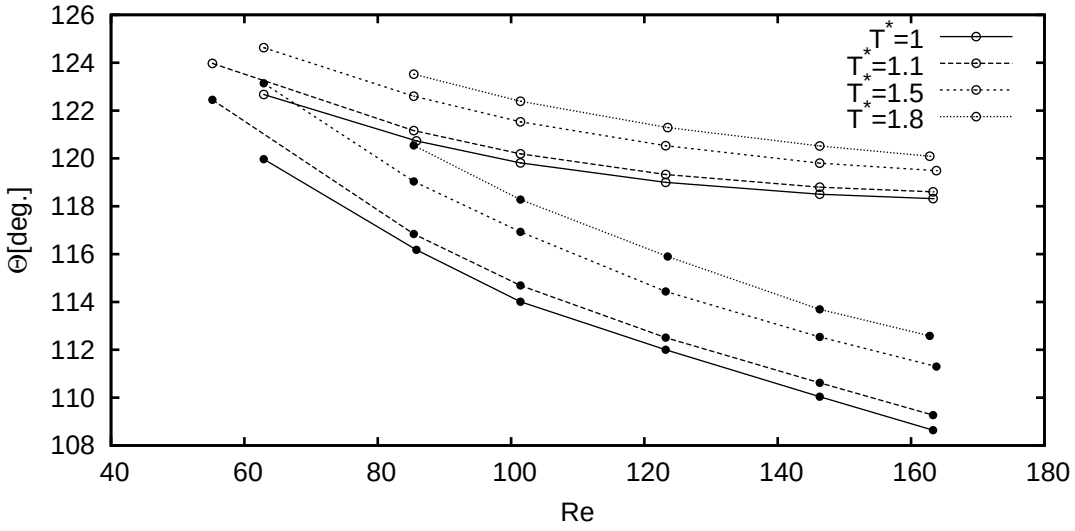


Figure 3.19: Temperature dependence of amplitude of the separation angle for various ratios of flow and cylinder temperatures $T^* = T_W/T_\infty$. Filled points refer to maximal separation angles, empty points denotes the minima.

T^*	Re	Θ_E	$Amp(\Theta_E)$	$Amp(\Theta_T)$
1	$62,9$	$121,32$	$2,7$	$1,2$
	$85,8$	$118,46$	$4,5$	$2,1$
	$101,4$	$116,91$	$5,8$	$2,6$
	$123,2$	$115,50$	$7,0$	$3,1$
	$146,3$	$114,27$	$8,5$	$3,6$
	$163,3$	$113,48$	$9,7$	$3,9$
1.1	$55,2$	$123,21$	$1,5$	$0,7$
	$85,4$	$119,00$	$4,3$	$2,0$
	$101,4$	$117,44$	$5,5$	$2,5$
	$123,2$	$115,92$	$6,8$	$3,0$
	$146,3$	$114,71$	$8,2$	$3,5$
	$163,3$	$113,93$	$9,3$	$3,9$
1.5	$62,9$	$123,89$	$1,5$	$0,8$
	$85,4$	$120,82$	$3,6$	$1,7$
	$101,4$	$119,23$	$4,6$	$2,1$
	$123,2$	$117,49$	$6,1$	$2,7$
	$146,3$	$116,17$	$7,3$	$3,2$
	$163,8$	$115,39$	$8,2$	$3,6$
1.8	$85,4$	$122,03$	$3,0$	$1,4$
	$101,4$	$120,34$	$4,1$	$1,9$
	$123,5$	$118,60$	$5,4$	$2,4$
	$146,3$	$117,11$	$6,8$	$2,9$
	$162,8$	$116,34$	$7,5$	$3,3$

Table 3.8: Resulting values from the computations: Θ_E denotes the time-averaged value of the separation angle, $Amp(\Theta_E)$ is the amplitude of oscillations of the separation angle, while $Amp(\Theta_T)$ is amplitude of the frontal stagnation point.

Conclusion

The thesis concerns a direct numerical simulation of instabilities in a flow of heated fluids. The system of equations was formulated from general physical laws, while neglecting insignificant phenomenons. The resulting system of evolutionary partial differential equations is strongly coupled and concerns temperature dependent material properties. Available results from mathematical analysis, concerning this system were collected, but up to the authors knowledge, there is still no complete analysis formulated, especially due to the restrictions in boundary conditions.

The computational algorithm for the incompressible Navier-Stokes-Fourier system in primitive variables was designed with intention to apply a high order discretisation in spatial coordinates. Properties of the high order discretisations were described on relevant examples, which emerge in the single steps of the time-marching algorithm. Various results of the spatial discretisation, exhibiting exponential convergence and machine precision accuracy, were presented. The spectra of a function in the transform space appeared as an additional tool for analysis of the results, since it contains information related to both the level of accuracy achieved and regularity of the function. Using this tool, an inaccuracy emerging from the incompatibility of the initial data and boundary conditions in the steps of the time-marching scheme, was recognized.

Most of the computations were performed using high order and extremely high order spectral element methods, providing comparison of both approaches. We observed, that a need of breaking the computational domain to elements and the total number of degrees of freedom in the algebraic system are reduced with increasing order of the approximation.

The results from computations, which use the designed algorithm, were presented and show, that the high order approximation give physically realistic results. Concerning the heated cylinder in flow of both air and water, the results reflect the influence of the heating in change of the Strouhal number, which is dependent on both the cylinders temperature and the Reynolds number. The solutions from the numerical simulations were compared with data from physical experiments and exhibit a satisfactory coincidence.

Presented results concerning dependence of the separation angle on temperature in flow around the cylinder have not been investigated in physical experiments so far.

Both the main aims of the work, application of a high order method and construction of an algorithm for temperature dependent fluid flow, were fulfilled and resulted in suggestions for further evolution and a number of possible improvements.

Bibliography

[1] Boyd, J. P.: *Chebyshev and Fourier Spectral Methods*, Dover Publications, Inc., Mineola, New York, 2000.

[2] Braack, M., Mucha, P. B.: Directional do-nothing condition for the Navier-Stokes equations. *J. Comp. Math.*, **32** (no.5, 2014), 507-521.

[3] Bruneau, C.-H., Fabrie, P.: Effective downstream boundary conditions for incompressible Navier-Stokes equations. *Int. J. Numer. Meth. Fl.*, **19** (1994), 693-705.

[4] Bulíček, M., Feireisl, E., Málek, J.: A Navier-Stokes-Fourier system for incompressible fluids with temperature dependent material coefficients. *Nonlinear Anal.-Real.*, **10** (2009), 992-1015.

[5] Burrage, K., Butcher, J. C.: Non-linear stability of a general class of differential equation methods. *BIT*, **20** (1980), 185.

[6] Dong, S., Karniadakis, G. E., Chryssostomidis, C.: A robust and accurate outflow boundary condition for incompressible flow simulations on severely-truncated unbounded domains. *J. Comp. Phys.*, **261** (2015), 83-105.

[7] Cantwell, C. D., Moxey, D., Comerford, A., *et. al.*: Nektar++: An open-source spectral/hp element framework. *Computer Physics Communications*, **192** (2015), 205-219.

[8] Canuto, C., Hussaini, M.Y., Quarteroni, A., Zang, T.A.: *Spectral Methods: Fundamentals in Single Domains*, Springer-Verlag Berlin Heidelberg, 2006.

[9] Canuto, C., Hussaini, M.Y., Quarteroni, A., Zang, T.A.: *Spectral Methods: Evolution to Complex Geometries and Applications to Fluid Dynamics*, Springer-Verlag Berlin Heidelberg, 2006.

[10] Fedorchenko, A., Trávníček, Z., Wang A.-B.: On the effective concept in the problem of laminar vortex shedding behind a heated circular cylinder. *Phys. fluids*, **19** (2007)

[11] Feireisl, E., Málek, J.: On the Navier-Stokes equations with temperature-dependent transport coefficients. *Differ. Equ. Nonlinear Mech.* (2006), 14 pp.(electronic) Art. ID 90616

[12] Forestier,M. Y., Pasquetti, R., Peyret, R., Sabbah, C.: Spatial development of wakes using a spectral multi-domain method. *Appl. Numer. Math.*, **33** (2000), 207-216.

[13] Gear, W.: *Numerical initial value problems in ordinary differential equations*, Englewood Cliffs, NJ: Prentice-Hall, 1971.

[14] Gebhart, B.: *Heat Conduction and Mass Diffusion*, McGraw-Hill, Inc., New York, 1993.

[15] Grisvard, P.: *Elliptic Problems in Nonsmooth Domains*, Pitman, London, 1985.

[16] Guermond, J.L., Shen, J.: Velocity-correction projection methods for incompressible flows. *SIAM J. Numer. Anal.*, **41** (2003), 112.

[17] Guermond, J.L., Shen, J.: A new class of truly consistent splitting schemes for incompressible flows. *J. Comp. Phys.*, **192** (2003), 262-276.

[18] Hesthaven, J. S., Gottlieb, S. G., Gottlieb, D.: *Spectral Methods for Time-Dependent Problems*, Cambridge University Press, 2007.

[19] Karban, P., Mach, F., Kůs, P., Pánek, D., Doležel, I.: Numerical solution of coupled problems using code Agros2D. *Computing*, **95** (2013), 381-408.

[20] John, V.: Reference values for drag and lift of a two-dimensional time-dependent flow around a cylinder. *Int. J. Numer. Meth. Fluids*, **44** (2004), 777-788.

[21] Maday, Y., Patera, A. T., Ronquist, E. M.: An operator-integration-factor splitting method for time-dependent problems: application to the incompressible fluid flow. *J. Sci. Comput.*, **5** no.4 (1990), 263-292.

[22] Karamanos, G., Sherwin, S.: A high order splitting scheme for the Navier-Stokes equations with variable viscosity. *Appl. Numer. Math.*, **33** (2000), 455.

[23] Karniadakis, G. E., Israeli, M., and Orszag, S. A.: High-order splitting methods for incompressible Navier-Stokes equations. *J. Comp. Phys.*, **97** (1991), 414.

[24] Karniadakis, G. E., Triantafyllou, G. S.: Three-dimensional dynamics and transition to turbulence in the wake of bluff objects. *J. Fluid Mech.*, **238** (1992), 1-30.

[25] Karniadakis, G. E., Sherwin, S.: *Spectral/hp Element Methods for Computational Fluid Dynamics.* Oxford University Press, 2005.

[26] Kundu, P., Cohen, I. M.: *Fluid Mechanics, third edition* Elsevier, 2004.

[27] Maršík, F., Trávníček, Z., Yen, R.-H., *et. al.*: Sr-Re-Pr relationship for a heated/cooled cylinder in laminar cross flow. *Proceedings of CHT-08 ICHMT International Symposium on Advances in Computational Heat Transfer*, 2008, Marrakech, Morocco

[28] Orszag, S. A., Israeli, M., Deville M. O.: Boundary Conditions for Incompressible Flows. *J. Sci. Comput.*. **1** (1986), 75-111.

[29] Pech, J.:On computations of temperature dependent incompressible flows by high order methods. *Proceedings of PANM 17 conference* 169-174 (2014), Dolní Maxov.

[30] Pérez, C. E., Thomas, J.-M., Blancher, S., *et. al.*: The steady Navier-Stokes/energy system with temperature-dependent viscosity–Part 1: Analysis of the continuous problem. *Int. J. Numer. Meth. Fluids.* **56** (2008), 63-89.

[31] Pérez, C. E., Thomas, J.-M., Blancher, S., *et. al.*: The steady Navier-Stokes/energy system with temperature-dependent viscosity–Part 2: The discrete problem and numerical experiments. *Int. J. Numer. Meth. Fluids.* **56** (2008), 91-114.

[32] Peyret, R.: *Spectral Methods for Incompressible Viscous Flow.* Springer-Verlag New York, Inc., 2002

[33] Quartapelle, L.: *Numerical Solution of the Incompressible Navier-Stokes Equations.* Birkhäuser, Basel, 1993.

[34] Quarteroni, A., Sacco, R., Saleri, F.: *Numerical Mathematics* Springer Verlag, New York, 2000.

[35] Ramires, M. L. V., Nieto de Castro, C. A., Nagasaka, Y., *et. al.*: Standard reference data for the thermal conductivity of water. *J. Phys. Chem. Ref. Data*,**24** (1995), 1377.

[36] Ren, M.: *3D Flow Transition behind a Heated Cylinder.* Eindhoven University Press, 2005.

[37] Schäfer, M., Turek S.: Benchmark Computations of Laminar Flow Around a Cylinder. *Flow Simulation with High-Performance Computers II, Notes on Numerical Fluid Mechanics (NNFM)* **48** (1996), 547.

[38] Šolín, P., Segeth, K., Doležel, I.: *High-Order Finite Element Methods.* Chapman & Hall/CRC, 2004.

[39] Timmermans, L.: *Analysis of spectral element methods with application to incompressible flow* Thesis Eindhoven (1994), Eindhoven University of Technology.

[40] Vít, T., Ren, M., Trávníček, Z., Maršík, F., Rindt, C.: The influence of temperature gradient on the Strouhal-Reynolds number relationship for water and air.*Exp. Therm. Fluid Sci.* **31** (2007), 751-760

[41] Vos, P. E. J., Chun, S., Bolis, A., *et. al.*: A generic framework for time-stepping partial differential equations (PDEs): general linear methods, object-oriented implementation and application to fluid problems. *Int. J. CFD*, **25** (2011), 107.

[42] Vosse van de, F. N., Minev, P. D.: *Spectral element methods: theory and applications*, Eindhoven University of Technology, 1996.

[43] Wang, A.-B., Trávníček, Z., Chia K.-C.: On the relationship of effective Reynolds number and Strouhal number for the laminar vortex shedding of a heated circular cylinder. *Phys. Fluids*, **12** no.6 (2000), 1401-1410.

[44] Williamson, C. H. K.: Vortex Dynamics in the Cylinder Wake. *Annu. Rev. Fluid. Mech.*, **28** (1996), 477-539.

[45] Williamson, C. H. K., Brown, G. L.: A Series in $(1/\sqrt{Re})$ to Represent the Strouhal-Reynolds Number Relationship of the Cylinder Wake. *J. Fluids Struct.*, **12** (1998), 1073-1085.

[46] Wu, M.-H., Wen C.-Y., Yen R.-H., *et. al.*: Experimental and numerical study of the separation angle for flow around a circular cylinder at low Reynolds number. *J. Fluid Mech.*,**515** (2004), 233.

Appendices

A Elements of convergence theory

Definition 4. *(Wellposedness)*
Equation

$$\begin{aligned} &\frac{\partial u(x,t)}{\partial t} = Lu(x,t), \quad x \in \Omega, \quad t \geq 0, \\ &Bu(x,t) = 0, \qquad\quad x \in \partial\Omega, \quad t > 0, \\ &u(x,0) = g(x), \qquad\quad x \in \Omega, \quad t = 0, \end{aligned} \tag{10}$$

L is independent of time and space and boundary operator B is possibly included in L, is wellposed if,for every $g \in C_0^r$ and for each time $T_0 > 0$ there exists an unique solution $u(x,t)$, which is a classical solution, and such that

$$\|u(t)\| \leq Ce^{\alpha t} \|g\|_{\mathcal{H}^p(\Omega)}, \quad 0 \leq t \leq T_0 \tag{11}$$

for $p \leq r$ and some positive constants C and α. It is strongly well posed if this is true for $p = 0$, i.e., when $\|\cdot\|_{\mathcal{H}^p(\Omega)}$ is the $\mathcal{L}^2$ norm.

Definition 5. *(Convegrence)*
An approximation is convergent if

$$\|u_N(t)P_N u(t)\| \to 0 \text{ as } N \to \infty, \tag{12}$$

$\forall t \in [0,T]$, $u(0) \in \mathcal{B}$ and $u_N(0) \in \mathcal{B}_N$.

Definition 6. *(Consistency)*
An approximation is consistent if

$$\left.\begin{aligned} &\|P_N L(I - P_N)u\| \to 0 \\ &\|P_N u(0) - u_N(0)\| \to 0 \end{aligned}\right\} \text{ as } N \to \infty, \tag{13}$$

$\forall u(0) \in \mathcal{B}$ and $u_N(0) \in \mathcal{B}_N$.

Definition 7. *(Stability)*
An approximation is stable if

$$\left\|e^{L_N t}\right\| \leq C(t), \ \forall N, \tag{14}$$

with the associated operator norm

$$\left\|e^{L_N t}\right\| = \sup_{u \in \mathcal{B}} \frac{\left\|e^{L_N t}u\right\|}{\|u\|}$$

and $C(t)$ is independent of N and bounded for any $t \in [0,T]$.

Theorem 9. *A consistent approximation to a linear wellposed partial differential equation is convergent if and only if it is stable.*

B Fractional step (operator splitting) techniques

An initial value problem in form

$$\begin{aligned} &\frac{\partial u}{\partial t} = \sum_{i=1}^{N} \mathbf{A}_i(u)\,, \quad t \in [t_n, t_n + \Delta t] = [t_n, t_{n+1}] \\ &y(t_n) = u_0 \end{aligned} \tag{15}$$

where u is the solution and $\mathbf{A}_i \quad i = 1, \ldots, N$ are differential operators, may be solved as a sequence of M initial value problems

$$\begin{aligned} &\frac{\partial u^{(j)}}{\partial t} = \mathbf{A}_i(u^{(j)})\,, \\ &t \in [t_1^{(j)}, t_2^{(j)}] \subseteq [t_n, t_{n+1}]\,, \\ &i = 1, \ldots, N\,,\ j = 1, \ldots, M\,,\ M \geq N\,. \end{aligned} \tag{16}$$

$u^{(j)}$ denotes the j-th solution and M is related to particular fractional step method (see below). This method is noted as *method of fractional steps* or *operator splitting*. The sub-problems are concatenated by passing the j-th result to the initial conditions of $(j+1)$-th problem

$$\begin{aligned} &t_1^{(1)} = t_n\,,\ u^{(1)}(t_1^{(1)}) = u_0\,, \\ &u^{(j+1)}(t_1^{(j+1)}) = u^{(j)}(t_2^{(j)})\,,\ \forall j = 1, \ldots, M-1 \\ &t_2^{(M)} = t_{n+1}\,,\ u(t_{n+1}) \approx u^{(M)}(t_2^{(M)})\,. \end{aligned}$$

Concerning properties of particular A_i, this approach allows to use different methods to solve every sub-problem (ODE), concerning properties of every particular A_i.

In case of linear operators, the *splitting error* is the error caused by the sequential application of the operators, depends on value of their commutator

$$[\mathbf{A}_i, \mathbf{A}_j] = \mathbf{A}_i\mathbf{A}_j - \mathbf{A}_j\mathbf{A}_i\,.$$

If the operators commute ($[\mathbf{A}_i, \mathbf{A}_j] = 0$) the splitting error vanishes, what is valid already for the splitting technique of lowest order.

There are various methods of various orders of accuracy.

The basic types of the method for $N = 2$, $u_n = u_0$ (initial condition) or $u_n = u(t_n)$ (solution at n-th step as an initial condition for the $n+1$-st step) are

1. *first order splitting*

   - Lie-Trotter

$$\frac{\partial u^{(1)}}{\partial t} = \mathbf{A}_1(u^{(1)})\,,\ u^{(1)}(t_n) = u_n \tag{17}$$

$$\frac{\partial u^{(2)}}{\partial t} = \mathbf{A}_2(u^{(2)})\,,\ u^{(2)}(t_n) = u^{(1)}(t_{n+1}) \tag{18}$$

- Additive splitting

$$\frac{\partial u^{(1)}}{\partial t} = \mathbf{A}_1(u^{(1)})\,,\; u^{(1)}(t_n) = u_n \tag{19}$$

$$\frac{\partial u^{(2)}}{\partial t} = \mathbf{A}_2(u^{(2)})\,,\; u^{(2)}(t_n) = u_n \tag{20}$$

$$u(t_{n+1}) = u^{(2)}(t_{n+1}) + u^{(1)}(t_{n+1}) - u_n \tag{21}$$

2. *second order splitting*

- The Strang's splitting performs the problem in three steps ($N = 2$, $M = 3$)

$$\frac{\partial u^{(1)}}{\partial t} = \mathbf{A}_1(u^{(1)})\,,\; u^{(1)}(t_n) = u_n \tag{22}$$

$$\frac{\partial u^{(2)}}{\partial t} = \mathbf{A}_2(u^{(2)})\,,\; u^{(2)}(t_n) = u^{(1)}(t_{n+1/2}) \tag{23}$$

$$\frac{\partial u^{(3)}}{\partial t} = \mathbf{A}_1(u^{(3)})\,,\; u^{(3)}(t_{n+1/2}) = u^{(2)}(t_{n+1}) \tag{24}$$

  Note, that the first and third step is discretized with half step since $t_{n+1/2} = t_n + \dfrac{\Delta t}{2}$.

- Symmetrically weighted splitting ($N = 2$,$M = 4$) performs firstly the Lie-Trotter splitting with result $u^{(2)}(t_{n+1})$

$$\frac{\partial u^{(1)}}{\partial t} = \mathbf{A}_1(u^{(1)})\,,\; u^{(1)}(t_n) = u_n \tag{25}$$

$$\frac{\partial u^{(2)}}{\partial t} = \mathbf{A}_2(u^{(2)})\,,\; u^{(2)}(t_n) = u^{(1)}(t_{n+1}) \tag{26}$$

  Second part is again the Lie-Trotter approach, but the operators act in reverse order

$$\frac{\partial u^{(3)}}{\partial t} = \mathbf{A}_2(u^{(3)})\,,\; u^{(3)}(t_n) = u_n \tag{27}$$

$$\frac{\partial u^{(4)}}{\partial t} = \mathbf{A}_4(u^{(4)})\,,\; u^{(4)}(t_n) = u^{(3)}(t_{n+1}) \tag{28}$$

  The solution is obtained as an average of both results

$$u_{n+1} = \frac{u^{(2)}(t_{n+1}) + u^{(4)}(t_{n+1})}{2} \tag{29}$$

Operator splitting technique based on integration factor was used in combination with spectral method in (Timmermans [39]) and is briefly introduced in the following section.

Operator splitting- formulation by integration factor

Splitting of a differential operator to a set of simpler problems, allows one to use different approximation methods in the solution to the sub-problems. Lets take the convection diffusion problem

$$\frac{\partial \mathbf{u}}{\partial t} = \mathcal{D}\mathbf{u} + \mathcal{C}\mathbf{u} + \mathbf{f} \tag{30}$$

for illustration. $\mathcal{D}\mathbf{u} = \nu\nabla^2\mathbf{u}$ is the diffusion operator and $\mathcal{C}\mathbf{u} = \mathbf{v}\cdot\nabla\mathbf{u}$ is (linearized) convection operator with known velocity $\mathbf{v}$. Following (Maday [21]) we introduce an integrating factor ${\mathcal{F}_\mathcal{C}}^{(t^*,t)}$ in $\mathcal{C}$, satisfying

$$\frac{\partial}{\partial t}{\mathcal{F}_\mathcal{C}}^{(t^*,t)} = -{\mathcal{F}_\mathcal{C}}^{(t^*,t)}\mathcal{C}, {\mathcal{F}_\mathcal{C}}^{(t^*,t^*)} = \mathbb{I} \tag{31}$$

to the (30)

$$\frac{\partial}{\partial t}\left({\mathcal{F}_\mathcal{C}}^{(t^*,t)}\mathbf{u}(t)\right) = {\mathcal{F}_\mathcal{C}}^{(t^*,t)}(\mathcal{D}\mathbf{u} + \mathbf{f}). \tag{32}$$

If the backward differences are used as approximation to the time derivative, we arrive to

$$\frac{\gamma_0\mathbf{u}^{n+1} - \sum_{q=0}^{J}\alpha_q{\mathcal{F}_\mathcal{C}}^{(t^{n+1},t^{n+1-q})}\mathbf{u}^{n+1-q}}{\Delta t} = \mathcal{D}\mathbf{u}^{n+1} + \mathbf{f}^{n+1}, \tag{33}$$

where condition ${\mathcal{F}_\mathcal{C}}^{(t^{n+1},t^{n+1})} = \mathcal{I}$ was applied in the first term of the nominator. To complete the scheme, we have to provide method of calculation of terms ${\mathcal{F}_\mathcal{C}}^{(t^{n+1},t^{n+1-q})}\mathbf{u}^{n+1-q}$. It is not necessary to explicitly construct the integrating factor $\mathcal{F}_\mathcal{C}$, but we arrive to associated initial value problem

$$\frac{\partial\tilde{\mathbf{u}}(s)}{\partial s} = \mathcal{C}\tilde{\mathbf{u}}(s), 0 < s < q\Delta t\tilde{\mathbf{u}}(0) = \mathbf{u}^{n+1-q} \tag{34}$$

and finally

$${\mathcal{F}_\mathcal{C}}^{(t^{n+1},t^{n+1-q})}\mathbf{u}^{n+1-q} = \tilde{\mathbf{u}}(q\Delta t) \tag{35}$$

C Other splitting schemes for the Inc. Navier-Stokes system

The splitting scheme designed for the incompressible Navier-Stokes equations (Karniadakis [23]) became the basis for the scheme solving the temperature dependent flow in this thesis. It also motivated development of more recent splitting schemes for the incompressible Navier-Stokes equations, which will be summarized in this paragraph. From comparison of the recent schemes it follows, that an important role in accuracy of the scheme plays satisfaction of the Babuška-Brezzi condition.

Babuška-Brezzi condition

$$\inf_{q\in\mathcal{L}^2(\Omega)}\sup_{\mathbf{v}\in\boldsymbol{\mathcal{H}}^1_{0,\Gamma_D}(\Omega)}\frac{(\operatorname{div}\mathbf{v}, q)}{\|\mathbf{v}\|_1\,\|q\|} \geq \beta \tag{36}$$

for some positive number β, reflects the fact, that the Navier-Stokes equations are solved with only the gradient of pressure. It was recognised, that those schemes, which satisfy this condition are of higher accuracy (Guermond [17]). This is closely related to the rotational form of the diffusion term introduced firstly in the pressure-Neumann condition (2.43).

In solution techniques to the Incompressible Navier-Stokes equations in primitive variables it has been common practice to extrapolate the non-linear term and proceed by solution of the linearized system, the incompressible Stokes problem (see section 2.1.1). This is the case of semi-implicit velocity-correction scheme (Karniadakis [23]), which firstly established a priori control on the divergence by introduction of the rotational form of the boundary condition 2.43. This pressure boundary condition controls the divergence boundary layer, but the vorticity is not still accurate. A strategy, which incorporates the rotational form of the diffusion term directly into the solved equations, leaves the pressure boundary condition and results in *consistent splitting* schemes, was proposed by Guermond ([17]). The key idea stays in testing the momentum equation against gradients $\nabla\phi$, when applying $\left(\frac{\partial \mathbf{u}}{\partial t}, \nabla\phi\right) = -\left(\nabla\cdot\left(\frac{\partial \mathbf{u}}{\partial t}\right), \phi\right) = 0$ to obtain

$$\int_\Omega \nabla p\cdot\nabla\phi = \int_\Omega (\nu\nabla^2\mathbf{u} - \mathbf{u}\cdot\nabla\mathbf{u} + \mathbf{f})\cdot\nabla\phi\,,\ \forall\phi\in\mathcal{H}^1(\Omega) \tag{37}$$

what together with substitution $\nabla^2\mathbf{u}\to -\nabla\times\nabla\times\mathbf{u}$ (this repeatedly removes the term $\nabla\nabla\cdot\mathbf{u}$) leads to

$$\frac{\sum_{q=0}^{Q}\alpha_q\mathbf{u}_{n+1-q}}{\Delta t} - \nu\nabla^2\mathbf{u}_{n+1} + \nabla p^\star = \mathbf{f}(t_{n+1}) - (\mathbf{u}\cdot\nabla\mathbf{u})^\star \tag{38}$$

$$(\nabla p_{n+1},\nabla\phi) = (\mathbf{f}(t_{n+1}) - (\mathbf{u}\cdot\nabla\mathbf{u})^\star - \nu\nabla\times\nabla\times\mathbf{u}_{n+1},\nabla\phi),\ \forall\phi\in\mathcal{H}^1(\Omega)\,, \tag{39}$$

where α_q are coefficients of the BDF (see table 2.1) and symbol $\star$ denotes extrapolation, e.g. $p^\star = \sum_{q=0}^{Q-1}\beta_q p_{n-q}$. The term $\nabla\times\nabla\times\mathbf{u}$ is problematic in spatial discretisation using methods providing only the C^0 continuity (FEM, SEM, ...). Its direct evaluation is therfore circumvented by subtracting the L^2 inner product of the equation 38 and gradients $\nabla\phi$ from 39, what results in new form of (39)

$$(\nabla\psi_{n+1},\nabla\phi) = \left(\frac{\sum_{q=0}^{Q}\alpha_q\mathbf{u}_{n+1-q}}{\Delta t},\nabla\phi\right)\,,\ \forall\phi\in\mathcal{H}^1(\Omega)\,, \tag{40}$$

where $p_{n+1} = \psi_{n+1} + p^\star - \nu\nabla\cdot\mathbf{u}_{n+1}$.

The "KIO" scheme was revised in [16]. The approach of rotational diffusion term ($\nabla^2\mathbf{u} = -\nabla\times\nabla\times\mathbf{u}$) was introduced directly in the formulation of the solved equations, but still it preserves splitting of the non-linearity from the Stokes system. In the first step the pressure-Poisson equation

$$\begin{cases} \dfrac{1}{\Delta t}\left(\alpha_0\mathbf{u}_{n+1} + \displaystyle\sum_{q=1}^{Q}\alpha_q\tilde{u}_{n+1-q}\right) + \nabla\times\nabla\times\tilde{\mathbf{u}}_n + \nabla p_{n+1} = f(t_{n+1}) \\ \nabla\cdot\mathbf{u}_{n+1} = 0 \\ \mathbf{u}_{k+1}\cdot\mathbf{n}|_\Gamma = 0 \end{cases} \tag{41}$$

is solved. This step also applies the compatibility condition $\mathbf{u}_{n+1} \cdot \mathbf{n} = 0$ and results in evaluation of the intermediate velocity $\mathbf{u}_{n+1}$. The correction comes in the second step

$$\begin{cases} \dfrac{\alpha_0}{\Delta t}(\tilde{\mathbf{u}}_{n+1} - \mathbf{u}_{n+1}) - \nabla^2 \tilde{\mathbf{u}}_{n+1} - \nabla \times \nabla \times \tilde{\mathbf{u}}_{n+1} = 0 \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \tilde{\mathbf{u}}_{n+1}|_\Gamma = 0 \end{cases} \tag{42}$$

This scheme fulfills

$$\frac{\partial p_{n+1}}{\partial \mathbf{n}}|_\Gamma = (f(t_{n+1}) + \nabla^2 \tilde{\mathbf{u}}_{n+1}) \cdot \mathbf{n}|_\Gamma \,, \tag{43}$$

which is a consistent Neumann pressure boundary condition. Finally, the KIO scheme is exposed as a representative of this rotational velocity-correction schemes and its analysis is provided.

More recently, Dong [6] proposed a scheme, which do not include any extrapolation neither on the convective term nor the pressure. The scheme is therefore expected to be unconditionally stable. Since it enhances the rotational velocity-correction scheme of Guermond ([16]) it adopts the similar estimates for order of accuracy.

$$\begin{gathered} \frac{1}{\Delta t}\left(\alpha_0 \tilde{\mathbf{u}}_{n+1} \textstyle\sum_{q=1}^{Q} \alpha_q \mathbf{u}_{n+1-q}\right) + \nabla p_{n+1} + \mathbf{u}_n \cdot \nabla \mathbf{u}_n + \nu \nabla \times \nabla \times \mathbf{u}_n = f(t_{n+1}) \\ \nabla \cdot \tilde{\mathbf{u}}_{n+1} = 0 \\ \mathbf{n} \cdot \tilde{\mathbf{u}}_{n+1}|_\Gamma = \mathbf{n} \cdot \mathbf{w}_{n+1} \,, \end{gathered} \tag{44}$$

where $\mathbf{w}$ refers to the nonhomogeneous Dirichlet boundary condition $\mathbf{u}|_\Gamma = \mathbf{w}$. The second substep then corrects the velocity

$$\begin{gathered} \frac{\alpha_0}{\Delta t}(\mathbf{u}_{n+1} - \tilde{\mathbf{u}}_{n+1}) - \nu \nabla^2 \mathbf{u}_{n+1} + \tilde{\mathbf{u}}_{n+1} \cdot \nabla \mathbf{u}_{n+1} - \mathbf{u}_n \cdot \nabla \mathbf{u}_n - \nu \nabla \times \nabla \times \mathbf{u}_n = 0 \\ \mathbf{u}_{n+1}|_\Gamma = \mathbf{w}_{n+1} \,. \end{gathered} \tag{45}$$

D Jacobi polynomials

Recursion relations ([25])

$$\begin{aligned} P_0^{\alpha,\beta}(x) &= 1 \,, \\ P_1^{\alpha,\beta}(x) &= \frac{1}{2}[\alpha - \beta + (\alpha + \beta + 2)x] \,, \\ a_n^1 \, P_{n+1}^{\alpha,\beta}(x) &= (a_n^2 + a_n^3 x) P_n^{\alpha,\beta}(x) - a_n^4 P_{n-1}^{\alpha,\beta}(x) \,, \end{aligned} \tag{46}$$

$$\begin{aligned} a_n^1 &= 2(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta) \,, \\ a_n^2 &= (2n+\alpha+\beta+1)(\alpha^2-\beta^2) \,, \\ a_n^3 &= (2n+\alpha+\beta)(2n+\alpha+\beta+1)(2n+\alpha+\beta+2) \,, \\ a_n^4 &= 2(n+\alpha)(n+\beta)(2n+\alpha+\beta+2) \,, \end{aligned}$$

Recursion relations for derivatives of Jacobi polynomials. These are needed in Galerkin formulation of second order PDEs.

$$\begin{aligned} b_n^1(x)\frac{\mathrm{d}}{\mathrm{d}x}P_n^{\alpha,\beta}(x) &= b_n^2(x)P_n^{\alpha,\beta}(x) + b_n^3(x)P_{n-1}^{\alpha,\beta}(x)\,, \\ b_n^1(x) &= (2n+\alpha+\beta)(1-x^2)\,, \\ b_n^2(x) &= n[\alpha-\beta-(2n+\alpha+\beta)x]\,, \\ b_n^3(x) &= 2(n+\alpha)(n+\beta) \end{aligned} \tag{47}$$

Another useful relations

1. Value at the end-point of the "standard" interval $[-1,1]$

$$P_n^{\alpha,\beta}(1) = \frac{(n+\alpha)!}{\alpha!n!}$$

2. "Symmetry"

$$P_n^{\alpha,\beta}(-x) = (-1)^n P_n^{\beta,\alpha}(x)$$

3. Relation among types of the Jacobi polynomials in sense of a k-th derivative

$$\frac{\mathrm{d}^k}{\mathrm{d}x^k}P_n^{\alpha,\beta}(x) = \left(\frac{1}{2}\right)^k \frac{\Gamma(\alpha+\beta+n+k+1)}{\Gamma(\alpha+\beta+n+1)}P_{n+k}^{\alpha+k,\beta+k}(x)$$

Consequences

- 1. & 2. $\Rightarrow P_n^{\alpha,\beta}(-1) = (-1)^n\frac{(n+\beta)!}{\beta!n!}$
- 2. $\Rightarrow$ ultraspherical polynomials are even (or odd) with respect to the origin

$$P_n^{\alpha}(-x) = (-1)^n P_n^{\alpha}(x)$$

E Extension to time integration methods

Variable timestep in IMEX methods

Let $\Delta t = t_{n+1} - t_n$ and $r_{n+2-m}\Delta t = t_{n+2-m} - t_{n+1-m}$ for $m \geq 2$. Coefficients of the second and third order stiffly-stable mixed schemes are as follows

the schemes still need values from multiple previous time steps

Table 9: Coefficients of variable time-step mixed stiffly-stable schemes ([32]).

Coefficient	2nd order	3rd order
γ_0	$\frac{2+r_n}{1+r_n}$	$1+\frac{1}{1+r_n}+\frac{1}{1+r_n+r_{n-1}}$
α_0	$-1-\frac{1}{r_n}$	$-\frac{(1+r_n)(1+r_n+r_{n-1})}{r_n(r_n+r_{n-1})}$
α_1	$\frac{1}{1+r_n}$	$\frac{1+r_n+r_{n-1}}{r_n r_{n-1}(1+r_n)}$
α_2		$-\frac{1+r_n}{r_{n-1}(r_n+r_{n-1})(1+r_n+r_{n-1})}$
β_0	$1+\frac{1}{r_n}$	$\frac{(1+r_n)(1+r_n+r_{n-1})}{r_n(r_n+r_{n-1})}$
β_1	$-\frac{1}{r_n}$	$-\frac{1+r_n+r_{n-1}}{r_n r_{n-1}}$
β_2		$\frac{1+r_n}{r_{n-1}(r_n+r_{n-1})}$