Numerical solution of the Fokker-Planck equation using physics-conforming finite element methods

---

Dissertation
zur Erlangung des akademischen Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

Der Fakultät für Mathematik der
Technischen Universität Dortmund
vorgelegt von

Katharina Theresa Wegener

im Juli 2024

Dissertation

Numerical solution of the Fokker-Planck equation using physics-conforming finite element methods

Fakultät für Mathematik
Technische Universität Dortmund

Erstgutachter: Prof. Dr. Dmitri Kuzmin
Zweitgutachter: Prof. Dr. Stefan Turek

Tag der mündlichen Prüfung: 20. September 2024

Acknowledgments

First of all, I would like to thank my supervisor Prof. Dr. Dmitri Kuzmin very much. Thanks to his expertise, things always moved forward, even when they seemed stuck. I am grateful, for example, looking back at the numerous mails with his answers to my questions about limiting during the pandemic home office period. I am also very grateful for his proofreading of this thesis. Without Dmitri's profound knowledge of the topics and his guidance, this thesis would not have come to be.

Two other people without whom this work would not have been possible are Peter Zajac and Dirk Ribbrock. In my early days at the LS3, they took a lot of time to introduce me to the 'Finite Element Analysis Toolkit 3' (FEAT3) software package (`http://www.featflow.de/en/software/feat3.html`). Their support ranged from explaining numerous mathematical backgrounds in FEAT3 to fixing Git issues.

Another big thank you goes to Dr. Christoph Lohmann, who was always available for discussions and questions and who also proofread parts of this thesis. I also would like to say thank to Prof. Dr. Stefan Turek for his support. Without him, I would not have started at the chair. Not least his way of running the LS3 the way he does probably contributes significantly to the atmosphere at the chair.

PD Dr. Andriy Sokolov gave a decisive impulse to the usage of finite elements on the sphere. Maximilian Esser took on the task of implementing a solver for the NSE and wrote it in such a way that we could quickly combine our components. Another thank you goes to the German Research Association (DFG), which financially supported this project 401649630, as well as to Dr. Kevin Breuer, who worked on the engineering side of the project.

Standing for the entire chair (a complete list would be too long), a special thanks goes to my current and former office colleagues Rida Ahmad, Michael Fast, Hannes Ruelmann and Dr. Florian Streitbürger. We are probably the office with the biggest appetite for cake, the most colorful blackboard and the most lovingly named plants. The LS3 is a chair with an atmosphere that makes you like to come to work.

Finally, a big thank you to my friends and my family, my parents and my brothers for being there. Special greetings go out to my math grandpa 'somewhere in heaven'...

Katharina Wegener

Contents

Abstract

In this work, a Fokker-Planck equation (FPE) is used to approximate the orientation distribution of fibers. FPEs combined with the Navier-Stokes equations (NSE) are widely used to predict the motion of the fibers in fiber suspension flows with low Reynolds numbers. The fibers align in response to the flow and randomize in response to fiber-fiber interactions. A precise formulation takes into account that the flow-fiber interaction is bilateral, so that the suspension rheology also depends on the fiber orientation.
Various approaches to model fiber suspensions, including the well-known Folgar-Tucker equation, which relies on orientation tensors, are reviewed. We aim to solve the FPE using the continuous Galerkin method. For each point in the 3d physical space, an equation on the surface of a unit sphere representing the orientation states is solved, while for each point on the sphere an advection equation in the 3d physical space has to be solved. We handle this in the framework of an alternating direction approach including subtime stepping. Algebraic flux correction is performed for each equation to ensure positivity preservation as well as the normalization property of the distribution function.
Numerical tests are performed for the individual subproblems. Finally, the velocity field is calculated by the incompressible Navier-Stokes equations (NSE), and benchmark problems for the coupled FPE-NSE system are solved. Thus, the relevance of this two-way coupling across the scales can be validated, and the effect of a different number of fibers is examined.

1 Introduction and motivation

1.1 Application: Fiber suspensions

Multiphase flows play a central role both in nature and in industrial applications. Such flows consist of different ingredients, referred to as phases, that interact with each other within the mixture. They occur in many chemical processes, in physics, in engineering and more. Even in the year 2024 multiphase flows pose an enormous challenge.
The simplest example of a multiphase flow is a two-phase flow. Among the first two-phase flows that were observed from a scientific point of view we find geophysical flows like dust in the air or clouds [Dre83]. The two-phase flow studied in this thesis is a fiber suspension, where solid fibers are dispersed in a fluid. Such flows laden with particles arise in the process of paper production [LU07, KOM09, HLHN11, Loh16b], or can be used for polymer turbulent drag reduction [ZMLD+12].

This work is motivated by the injection molding of fiber-reinforced plastics, see, e.g., [Ver98]. Fibers are added to the melt used within the process of injection molding to improve the properties of the resulting material. The reinforcement with fibers can offer both mechanical and economic advantages [Fol82, LDHB88]. So engineers need reliable simulation tools to predict the properties of the plastics. In fact, the fibers significantly influence the microstructure of the composite, which in turn controls mechanical properties like stiffness and strength, or the electrical and thermal conductivity of the finished product [GGOS20].
The microstructural change investigated in this work is the orientational structure of the fibers. Our material is stronger and stiffer in the direction in which more fibers are oriented [FT84]. Imagine many straws parallel to each other. They are most stable along the orientation direction, which is perpendicular to the orientation direction of the straws. Whether a high fiber alignment or a uniform fiber orientation is preferred depends on the application [OFCH04].

An important insight, which has been established for fiber suspension rheology over the last few decades, is that not only the flow influences the fibers ('one-way coupling'), but that vice versa the flow field depends on the orientation state of the suspension ('two-way coupling') [Bat70, LDHB88, VT02, Lin08].

1.2 General modeling approaches

The main objective of this thesis is to describe the behavior of fiber suspension flows and in particular the orientation of these fibers. If we detach a little from our specific application, there are related applications that can be modeled in a similar

way. In addition to fibers, particles of different nature such as polymers can be considered. The unifying point is that we have both a fluid phase and a particle phase. Therefore our problem can be embedded in a broader context of formulations and solution strategies. From a mathematical perspective the distinction between fully macroscopic approaches and so-called multiscale methods is of interest. For a detailed review, the survey article [Keu04] is recommended. Even though it was written two decades ago, it still represents an informative summary. Following that article, three basic approaches can be identified, namely

1.) fully macroscopic approaches,
2.) stochastic approaches for multiscale methods,
3.) fully deterministic approaches for multiscale methods.

Fully macroscopic approaches. Fully macroscopic approaches take a macroscopic perspective for both phases, that is, for the fluid as well as for the particle ensembles. They are also known as continuum approaches, since they assume that a composite can be considered as a continuum [OP02]. This description is the oldest one as it requires comparatively little computing power. However, even this system represents a highly complex mathematical problem.
Two competing derivations are distinguished: postulation and averaging [Dre83, DP99]. In the first case the equation is postulated without any references to the microscopic scale. The procedure is purely phenomenological. The alternative are averaging processes, where we may gain an insight into the relationship between micro- and macroscopic variables [DP99]. Both derivations should ultimately yield the same result.

Multiscale approaches. Multiscale approaches emerged increasingly around the turn of the millennium. A coupling across the scales is inherent to fluid dynamics of particles [Kne06]. On the one hand, there is an equation for the particles suspended in the fluid. On the other hand, there are the governing equations for the carrier fluid. The coupled equations form a multiscale problem. Hence, we have a micro-macro coupling, that is, in our application a fiber-flow modeling.

Regarding the scale to describe the particle phase there are very different potential levels of consideration. Overall, there is a tension between the degree of sophistication of the model and its computational tractability [OP02]. Starting at the bottom level, atomistic models might be developed. Even though they are hypothetically interesting to gain an in-depth insight of some phenomena, the idea to use such models has to be discarded because of the computational resources this would consume [Keu04]. In addition, too many details may hide the main result [DP99]. The next level is a coarse-grained description of molecular configurations. Such a macroscopic description stays computationally intensive, but it is within the realm of possibility. Furthermore, we work with statistical quantities, that do not mask the main information. These quantities range from averaged particle velocity to a

particle orientation distribution [ZMLD+12]. Sometimes they are more accurately referred to as mesoscopic part.

The governing equations, which model the surrounding fluid, on the macroscopic level are shared by all the three basic approaches listed above. By default, these equations are the conservation laws for mass and momentum, combined within the framework of (Navier-)Stokes equations, and supplemented by a constitutive law for the stress. While the carrier fluid considered on its own is a Newtonian fluid, it becomes a non-Newtonian fluid in the mixture.

Stochastic vs. deterministic approaches. The focus of this work is on the microscopic ingredients of the system. For the corresponding equations, we distinguish between stochastic and deterministic approaches. A natural starting point to compare the two approaches is the Brownian motion. Considering a single particle in a fluid, Brownian motion is caused by collisions of this particle with the molecules of the fluid. The effect of the resulting rapid fluctuations, which occur on extremely short length scales, is usually taken into account through random forces on a coarser level [Ött12]. Such phenomena can be modeled by stochastic differential equations (SDEs), where a so-called noise term represents the fluctuating random forces. Since in each SDE at least one of the terms and, hence, the solution are of stochastic nature, this calls for stochastic numerical methods.

An alternative to the stochastic simulations are Fokker-Planck equations (FPEs), which are deterministic PDEs. A whole book was dedicated to these equations by Risken [Ris96]. The book contains an extensive discussion of the derivation of the FPE, some of its applications, and methods to solve the equation. In particular, exact solutions and expressions for some special cases are derived. In this thesis, the possibility of deriving such a solution is seen in the example of the extended Jeffery equation. However, analytical solutions are an exception.

We can draw a link between the stochastic and the deterministic approach. For instance, the theses [Loz03, Kne08] contain derivations of a deterministic FPE from a stochastic approach. Under certain conditions [OP02, Th.11.1], there is a formal equivalence between an FPE and its associated SDE [LC03, Ött12]. Nevertheless, there is no one-to-one correspondence between SDEs and FPEs. An SDE, on the one hand, specifies actual trajectories. In our case of fiber suspensions, the fibers themselves can only be directly visualized using a stochastic approach. The FPE, on the other hand, describes the corresponding probability distribution of the stochastic process [Ött12]. Once the distribution function is known, the relevant macroscopic variables, given by so-called orientation tensors in this thesis, can be calculated as statistical averages [Ris96, OP02].

Stochastic methods. The idea of using a stochastic approach for quantitative prediction purposes may appear odd at first. In fact, however, an exact solution can be approximated by examining a representative sample of states, and the more states are added, the smaller the error becomes [Ött12]. In the context of our micro-macro approaches, a well-known methodology bears the acronym CONNFFESSIT ('Calculation of non-Newtonian Flow: Finite Elements and Stochastic Simulation

Technique'). The method can be described in three steps, which are repeated until convergence is reached [LC03, Keu04]. First, the conservation laws are solved with a finite element approach to obtain velocity and pressure fields. Then, the paths of the model particles are calculated using the corresponding SDE. Finally, the stress field, needed for the momentum equation in the first step, is obtained based on these numerous configurations. In [NKA20] another stochastic approach for the FPE with fiber suspensions as application is given.

Although SDEs might be numerically more tractable than the associated FPE, there are several shortcomings [Loz03, Keu04]. Due to a large number of necessary trajectories the computational effort and the memory requirement are huge. In addition, the stress field is typically not smooth. Last but not least, averaging introduces a statistical error, so that there is statistical noise in the computed extra stress.

Eulerian and Lagrangian approaches. Another well-established distinction for the disperse phase is between Lagrangian and Eulerian methods [OFCH04, KOM09].

In the Lagrangian perspective the individual particles are tracked. For each particle of the suspension an equation of motion is solved for a given velocity field. In the case of a parallelized code, up to ten million particles or particle ensembles can be considered and the results can be expected to be quite accurate. However, the Lagrangian approach always requires an update of the particle position and therefore is not appropriate for steady-state computations. Moreover, the computational effort is proportional to the number of particles. A Lagrangian treatment of a fiber suspension can be found in [NKA20]. There, a relatively simple model is used by assuming that the particles move with the velocity field, and ignoring other forces. The fiber-fiber interactions are modeled by adding suitable random perturbations to the orientation angles.

Using the Eulerian perspective, we focus on a fixed location in space through which the mixture flows [Ran17c]. We determine the orientation distribution of the particles by solving a Fokker-Planck equation. Since the computational cost does not directly depend on the number of particles, in particular for non-dilute suspensions in complex geometries, the Eulerian approach is computationally more efficient than the Lagrangian approach.

1.3 Approaches for Fokker-Planck equation

We present different approaches from the literature for solving a Fokker-Planck equation in the context of fiber suspension. Our FPE models the development of an orientation distribution function, which in turn specifies the local orientation of the fibers suspended in the fluid [KOM09].

The beginning. At the turn of the millennium, there was a number of stochastic methods to simulate different mixtures numerically, whereas deterministic approaches based on the FPE were rare in the literature. Warner's work from 1972 [WJ72] is an exception. It considers an FPE without a spatial convection term,

but in the three-dimensional space. The pioneering work of Warner was followed by different works of Fan [Fan85, Fan89], who adopted and expanded the idea. In [Fan85], Warner's original idea was improved under the assumption that the probability density function is smooth, because otherwise the Galerkin method produces incorrect results in the case of a singularity. Presumably the first attempt to solve the FPE in a complex geometry was published in [Fan89].

Basic deterministic methods. First, the orientation distribution function describes the probability of the fibers to be present at a particular location in a given domain, the physical space. In addition, each fiber has a specific orientation. The space of all possible orientations is called configuration space [Loz03, KS09b] or orientation space [PT09]. In our case, it is simply the surface of the unit sphere.

In what follows, we compare deterministic approaches to solve the FPE. We categorize them into finite element, finite volume and spectral methods.

Let us start with the finite element method (FEM). To some extent for the FPE finite elements might be considered as a brute force method. However, e.g., in [FVDC92, Kne08, KS09a] and [ADS+24] the FPE or at least parts of the FPE are solved using this method. All in all, FEM is a very general approach, that can be adapted and extended to specific requirements and is flexible with respect to different domains covered by the FPE. Therefore physics-conforming finite element discretizations are considered in this thesis.

The alternative finite volume method (FVM) is particularly well suited for convection dominated problems. It is well known to preserve local conservation properties. Hence, it has been widely used in the literature for the FPE, for example, in [HO06]. In [FHH+08], a large spectrum of Péclet numbers, which characterize the ratio between the convective and the diffusive transport rate, is examined. The finite volume approach in [ZMLD+12] provides the opportunity to adapt the spherical mesh in regions of steep gradients.

Spectral methods are widely used to solve the part of the Fokker-Planck equation in the configuration space [Loz03, KS09b]. Spherical harmonics, which are the eigenfunctions of the Laplace operator on the sphere, are expected to be the optimal choice for the basis functions [Keu04]. In fact, that is true for the diffusion dominated case corresponding to a small Péclet number [HO06]. If this changes, the spherical harmonics deviate from being eigenfunctions [Keu04, ZMLD+12], so that a high number of basis functions is required to obtain reliable results. The work [Loh16b] uses a spectral method and considers the FPE without diffusion. In [DH22] the FPE including diffusivity is solved and therefore relatively few spherical harmonics are needed. Another downside of conventional spectral methods is that the basis functions often have a global support and therefore yield dense instead of sparse matrices. A counter-example is given by [Loh16b], where Fourier basis functions are applied and their orthogonality property is employed. An alternative are spectral

element methods, where approximations are developed separately on subdomains and subsequently patched together [OP02].

The Folgar-Tucker equation (FTE) can be understood as a low-order spectral method for the FPE [Kuz18], exploiting an approximation with only a few spherical harmonics. It has already been stated in [AT87] that the orientation tensors within the FTE are related to the Fourier series expansion coefficients of the distribution function. Further sources, which describe how to simulate fiber suspensions by solving the FTE, are [Tuc91, CT95]. A more recent work, which solves the FTE by means of finite elements and additionally takes into account the preservation of physical properties, is [Loh19].

Further literature. In a series of papers published by Chauvière and Lozinski shortly after the turn of the millennium, both stochastic and deterministic approaches to solve the FPE were examined [LC03, Loz03, CL04a, CL04b]. The authors significantly advanced the state of art for deterministic approaches. They introduced an operator splitting approach, applied a Galerkin spectral method in configuration space and a streamline upwind Petrov-Galerkin (SUPG) spectral element method in physical space. It was shown that deterministic approaches can outperform traditional stochastic simulations in the case of a low dimensional configuration space, i.e., if the dimension is not greater than three. Consequently, for our orientation space the deterministic approach is competitive. However, the question remains whether it can be recommended for more configurational degrees of freedom.

The contributions of Knezevic and Süli [Kne06, Kne08, KS09a, KS09b] are of particular importance to us as well. Several theoretical results about stability or convergence are found in [Kne08]. In all these works the specific formulation of the FPE differs from our model. Nevertheless, there are essential parallels to our approach, namely the splitting into configuration and physical space, and the coupling to the Navier-Stokes equations.

In the review article [LOP11], different approaches for the Fokker-Planck equation are contrasted. In [FMAA20], fiber suspension flows through different axisymmetric geometries are examined. Setups with and without coupling of flow and fiber orientation are compared. Another current approach can be found in [ADS+24].

1.4 Outline

We started with a detailed literature review. Different possibilities to model mixtures like fiber suspensions, reaching from stochastic to deterministic and from Lagrangian to Eulerian approaches, were introduced. The focus of this work is on the treatment of the Fokker-Planck equation (FPE) as a deterministic Eulerian description.

The mathematical modeling of the chosen approach is discussed in **Chapter 2**. Our FPE describes the orientation distribution function (ODF) for fibers suspended in a fluid. Its definition is based on Jeffery's equation. The classical Jeffery equation describes the translational movement of a single fiber. A phenomenological diffusion

term must be added to take into account fiber-fiber interactions. Apart from this, the classical Jeffery equation can be extended to a conservation law. The resulting generalization can be interpreted as a simplified FPE. An attractive feature of this model is the existence of an analytical solution.

A simplification to the FPE and a traditional approach to simulate fiber suspensions is the Folgar-Tucker equation (FTE). The FTE can be derived from the FPE but it describes the orientation with tensors instead of an ODF. This leads to a significantly reduced computational time but also to a loss of accuracy.
An introduction to the Navier-Stokes equations (NSE) is included in Chapter 2, because only the coupled NSE-FPE system takes into account the fluid-fiber interaction. For this purpose, the NSE for Newtonian fluids needs to be replaced by a mixture model.

In **Chapter 3**, we deal with the finite element method (FEM), as it is the basis for all numerical approaches of this thesis, no matter which PDE is considered. The chapter summarizes existing knowledge with special emphasis on the aspects that are needed in the further course of the thesis. We apply the baseline FEM to a convection-diffusion equation that represents a prototype of the FPE. We also consider the properties of the finite element matrices, since these are important for the construction of property-preserving FEM studied in the following chapter. Finally, time stepping methods are addressed.

We continue with the introduction of physics-compatible FEM in **Chapter 4**. The use of bound-preserving approximations is mandatory if steep gradients arise in the calculation of numerical solutions. Otherwise, convective terms tend to produce inaccurate and unphysical results. This manifests itself in spurious oscillations up to completely wrong results. Moreover, the method may fail to preserve basic physical properties.
The Fokker-Planck equation has two convective terms, one with respect to the chosen geometry in space and one with respect to the unit sphere. Therefore 'limiting', which seeks to resolve the afore mentioned issues, is a central aspect of this work. An overview of the historical development and of some modern limiting techniques is given. Milestones in the development were, for instance, total variation diminishing (TVD) and flux corrected transport (FCT) schemes. Although these concepts provide central ideas, they have different disadvantages. While TVD schemes are primarily designed for 1d problems, FCT schemes inhibit convergence to steady-state solutions.
Thus, we focus on the class of algebraic flux correction (AFC) schemes, which do not have any of the aforementioned drawbacks. No heuristic parameters are needed. The modifications happen on the level of the FE matrices and guarantee the validity of discrete maximum principles. Last but not least, AFC schemes are basically applicable to any meshes. This is beneficial for our FPE solver, since we can consistently apply the same limiting method for the different subproblems. As a concrete realization of AFC, we chose the monolithic convex limiting (MCL) strategy. A virtue of this method are the built-in corrections for space discretization. Numerical studies complete the chapter. There selected limiting schemes are investigated for different time stepping schemes.

The space-independent part of our FPE is a PDE on a manifold, namely a convection-diffusion equation on the sphere. In the context of finite elements, the structure of the bilinear form for the corresponding PDE deserves a detailed discussion. Therefore, **Chapter 5** addresses basic concepts of differential geometry ranging from differential operators defined on manifolds to generalized integration formulas. A comprehensive overview of the mathematical background is combined with practically applicable formulas. In Section 5.6, numerical tests for different types of PDEs on the sphere are performed, incorporating limiting strategies whenever necessary.

Chapter 6 is dedicated to techniques for the fully coupled FPE-NSE system. Section 6.1 considers the operator splitting that is used to decompose the FPE-NSE system into simpler subproblems, which can be solved in parallel in a reasonable time. Already the previous chapters of the thesis build on the splitting of the FPE in a purely space-dependent and a space-independent part. This section explains in detail how to solve these subproblems in an alternating manner to obtain an approximate solution for the full FPE.
When the subproblems of the FPE are combined, new issues arise, even if the individual subproblems are suitably treated. Namely, the normalization property of the ODF may be lost. To prevent this, Section 6.2 extends the methodology of Chapter 4 to a new tailor-made limiting algorithm for the full FPE. In more advanced applications, the velocity field of the fiber suspension is no longer specified analytically, but it is calculated using the NSE. This means that the Jacobians of the velocity fields have to be reconstructed. An algorithm for that is given in Section 6.3. Last but not least, we present paradigms and summarize strategies to solve the NSE numerically in Section 6.4.

Chapter 7 finally brings together all the previous considerations and techniques. In Section 7.1, the 'Analytical Jeffery Benchmark' is examined. This benchmark still contains some simplifications and omits the NSE. At the same time, it allows to solve the full FPE and to compare it with an analytical reference solution.
Section 7.2 considers the 'Axisymmetric Contraction Benchmark', which is well established in the literature. The axisymmetric contraction geometry is well suited for validating rheological models since experimental data and numerical reference solutions are available. We use this example to demonstrate that fiber-induced stresses have a significant impact on the steady-state flow behavior of fiber suspensions. A comparison with results from the literature is used for validation purposes.

Summing up, several aspects have to be considered and different tools have to be combined to construct a numerical solver for the coupled system. All in all, a numerical simulation tool for the direct computation of a space- and time-dependent ODF was developed as a promising but costly alternative to empirical reconstructions. Parts of this thesis were published in [WKT24].

1.5 Abbreviations, symbols, notations

abbreviation	meaning
AFC	algebraic flux correction
CFD	computational fluid dynamics
DMP	discrete maximum principle
DOF	degrees of freedom
EOC	experimental order of convergence
FCT	flux corrected transport
FD(M)	finite difference (method)
FE(M)	finite element (method)
FPE	Fokker-Planck equation
FTE	Folgar-Tucker equation
FV(M)	finite volume (method)
IDP	invariant domain preserving
LED	local extremum diminishing
MCL	monolithic convex limiting
NSE	Navier-Stokes equations
ODE	ordinary differential equation
ODF	orientation distribution function
PDE	partial differential equation
PDF	probability density function
PSC	pressure Schur complement
RK	Runge-Kutta (scheme)
SDE	stochastic differential equation
SSP	strong stability preserving
SUPG	streamline upwind Petrov-Galerkin
TVD	total variation diminishing
b.c.	boundary condition
e.g.	for example (Latin: ‘exempli gratia’)
i.e.	that is (Latin: ‘id est’)
1d (2d, 3d)	one- (two-, three-)dimensional
err	error
min	minimum
max	maximum
inf	infimum
sup	supremum

symbol	meaning
$\mathbb{N}$	natural numbers
$\mathbb{R}$	real numbers
$\mathbb{N}_{>0}$	positive natural numbers
$\mathbb{R}_0^+$	non-negative real numbers
d	space dimension $d \in \{1, 2, 3\}$, usually $d = 3$
$\mathbf{e}_i$	i^{th} unit vector
$\mathbf{I}_d$	$d \times d$ identity matrix
$\operatorname{tr} A$	trace of square matrix A
$\mathcal{O}(\cdot)$	big O notation for terms depending on small parameters
Ω	spatial domain in $\mathbb{R}^d$
$\partial\Omega$	boundary of Ω
$\partial\Omega_D$	Dirichlet boundary part
$\partial\Omega_N$	Neumann boundary part
$\mathbb{S}^{d-1}$	surface of the unit sphere, i.e., $\{\mathbf{x} \in \mathbb{R}^d \mid \|\mathbf{x}\|_2 = 1\}$
M	manifold (in the context of surfaces; see also below)
τ	parametrization, $\tau : \Omega \to M$
$\mathbf{u}, \mathbf{v}$	velocity fields $\mathbf{u}, \mathbf{v} : \Omega \to \mathbb{R}^d$
$\mathbf{p}$	orientation vector, $\mathbf{p} \in \mathbb{S}^{d-1}$
t	time instant
T	end of a time interval
$\mathcal{T}_h$	triangulation
N	number of nodes in spatial meshes
M	number of nodes in the spherical mesh (see also above)
i	grid point index for a spatial mesh
k	grid points index for a spherical mesh
u	solution of a general PDE
ψ	solution of the FPE (probability density function)
φ	trial function (in the context of FEM)
ψ	test function (in the context of FEM)
$C^m(\Omega)$	continuously differentiable functions $v : \Omega \to \mathbb{R}$ such that all partial derivatives of degree up to m are continuous in Ω
$C_c(\Omega)$	continuous functions $v : \Omega \to \mathbb{R}$ with compact support
$L^p(\Omega)$	Lebesgue space, $p \in [1, \infty)$
$W^{k,p}(\Omega)$	Sobolev space, $p \in [1, \infty)$, $k \in \mathbb{N}_0$
$\|\cdot\|_2$	Euclidean norm
$\|\cdot\|_{L^p(\Omega)}$	norm on the space $L^p(\Omega)$
$\|\cdot\|_{W^{k,p}(\Omega)}$	norm on the space $W^{k,p}(\Omega)$

$\|\cdot\|_{H^1(\Omega)}$	norm on the space $H^1(\Omega) = W^{1,2}(\Omega)$
$\|\cdot\|_F$	Frobenius norm
$\mathcal{P}_r$	polynomial finite element space, see (3.8)
$\mathcal{Q}_r$	polynomial finite element space, see (3.9)
$\mathbf{P}_{1(2)}$	space of linear (quadratic) Lagrangian functions
$\mathbf{Q}_{1(2)}$	space of multilinear (multiquadratic) Lagrangian functions
∂	partial derivative
∂_p	partial derivative on the sphere
$\frac{\mathrm{d}}{\mathrm{d}t}$	material/substantial/convective derivative
$\dot{\mathbf{u}}$	temporal derivative of $\mathbf{u}$, $\dot{\mathbf{u}} = \mathrm{d}\mathbf{u}/\mathrm{d}t$
$\nabla, \nabla\cdot, \Delta$	(standard) gradient, divergence, and Laplace operator
$\nabla_{\mathbf{x}}, \nabla_{\mathbf{x}}\cdot, \Delta_{\mathbf{x}}$	spatial differential operators
$\nabla_{\mathbf{p}}, \nabla_{\mathbf{p}}\cdot, \Delta_{\mathbf{p}}$	spherical differential operators
$\nabla_M, \nabla_M\cdot, \Delta_M$	differential operators on manifold M
f_{ij}	internodal flux (in the context of limiting)
$\mathbf{x}_i$	nodal point with index i
ε_i	element patch containing $\mathbf{x}_i$
$\mathcal{N}_i$	nodal stencil
$\mathcal{N}^e$	element stencil
a	aspect ratio
C_I	interaction coefficient
D_r	rotational diffusion coefficient
λ_e	shape parameter
$\boldsymbol{\tau}$	stress tensor, $\boldsymbol{\tau} \in \mathbb{R}^{d\times d}$
N_p	particle number
N_s	shear number
ϕ	volume fraction
Re	Reynolds number

2 Mathematical modeling

In this thesis, we model and simulate the orientation of fibers in a suspension. Two models of orientation dynamics are the Fokker-Planck and the Folgar-Tucker equation. We use the former one to evolve the space- and time-dependent orientation distribution function (ODF). The computational effort is extremely high, but at the same time the ODF contains a large amount of information and yields accurate results.

In this chapter, we address derivation, different versions, physical meaning and mathematical properties of our four equations, that is, the Fokker-Planck equation itself, the Jeffery equation, the Folgar-Tucker equation and the Navier-Stokes equations. The Jeffery equation acts as a complementary equation for the Fokker-Planck equation in our application. To take into account that the fibers within the suspension align with the streamlines of the flow field and that, conversely, fiber-induced stresses influence the fluid flow, the coupled FPE-NSE system has to be considered.

2.1 Fokker-Planck equation

A Fokker-Planck equation can be set up whenever a phenomenon is described by a probability density. Hence, this partial differential equation has numerous applications in natural science. The wide range of its applicability is is not only reflected in its various versions but also in a variety of names. Our terminology dates back to the works of Fokker [Fok14] and Planck [Pla17]. In stochastics, the PDE is called Kolmogorov forward equation [MB05]. It is known as a diffusion equation in polymeric kinetic theory [LC03, HO06]. In the theory of Brownian motion, the names used are Klein-Kramers equation when the particles are exposed to an external field, or Smoluchowski equation when only position variables but no momentum variables are involved [Ött12]. Sometimes the FPE is also called drift-diffusion equation. This is explained by its generic formulation [Ris96, OP02, Keu04]

$$\frac{\partial}{\partial t}\psi(\mathbf{X},t) + \frac{\partial}{\partial \mathbf{X}}\{\mathbf{A}(\mathbf{X},t)\psi(\mathbf{X},t)\} = \frac{1}{2}\frac{\partial}{\partial \mathbf{X}}\frac{\partial}{\partial \mathbf{X}} : \{\mathbf{D}(\mathbf{X},t)\psi(\mathbf{X},t)\}, \tag{2.1}$$

where $\psi(\mathbf{X},t)$ is a probability distribution function (PDF) at time t. The vector $\mathbf{X}$ consists of N_C macroscopic variables, whose meaning ranges from space and velocity to orientation. The drift term $\mathbf{A}(\mathbf{X},t) \in \mathbb{R}^{N_C}$ is deterministic, whereas the symmetric positive definite diffusion tensor $\mathbf{D}(\mathbf{X},t) \in \mathbb{R}^{N_C \times N_C}$ is the stochastic contribution to the model. The colon operator ':' indicates the divergence of the divergence of a twice differentiable tensor [LOP11].
For our application to fiber suspensions, we have $\mathbf{X} = (\mathbf{x}, \mathbf{p})$, so that $\psi = \psi(\mathbf{x}, \mathbf{p}, t)$ is the main quantity of interest. The spatial variable $\mathbf{x} \in \mathbb{R}^d$, usually with $d = 3$,

and the time variable $t \in [0, T] \subset \mathbb{R}_0^+$ are the independent variables of many PDEs. Additionally, the configuration variable $\mathbf{p}$ is defined on the unit sphere

$$\mathbb{S}^{d-1} := \{\mathbf{q} \in \mathbb{R}^d \mid \|\mathbf{q}\|_2 = 1\}.$$

A vector $\mathbf{p} \in \mathbb{S}^{d-1}$ is a radially directed unit vector in the Euclidean space. It can also be uniquely identified by two orientation angles in the spherical coordinate system. Each point on the sphere represents a specific orientation. Thus, the function $\psi(\mathbf{x}, \mathbf{p}, t)$ describes the probability to find a fiber parallel to the orientation vector $\mathbf{p}$ at position $\mathbf{x}$ and time t. The specific Fokker-Planck equation reads

$$\frac{\partial \psi}{\partial t} + \nabla_{\mathbf{x}} \cdot (\mathbf{u}\psi) + \nabla_{\mathbf{p}} \cdot (\dot{\mathbf{p}}\psi) \;=\; \Delta_{\mathbf{p}}(D_r \psi) \quad \text{in } \Omega \times \mathbb{S}^2 \times (0, T]. \tag{2.2}$$

It has the structure of a time-dependent convection-diffusion equation. There are two convective terms, one with respect to space, the other with respect to orientation. The generic drift term from equation (2.1) can be found in equation (2.2) in the form of the velocity fields $\mathbf{u}$ and $\dot{\mathbf{p}}$. The fibers are assumed to be convected by the macroscopic velocity field $\mathbf{u}$, which can be specified analytically or calculated using the Navier-Stokes equations. Velocity field $\dot{\mathbf{p}}$ will be defined by the Jeffery equation as presented in Section 2.2.1.
The general formulation (2.1) of the Fokker-Planck equation (2.2) contains not only the configurational but also a spatial diffusion term. This spatial term plays a role for the analysis, but in the computational framework its effect is negligible, since the associated coefficient is typically in the order of 10^{-8} [Kne08]. This is the reason why we ignore it for the time being. If we assume that diffusion is isotropic, the remaining diffusion tensor $\mathbf{D}(\mathbf{x}, t)$ reduces to a positive scalar rotary diffusion coefficient D_r multiplied by the identity tensor. The resulting spherical Laplacian $\Delta_{\mathbf{p}}(D_r\psi)$ models the fiber-fiber interactions in non-dilute suspensions. Section 2.2.3 is devoted to this topic in detail. However, the influence of this diffusive term is comparatively small as well due to $D_r \ll 1$.

Since the PDF ψ describes the orientation dynamics of fibers, it is also called ODF here. As such it has to meet the following three requirements:

$$\psi(\mathbf{x}, \mathbf{p}, t) \;\geq\; 0 \qquad \forall\, \mathbf{x}, \mathbf{p}, t \qquad \text{(non-negativity)} \tag{2.3a}$$

$$\int_{\mathbb{S}^{d-1}} \psi(\mathbf{x}, \mathbf{p}, t)\, \mathrm{d}\mathbf{p} \;=\; 1 \qquad \forall\, \mathbf{x}, t \qquad \text{(normalization)} \tag{2.3b}$$

$$\psi(\mathbf{x}, \mathbf{p}, t) = \psi(\mathbf{x}, -\mathbf{p}, t) \qquad \forall\, \mathbf{x}, \mathbf{p}, t \qquad \text{(symmetry)} \tag{2.3c}$$

Non-negativity and normalization directly follow from ψ being a PDF, whereas symmetry is characteristic for our model, where we do not distinguish between front and back of the fibers. It might be worth mentioning that even though ψ describes a probability density, $\psi(\mathbf{x}, \mathbf{p}, t) \leq 1$ is not required.

When defining the orientation tensors or when introducing concepts of differential geometry later, we will stick to a general d. For the sake of brevity, however, and because the orientation is always defined as a point on the unit sphere $\mathbb{S}^2 \subset \mathbb{R}^3$, the following considerations are restricted to $d = 3$. In the three-dimensional space

the probability density $\psi = \psi(\mathbf{x}, \mathbf{p}, t) : \Omega \times \mathbb{S}^2 \times [0, T] \to \mathbb{R}_0^+$ is a function with respect to six dimensions, because $\Omega \subset \mathbb{R}^3$ and $\dim(\mathbb{S}^2) = 2$. The number of unknowns demonstrates the sheer computational effort we face when we solve the Fokker-Planck equation numerically. To handle the problem, we apply an alternating direction approach, which realizes a 'divide and conquer' strategy, see Section 6.1. The main point is that the full space-dependent Fokker-Planck equation (2.2) is replaced by two subproblems,

$$\text{the spatial advection equation} \qquad \frac{\partial \psi}{\partial t} + \nabla_{\mathbf{x}} \cdot (\mathbf{u}\psi) = 0 \tag{2.4a}$$

$$\text{and the space-independent FPE} \qquad \frac{\partial \psi}{\partial t} + \nabla_{\mathbf{p}} \cdot (\dot{\mathbf{p}}\psi) = \Delta_{\mathbf{p}}(D_r \psi). \tag{2.4b}$$

The studies of these two equations are the main topics of this thesis. The linear advection/convection/transport equation (2.4a) is a first-order hyperbolic equation in the physical space, whereas (2.4b) is a convection-diffusion equation on the sphere. Both can be interpreted as unsteady conservation laws of the form

$$\frac{\partial u}{\partial t} + \nabla \cdot \mathbf{f}(u) = 0,$$

where u is the scalar quantity to be conserved and $\mathbf{f} = (f_1, \ldots, f_d)$ is a given flux function. For a linear advection equation the flux is specified as $\mathbf{f} = \mathbf{v}u$ for a velocity field $\mathbf{v}$. Extended by a diffusive term with a coefficient ε, the flux reads $\mathbf{f} = \mathbf{v}u - \varepsilon \nabla u$. In the absence of reactive terms (sources and/or sinks), a conservation law states that no mass is produced or destroyed in the interior of the domain.

2.2 Jeffery equation

2.2.1 Classical Jeffery equation

The so-called Jeffery equation is a complementary equation to the Fokker-Planck equation and defines the rotation velocity $\dot{\mathbf{p}}$. Specifically, the quantity $\dot{\mathbf{p}}$ describes the change in orientation for a single fiber under the influence of the velocity gradients from the carrier fluid. Jeffery was the first author, who developed an expression for the motion of an ellipsoidal particle immersed in a fluid in the absence of external torque [Jef22]. Following [LDHB88], we formulate Jeffery's equation as

$$\dot{\mathbf{p}} = \mathbf{W}\mathbf{p} + \lambda_e \left[\mathbf{D}\mathbf{p} - (\mathbf{D} : (\mathbf{p} \otimes \mathbf{p}))\, \mathbf{p}\right] \tag{2.5a}$$

$$:= \mathbf{W}\mathbf{p} + \lambda_e \left[\mathbf{D}\mathbf{p} - \mathbf{D}\mathbf{p}\mathbf{p}\mathbf{p}\right] \tag{2.5b}$$

$$:= \boldsymbol{\kappa}\mathbf{p} - \boldsymbol{\kappa}\mathbf{p}\mathbf{p}\mathbf{p}, \quad \text{where } \boldsymbol{\kappa} := \mathbf{W} + \lambda_e \mathbf{D} \tag{2.5c}$$

The strain-rate/deformation tensor

$$\mathbf{D} = (d_{ij})_{i,j=1}^3 := \frac{1}{2}(\nabla \mathbf{u} + \nabla \mathbf{u}^T) \in \mathbb{R}^{3\times 3} \tag{2.6}$$

and the spin/vorticity tensor

$$\mathbf{W} = (\omega_{ij})_{i,j=1}^3 := \frac{1}{2}(\nabla \mathbf{u} - \nabla \mathbf{u}^T) \in \mathbb{R}^{3\times 3} \tag{2.7}$$

split the Jacobian $\nabla \mathbf{u}$ into a symmetric and a skew-symmetric part, that is,

$$\nabla \mathbf{u} = \mathbf{D} + \mathbf{W}, \quad \text{where } \mathbf{D} = \mathbf{D}^T \text{ and } \mathbf{W}^T = -\mathbf{W}.$$

It remains to define the multiplication operators that are used in (2.5a). For two vectors $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ the dyadic product is defined by

$$\mathbf{a} \otimes \mathbf{b} := \mathbf{a}\mathbf{b}^T \in \mathbb{R}^{n \times n}, \quad \text{that is,} \quad (\mathbf{a} \otimes \mathbf{b})_{ij} := a_i b_j,$$

whereas the tensor contraction/double-dot product for two tensors $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$ is defined by the sum of the corresponding components, that is,

$$\mathbf{A} : \mathbf{B} := \operatorname{tr}(\mathbf{A}^T \mathbf{B}) = \sum_{ij} a_{ij} b_{ij}.$$

The Jeffery equation given by (2.5b) and (2.5c) uses a compressed notation, which can also be found in [DA84, LDHB88, PT09]. The final formulation (2.5c) requires that $-\mathbf{W}\mathbf{p}\mathbf{p}\mathbf{p} = 0$, which will follow immediately from Lemma 2.1 below and the properties of tensor $\mathbf{W}$.
A fiber is modeled by an ellipsoidal rigid particle with length l and diameter d. Using the aspect ratio $a := l/d \in [1, \infty)$, the shape parameter λ_e is defined by

$$\lambda_e := \frac{a^2 - 1}{a^2 + 1} \in [0, 1).$$

For $l = d$, which is the limit of a sphere, we obtain $\lambda_e = 0$, whereas for $l \gg d$ the shape factor λ_e converges to 1. A typical choice for our application is $a = 10$, so that $\lambda_e = 99/101$. For an infinite aspect ratio, the tensor $\boldsymbol{\kappa} := \mathbf{W} + \lambda_e \mathbf{D}$ would be equal to $\nabla \mathbf{u}$.

The following relationship is a helpful tool, which can be applied to $\mathbf{D}$ and $\mathbf{W}$:

Lemma 2.1 (Calculation rule)**.** *For a tensor $\mathbf{T} \in \mathbb{R}^{n \times n}$ and $\mathbf{p}, \mathbf{q} \in \mathbb{R}^n$, we have*

$$\mathbf{T} : (\mathbf{p} \otimes \mathbf{q}) = \langle \mathbf{p}, \mathbf{T}\mathbf{q} \rangle.$$

Proof. $\mathbf{T} : (\mathbf{p} \otimes \mathbf{q}) = \sum_{ij} t_{ij} p_i q_j = \sum_i p_i \sum_j t_{ij} q_j = \langle \mathbf{p}, \mathbf{T}\mathbf{q} \rangle$. □

Consequently, the symmetric strain-rate tensor $\mathbf{D}$ satisfies

$$\mathbf{D} : (\mathbf{p} \otimes \mathbf{q}) = \langle \mathbf{p}, \mathbf{D}\mathbf{q} \rangle = \langle \mathbf{q}, \mathbf{D}\mathbf{p} \rangle, \tag{2.8a}$$
$$\mathbf{D} : (\mathbf{p} \otimes \mathbf{p}) = \langle \mathbf{p}, \mathbf{D}\mathbf{p} \rangle. \tag{2.8b}$$

For the skew-symmetric vorticity tensor $\mathbf{W}$, on the other hand, we obtain

$$\mathbf{W} : (\mathbf{p} \otimes \mathbf{q}) = \langle \mathbf{p}, \mathbf{W}\mathbf{q} \rangle = -\langle \mathbf{q}, \mathbf{W}\mathbf{p} \rangle, \tag{2.8c}$$
$$\mathbf{W} : (\mathbf{p} \otimes \mathbf{p}) = \langle \mathbf{p}, \mathbf{W}\mathbf{p} \rangle = 0. \tag{2.8d}$$

The last relation holds since $\langle \mathbf{p}, \mathbf{W}\mathbf{p} \rangle = \langle \mathbf{W}\mathbf{p}, \mathbf{p} \rangle = \langle \mathbf{p}, \mathbf{W}^T \mathbf{p} \rangle = -\langle \mathbf{p}, \mathbf{W}\mathbf{p} \rangle$ due to the skew-symmetry of $\mathbf{W}$.

Now we can also show that $\dot{\mathbf{p}}$ is perpendicular to $\mathbf{p}$, i.e., $\langle \mathbf{p}, \dot{\mathbf{p}} \rangle = 0$. This property is important later, when $\dot{\mathbf{p}}$ is used as a velocity field for the convection on the sphere and therefore has to be normal to it.

Lemma 2.2 (Orthogonality)**.** *The vector* $\dot{\mathbf{p}}$ *given by the Jeffery equation* (2.5) *is perpendicular to* $\mathbf{p}$*, that is,* $\langle \mathbf{p}, \dot{\mathbf{p}} \rangle = 0$*.*

Proof. Considering

$$\dot{\mathbf{p}} = \underbrace{\mathbf{W}\mathbf{p}}_{(\star)} + \lambda_e \underbrace{[\mathbf{D}\mathbf{p} - (\mathbf{D} : (\mathbf{p} \otimes \mathbf{p}))\mathbf{p}]}_{(\star\star)}$$

the orthogonality even holds for each summand. The validity of $\langle \mathbf{p}, (\star) \rangle = 0$ is implied by (2.8d). To check that $\langle \mathbf{p}, (\star\star) \rangle = 0$, we use $\|\mathbf{p}\| = 1$ and consequently $\langle \mathbf{p}, \mathbf{p} \rangle = \|\mathbf{p}\|^2 = 1$. Applying equation (2.8b) as well, this yields

$$\begin{aligned} \langle \mathbf{p}, (\star\star) \rangle &= \langle \mathbf{p}, \mathbf{D}\mathbf{p} \rangle - \langle \mathbf{p}, \underbrace{(\mathbf{D} : (\mathbf{p} \otimes \mathbf{p}))}_{\in \mathbb{R}} \mathbf{p} \rangle \\ &= \langle \mathbf{p}, \mathbf{D}\mathbf{p} \rangle - (\mathbf{D} : (\mathbf{p} \otimes \mathbf{p})) \underbrace{\langle \mathbf{p}, \mathbf{p} \rangle}_{=1} \quad = \; 0, \end{aligned}$$

which is the desired result. □

Finally, let us address a solution formula for Jeffery's equation.

Lemma 2.3 (Solution of Jeffery's equation)**.** *An analytical solution of Jeffery's equation* (2.5) *is given by*

$$\mathbf{p} = \frac{\mathbf{q}}{\|\mathbf{q}\|}, \quad \textit{where } \mathbf{q} \textit{ is a solution of the ODE system } \dot{\mathbf{q}} = \boldsymbol{\kappa}\mathbf{q}.$$

Proof. Using the results of our preliminary work, the formula can be proven by simple substitution:

$$\begin{aligned} \dot{\mathbf{p}} \quad &= \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\mathbf{q}}{\|\mathbf{q}\|}\right) = \frac{\dot{\mathbf{q}}\|\mathbf{q}\| - \mathbf{q}\frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{q}\|}{\|\mathbf{q}\|^2} && \left| \; \frac{\mathrm{d}}{\mathrm{d}t}\|\mathbf{q}\| = \frac{\mathbf{q}\cdot\dot{\mathbf{q}}}{\|\mathbf{q}\|^2} \right. \\ &= \frac{\dot{\mathbf{q}}}{\|\mathbf{q}\|} - \frac{\mathbf{q}}{\|\mathbf{q}\|^2}\frac{\mathbf{q}\cdot\dot{\mathbf{q}}}{\|\mathbf{q}\|} && \left| \; \dot{\mathbf{q}} = \boldsymbol{\kappa}\mathbf{q} \right. \\ &= \frac{\boldsymbol{\kappa}\mathbf{q}}{\|\mathbf{q}\|} - \frac{\mathbf{q}}{\|\mathbf{q}\|}\left\langle \frac{\mathbf{q}}{\|\mathbf{q}\|}, \boldsymbol{\kappa}\frac{\mathbf{q}}{\|\mathbf{q}\|} \right\rangle && \left| \; \mathbf{p} = \frac{\mathbf{q}}{\|\mathbf{q}\|} \right. \\ &= \boldsymbol{\kappa}\mathbf{p} - \boldsymbol{\kappa}\mathbf{p}\mathbf{p}\mathbf{p} \end{aligned}$$

Thus, we obtain (2.5c), which finishes the proof. □

2.2.2 Extended Jeffery equation

In the previous section, we examined the Jeffery equation as an equation that describes the temporal change of the orientation vector $\mathbf{p}$. Another equation that is often called Jeffery equation in the literature reads

$$\frac{\mathrm{d}\psi}{\mathrm{d}t} + \nabla_{\mathbf{p}} \cdot (\dot{\mathbf{p}}\psi) = 0. \tag{2.9}$$

It can be considered as an extension of the original Jeffery equation (2.5) to a conservation law for ψ. We interpret the temporal derivative as the so-called material, substantial or convective derivative, that is,

$$\frac{\mathrm{d}(\cdot)}{\mathrm{d}t} = \frac{\partial(\cdot)}{\partial t} + (\mathbf{u} \cdot \nabla_{\mathbf{x}})(\cdot). \tag{2.10}$$

Then, the extended Jeffery equation (2.9) equals the full space-dependent FPE with $D_r = 0$. Setting $\mathbf{u} = 0$, we obtain the space-*in*dependent FPE. The formulation with the material derivative is also referred to as Eulerian form, whereas the version with the partial derivative is called Lagrangian form. Overall, the usefulness of the extended Jeffery equation in the Lagrangian form lies in the possibility of obtaining an analytical solution for fixed tensors $\mathbf{D}$ and $\mathbf{W}$. It is used as a reference solution several times in this thesis.

In the following Theorem 2.4, an analytical solution is given for the space-independent FPE with $D_r = 0$. The theorem is proven afterwards and certain properties of the solution are explored.

Theorem 2.4 (Analytical solution of the extended Jeffery equation)**.** *For the initial value problem*

$$\begin{cases} \dfrac{\partial \psi}{\partial t} + \nabla_{\mathbf{p}} \cdot (\dot{\mathbf{p}}\psi) = 0 \quad \text{on } \mathbb{S}^{d-1} \times (0, T], & \text{(2.11a)} \\ \psi|_{t=0} = \psi_0 \quad \text{on } \mathbb{S}^{d-1}, & \text{(2.11b)} \end{cases}$$

and $d \in \{2, 3\}$, an exact solution is given by [DA84, MSHJS11, Kuz18]

$$\begin{cases} \psi(\mathbf{p}) = \psi_0 \left(\dfrac{\mathbf{C}\mathbf{p}}{\|\mathbf{C}\mathbf{p}\|} \right) \dfrac{1}{\|\mathbf{C}\mathbf{p}\|^d}, & \text{(2.12a)} \\ \dfrac{\partial \mathbf{C}}{\partial t} = -\mathbf{C} \cdot (\mathbf{W} + \lambda_e \mathbf{D}), \ \mathbf{C}|_{t=0} = \mathbf{I}_d, & \text{(2.12b)} \end{cases}$$

where $\mathbf{I}_d$ is the identity tensor.

With additional assumptions the exact solution can be simplified.

Corollary 2.5 (Special cases of Theorem 2.4)**.** *In the case of an isotropic/random initial orientation distribution, that is, for*

$$\psi_0(\mathbf{p}) = \frac{1}{2(d-1)\pi}, \quad d \in \{2, 3\}, \tag{2.13}$$

the analytical solution (2.12a) can be rewritten as

$$\psi(\mathbf{p}) = \frac{1}{2(d-1)\pi} \frac{1}{\|\mathbf{C}\mathbf{p}\|^d} \tag{2.14}$$

If the velocity field $\nabla \mathbf{u}$, the key ingredient to define the tensors $\mathbf{W}$ and $\mathbf{D}$, is constant, formula (2.12b) simplifies to

$$\mathbf{C}(t) = \exp(-t(\mathbf{W} + \lambda_e \mathbf{D})) = \exp(-t\,\boldsymbol{\kappa}). \tag{2.15}$$

In the important 3d case with an isotropic initial distribution $\psi_0 = \frac{1}{4\pi}$, the analytical solution reads

$$\psi(\mathbf{p}) = \frac{1}{4\pi \|\mathbf{C}\mathbf{p}\|^3}.$$

We give a proof of Theorem 2.4 for the special case of the isotropic initial distribution in 3d. The basic idea follows a very similar proof given in [DA84], although we modify it and add details. An alternative approach can be found in [MSHJS11]. Let us insert a helpful lemma in advance.

Lemma 2.6. *For the basis vectors*

$$\mathbf{e}_r = \begin{pmatrix} \sin\theta\cos\varphi \\ \sin\theta\sin\varphi \\ \cos\theta \end{pmatrix}, \quad \mathbf{e}_\varphi = \begin{pmatrix} -\sin\varphi \\ \cos\varphi \\ 0 \end{pmatrix}, \quad \mathbf{e}_\theta = \begin{pmatrix} \cos\theta\cos\varphi \\ \cos\theta\sin\varphi \\ -\sin\theta \end{pmatrix},$$

see also Appendix 10.1, and for the tensors $\mathbf{D}$ and $\mathbf{W}$ it holds true that

a) $\mathbf{e}_r \otimes \mathbf{e}_r + \mathbf{e}_\varphi \otimes \mathbf{e}_\varphi + \mathbf{e}_\theta \otimes \mathbf{e}_\theta = \mathbf{I}_3$,

b) $(\mathbf{e}_k \otimes \mathbf{e}_k)\mathbf{D}\mathbf{e}_r = \mathbf{D}\mathbf{e}_r\mathbf{e}_k\mathbf{e}_k$ *and* $(\mathbf{e}_k \otimes \mathbf{e}_k)\mathbf{W}\mathbf{e}_r = -\mathbf{W}\mathbf{e}_r\mathbf{e}_k\mathbf{e}_k$, $k \in \{r, \varphi, \theta\}$.

Proof. a) The relationship can be verified by using concrete spherical coordinates and applying the Pythagorean trigonometric identity $\sin^2\alpha + \cos^2\alpha = 1$. Combined within a sum the three individual dyadic products give the unit matrix.

b) We define $\mathbf{M} := \mathbf{e}_k \otimes \mathbf{e}_k$, $k \in \{r, \varphi, \theta\}$, and $\mathbf{G} := \mathbf{M}\mathbf{T}$, $\mathbf{T} \in \{\mathbf{D}, \mathbf{W}\}$. Then

$$\begin{aligned}((\mathbf{e}_k \otimes \mathbf{e}_k)\mathbf{T}\mathbf{e}_r)_i &= \sum_j g_{ij}(\mathbf{e}_r)_j = \sum_{j,l} m_{il}t_{lj}(\mathbf{e}_r)_j = \sum_{j,l}(\mathbf{e}_k)_i(\mathbf{e}_k)_l t_{lj}(\mathbf{e}_r)_j \\ &= \sum_{j,l} t_{lj}(\mathbf{e}_r)_j(\mathbf{e}_k)_l(\mathbf{e}_k)_i = ((\mathbf{T}^T : (\mathbf{e}_r \otimes \mathbf{e}_k))\mathbf{e}_k)_i\end{aligned}$$

Since $\mathbf{D}^T = \mathbf{D}$ and $\mathbf{W}^T = -\mathbf{W}$, this yields the desired result.

□

Proof. (of Theorem 2.4) Starting with a suitable expression for $\dot{\mathbf{p}}$, we derive an expression for $\nabla_\mathbf{p} \cdot (\dot{\mathbf{p}}\psi)$ with unspecified ψ. Finally, substituting the analytical solution ψ, we show that it satisfies the extended Jeffery equation $\frac{\partial\psi}{\partial t} + \nabla_\mathbf{p} \cdot (\dot{\mathbf{p}}\psi) = 0$. In what follows, we often use the fact that $\mathbf{p} = \mathbf{e}_r$, i.e., that the orientation vector $\mathbf{p} \in \mathbb{S}^2$ and the radially directed unit vector $\mathbf{e}_r$ are equivalent.

We start with Jeffery's equation (2.5c) and insert the identity tensor $\mathbf{I}_3$ as defined in Lemma 2.6a). Directly afterwards, Lemma 2.6b) is used. Finally, we exploit the fact that $\mathbf{W}\mathbf{e}_r\mathbf{e}_r\mathbf{e}_r = 0$ by Lemma 2.1. All in all, we find that

$$\begin{aligned}\dot{\mathbf{p}} &= \boldsymbol{\kappa}\mathbf{p} - \boldsymbol{\kappa}\mathbf{p}\mathbf{p}\mathbf{p} = \boldsymbol{\kappa}\mathbf{e}_r - \boldsymbol{\kappa}\mathbf{e}_r\mathbf{e}_r\mathbf{e}_r \\ &= [\mathbf{e}_r \otimes \mathbf{e}_r + \mathbf{e}_\varphi \otimes \mathbf{e}_\varphi + \mathbf{e}_\theta \otimes \mathbf{e}_\theta]\,\boldsymbol{\kappa}\mathbf{e}_r - \boldsymbol{\kappa}\mathbf{e}_r\mathbf{e}_r\mathbf{e}_r \\ &= \; -\mathbf{W}\mathbf{e}_r\mathbf{e}_r\mathbf{e}_r \; - \mathbf{W}\mathbf{e}_r\mathbf{e}_\varphi\mathbf{e}_\varphi \; - \mathbf{W}\mathbf{e}_r\mathbf{e}_\theta\mathbf{e}_\theta \; - \mathbf{W}\mathbf{e}_r\mathbf{e}_r\mathbf{e}_r\end{aligned}$$

$$
\begin{aligned}
&+ \lambda_e \mathbf{D}\mathbf{e}_r\mathbf{e}_r\mathbf{e}_r + \lambda_e \mathbf{D}\mathbf{e}_r\mathbf{e}_\varphi\mathbf{e}_\varphi + \lambda_e \mathbf{D}\mathbf{e}_r\mathbf{e}_\theta\mathbf{e}_\theta - \lambda_e \mathbf{D}\mathbf{e}_r\mathbf{e}_r\mathbf{e}_r \\
&= \widetilde{\boldsymbol{\kappa}}\mathbf{e}_r\mathbf{e}_\varphi\mathbf{e}_\varphi + \widetilde{\boldsymbol{\kappa}}\mathbf{e}_r\mathbf{e}_\theta\mathbf{e}_\theta, \quad \text{where } \widetilde{\boldsymbol{\kappa}} := -\mathbf{W} + \lambda_e \mathbf{D}. \qquad (2.16)
\end{aligned}
$$

The use of the product rule reveals that

$$
\nabla_{\mathbf{p}} \cdot (\dot{\mathbf{p}}\psi) = \nabla_{\mathbf{p}}\psi \cdot \dot{\mathbf{p}} + (\nabla_{\mathbf{p}} \cdot \dot{\mathbf{p}})\psi. \qquad (2.17)
$$

The formula for the spherical differential operator $\nabla_{\mathbf{p}}$ can be found in the literature, but we also derive it below, see (5.14). The spherical coordinates representation of $\nabla_{\mathbf{p}}$ reads

$$
\nabla_{\mathbf{p}} := \mathbf{e}_\theta \frac{\partial}{\partial\theta} + \mathbf{e}_\varphi \frac{1}{\sin\theta}\frac{\partial}{\partial\varphi}.
$$

Applying this operator to the different terms in (2.17), some of them are considerably simplified, when we take into account that $\mathbf{e}_r, \mathbf{e}_\varphi$ and $\mathbf{e}_\theta$ are orthonormal. On the one hand, applying $\nabla_{\mathbf{p}}$ to ψ and forming the dot product with $\dot{\mathbf{p}}$ as given by (2.16), we deduce that

$$
\begin{aligned}
\nabla_{\mathbf{p}}\psi \cdot \dot{\mathbf{p}} &= \left(\mathbf{e}_\theta \frac{\partial\psi}{\partial\theta} + \mathbf{e}_\varphi \frac{1}{\sin\theta}\frac{\partial\psi}{\partial\varphi}\right) \cdot (\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\varphi\mathbf{e}_\varphi + \widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\theta\mathbf{e}_\theta) \\
&= (\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\theta)\frac{\partial\psi}{\partial\theta} + \frac{1}{\sin\theta}(\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\varphi)\frac{\partial\psi}{\partial\varphi}.
\end{aligned}
$$

On the other hand, the divergence $\nabla_{\mathbf{p}} \cdot \dot{\mathbf{p}}$ can be calculated by computing the dot product of the gradient operator $\nabla_{\mathbf{p}}$ and $\dot{\mathbf{p}}$. For this purpose, the basis vectors have to be differentiated with respect to φ and θ. A collocation of the basis vectors and their derivatives can be found in Appendix 10.1. The individual terms of the dot product are

$$
\begin{aligned}
\mathbf{e}_\theta\left[\frac{\partial}{\partial\theta}(\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\varphi\mathbf{e}_\varphi)\right] &= \mathbf{e}_\theta\left[\frac{\partial}{\partial\theta}(\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\varphi)\mathbf{e}_\varphi + (\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\varphi)\underbrace{\frac{\partial\mathbf{e}_\varphi}{\partial\theta}}_{=0}\right] = 0, \\
\mathbf{e}_\theta\left[\frac{\partial}{\partial\theta}(\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\theta\mathbf{e}_\theta)\right] &= \mathbf{e}_\theta\left[\frac{\partial}{\partial\theta}(\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\theta)\mathbf{e}_\theta + (\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\theta)\underbrace{\frac{\partial\mathbf{e}_\theta}{\partial\theta}}_{=-\mathbf{e}_r}\right] = \frac{\partial}{\partial\theta}(\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\theta), \\
\frac{1}{\sin\theta}\mathbf{e}_\varphi\left[\frac{\partial}{\partial\varphi}(\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\varphi\mathbf{e}_\varphi)\right] &= \frac{1}{\sin\theta}\mathbf{e}_\varphi\left[\frac{\partial}{\partial\varphi}(\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\varphi)\mathbf{e}_\varphi + (\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\varphi)\underbrace{\frac{\partial\mathbf{e}_\varphi}{\partial\varphi}}_{=-\sin\theta\,\mathbf{e}_r-\cos\theta\,\mathbf{e}_\theta}\right] \\
&= \frac{1}{\sin\theta}\frac{\partial}{\partial\varphi}(\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\varphi), \\
\frac{1}{\sin\theta}\mathbf{e}_\varphi\left[\frac{\partial}{\partial\varphi}(\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\theta\mathbf{e}_\theta)\right] &= \frac{1}{\sin\theta}\mathbf{e}_\varphi\left[\frac{\partial}{\partial\varphi}(\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\theta)\mathbf{e}_\theta + (\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\theta)\underbrace{\frac{\partial\mathbf{e}_\theta}{\partial\varphi}}_{=\cos\theta\,\mathbf{e}_\varphi}\right] \\
&= \frac{\cos\theta}{\sin\theta}(\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\theta).
\end{aligned}
$$

Using Lemma 2.1, the product rule with respect to the dot product and the differentiation rules for the basis vectors one more time, see Appendix 10.1, we find that

$$\frac{\partial}{\partial\theta}(\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\theta) = \frac{\partial}{\partial\theta}\langle \mathbf{e}_r, \widetilde{\boldsymbol{\kappa}}\mathbf{e}_\theta\rangle = \Big\langle \underbrace{\frac{\partial \mathbf{e}_r}{\partial\theta}}_{=\mathbf{e}_\theta}, \widetilde{\boldsymbol{\kappa}}\mathbf{e}_\theta \Big\rangle + \Big\langle \mathbf{e}_r, \widetilde{\boldsymbol{\kappa}} \underbrace{\frac{\partial \mathbf{e}_\theta}{\partial\theta}}_{=-\mathbf{e}_r} \Big\rangle = \widetilde{\boldsymbol{\kappa}} : \mathbf{e}_\theta\mathbf{e}_\theta - \widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_r$$

and analogously

$$\begin{aligned}\frac{1}{\sin\theta}\frac{\partial}{\partial\varphi}(\widetilde{\boldsymbol{\kappa}}\mathbf{e}_r\mathbf{e}_\varphi) = \frac{1}{\sin\theta}\langle \mathbf{e}_r, \widetilde{\boldsymbol{\kappa}}\mathbf{e}_\varphi\rangle &= \frac{1}{\sin\theta}\Big(\Big\langle \underbrace{\frac{\partial \mathbf{e}_r}{\partial\varphi}}_{=\sin\theta\mathbf{e}_\varphi}, \widetilde{\boldsymbol{\kappa}}\mathbf{e}_\varphi\Big\rangle + \Big\langle \mathbf{e}_r, \widetilde{\boldsymbol{\kappa}} \underbrace{\frac{\partial \mathbf{e}_\varphi}{\partial\varphi}}_{=-\sin\theta\mathbf{e}_r-\cos\theta\mathbf{e}_\theta} \Big\rangle\Big)\\ &= \widetilde{\boldsymbol{\kappa}} : \mathbf{e}_\varphi\mathbf{e}_\varphi \;-\; \widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_r - \frac{\cos\theta}{\sin\theta}\,\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\theta.\end{aligned}$$

Combining the previous results, and taking into account that two terms cancel out, we obtain

$$\begin{aligned}\nabla_{\mathbf{p}} \cdot (\dot{\mathbf{p}}\psi) = (\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\theta)\frac{\partial\psi}{\partial\theta} + \frac{1}{\sin\theta}(\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\varphi)\frac{\partial\psi}{\partial\varphi}\\ - 2(\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_r)\psi + (\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_\theta\mathbf{e}_\theta)\psi + (\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_\varphi\mathbf{e}_\varphi)\psi.\end{aligned}$$

The second line of the above expression can be further simplified to

$$-3(\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_r)\psi + \underbrace{(\widetilde{\boldsymbol{\kappa}} : \mathbf{I}_3)}_{=0}\psi = -3(\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_r)\psi,$$

where we used Lemma 2.6a) as well as the assumption that the fluid is incompressible, that is, $\nabla \cdot \mathbf{u} = 0$, and consequently $\operatorname{tr}(\nabla\mathbf{u}) = 0$. Moreover,

$$\widetilde{\boldsymbol{\kappa}} : \mathbf{I}_3 = \operatorname{tr}(-\mathbf{W}) + \lambda_e \operatorname{tr}(\mathbf{D}) = 0.$$

Summing up,

$$\nabla_{\mathbf{p}} \cdot (\dot{\mathbf{p}}\psi) \;=\; -3(\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_r)\psi + (\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\theta)\frac{\partial\psi}{\partial\theta} + \frac{1}{\sin\theta}(\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_\varphi)\frac{\partial\psi}{\partial\varphi}. \tag{2.18}$$

With all this preliminary work, we can finally show that $\frac{\partial\psi}{\partial t} + \nabla_{\mathbf{p}} \cdot (\dot{\mathbf{p}}\psi) = 0$ holds for ψ defined by (2.12). For this purpose we calculate $\frac{\partial\psi}{\partial\theta}$, $\frac{\partial\psi}{\partial\varphi}$, and $\frac{\partial\psi}{\partial t}$. As announced we stick to the 3d case with isotropic initial distribution. Thus,

$$\psi(\mathbf{p}) \overset{(2.12a)}{=} \frac{1}{4\pi\|\mathbf{C}\mathbf{p}\|^3} = \frac{1}{4\pi}\langle \mathbf{C}\mathbf{p}, \mathbf{C}\mathbf{p}\rangle^{-\frac{3}{2}}.$$

As 'inner derivatives', we obtain

$$\begin{aligned}\frac{\partial}{\partial\theta}\langle \mathbf{C}\mathbf{p}, \mathbf{C}\mathbf{p}\rangle \;&=\; \Big\langle \mathbf{C}\frac{\partial \mathbf{e}_r}{\partial\theta}, \mathbf{C}\mathbf{e}_r\Big\rangle + \Big\langle \mathbf{C}\mathbf{e}_r, \mathbf{C}\frac{\partial \mathbf{e}_r}{\partial\theta}\Big\rangle = 2\Big\langle \mathbf{C}\mathbf{p}, \mathbf{C}\frac{\partial \mathbf{e}_r}{\partial\theta}\Big\rangle \;=\; 2\langle \mathbf{C}^T\mathbf{C}\mathbf{p}, \mathbf{e}_\theta\rangle,\\ \frac{\partial}{\partial\varphi}\langle \mathbf{C}\mathbf{p}, \mathbf{C}\mathbf{p}\rangle \;&=\; 2\Big\langle \mathbf{C}\mathbf{p}, \mathbf{C}\frac{\partial \mathbf{e}_r}{\partial\varphi}\Big\rangle \;=\; 2\sin\theta\langle \mathbf{C}^T\mathbf{C}\mathbf{p}, \mathbf{e}_\varphi\rangle,\end{aligned}$$

$$\frac{\partial}{\partial t}\langle \mathbf{Cp}, \mathbf{Cp}\rangle \;=\; 2\Big\langle \mathbf{Cp}, \frac{\partial}{\partial t}\mathbf{Cp}\Big\rangle \;=\; 2\,\langle \mathbf{Cp}, -\mathbf{C}\boldsymbol{\kappa}\mathbf{p}\rangle \;=\; -2\langle \mathbf{C}^T\mathbf{C}, \boldsymbol{\kappa}\mathbf{p}\rangle.$$

In particular, for the temporal derivative, the equation (2.12b) for $\mathbf{C}$ was used. Since the partial temporal derivative is applied, we do not obtain $\dot{\mathbf{p}}$, but $\frac{\partial \mathbf{p}}{\partial t}$, which equals zero. Combined with the 'exterior derivative'

$$\psi'(\mathbf{p}) = \frac{-3}{2\cdot 4\pi}\langle \mathbf{Cp}, \mathbf{Cp}\rangle^{-\frac{5}{2}} = \frac{-3\psi}{2\|\mathbf{Cp}\|^2},$$

we end up with

$$\frac{\partial \psi}{\partial \theta} = \frac{-3\,\langle \mathbf{C}^T\mathbf{Cp}, \mathbf{e}_\theta\rangle\,\psi}{\|\mathbf{Cp}\|^2}, \qquad \frac{1}{\sin\theta}\frac{\partial \psi}{\partial \varphi} = \frac{-3\,\langle \mathbf{C}^T\mathbf{Cp}, \mathbf{e}_\varphi\rangle\,\psi}{\|\mathbf{Cp}\|^2},$$
$$\text{and} \qquad \frac{\partial \psi}{\partial t} = \frac{3\langle \mathbf{C}^T\mathbf{C}, \boldsymbol{\kappa}\mathbf{p}\rangle}{\|\mathbf{Cp}\|^2}.$$

Substitution of $\frac{\partial \psi}{\partial \theta}$ and $\frac{\partial \psi}{\partial \varphi}$ into (2.18) yields

$$\begin{aligned}\nabla\cdot(\dot{\mathbf{p}}\psi) \;=\; &\frac{-3\psi}{\|\mathbf{Cp}\|^2}\Big((\boldsymbol{\kappa} : \mathbf{e}_r\mathbf{e}_r)\langle \mathbf{C}^T\mathbf{Cp}, \mathbf{Ce}_r\rangle \\ &+ (\boldsymbol{\kappa} : \mathbf{e}_r\mathbf{e}_\theta)\langle \mathbf{C}^T\mathbf{Cp}, \mathbf{e}_\theta\rangle + (\boldsymbol{\kappa} : \mathbf{e}_r\mathbf{e}_\varphi)\langle \mathbf{C}^T\mathbf{Cp}, \mathbf{e}_\varphi\rangle\Big).\end{aligned}$$

On the other hand,

$$\frac{\partial \psi}{\partial t} \;=\; \frac{3\psi}{\|\mathbf{Cp}\|^2}\,\big\langle \mathbf{C}^T\mathbf{Cp}, \boldsymbol{\kappa}\mathbf{p}\big\rangle.$$

Finally, it must be verified that

$$\begin{aligned}\frac{\partial \psi}{\partial t} \;&=\; -\nabla_{\mathbf{p}}\cdot(\dot{\mathbf{p}}\psi) \\ \iff \qquad \big\langle \mathbf{C}^T\mathbf{Cp}, \boldsymbol{\kappa}\mathbf{p}\big\rangle &= \sum_{k\in\{r,\theta,\varphi\}} (\boldsymbol{\kappa} : \mathbf{e}_r\mathbf{e}_k)\big\langle \mathbf{C}^T\mathbf{Cp}, \mathbf{e}_k\big\rangle.\end{aligned}$$

We complete the proof in the same way we started it, that is, we insert the unit matrix from Lemma 2.6a) in the above left hand side. Then

$$\begin{aligned}\big\langle \mathbf{C}^T\mathbf{Cp}, \boldsymbol{\kappa}\mathbf{p}\big\rangle &= \big\langle \mathbf{C}^T\mathbf{Cp},\ [\mathbf{e}_r\otimes\mathbf{e}_r + \mathbf{e}_\varphi\otimes\mathbf{e}_\varphi + \mathbf{e}_\theta\otimes\mathbf{e}_\theta]\,\boldsymbol{\kappa}\mathbf{p}\big\rangle \\ &= \Big\langle \mathbf{C}^T\mathbf{Cp},\ \sum_{k\in\{r,\theta,\varphi\}} (\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_k)\mathbf{e}_k\Big\rangle \quad = \sum_{k\in\{r,\theta,\varphi\}} (\widetilde{\boldsymbol{\kappa}} : \mathbf{e}_r\mathbf{e}_k)\,\big\langle \mathbf{C}^T\mathbf{Cp},\ \mathbf{e}_k\big\rangle\end{aligned}$$

This is equal to the right hand side, which finishes the proof. □

We conclude this section with some remarks regarding the exact solution.

Remarks 2.7.

a) (Matrix exponential.) In formula (2.15) 'exp' is not the common scalar exponential function but a matrix valued generalization. Like its scalar counterpart, the exponential of a matrix can be defined by a power series:

$$\exp(\mathbf{A}) := \sum_{k=0}^{\infty} \frac{\mathbf{A}^k}{k!} = \mathbf{I} + \mathbf{A} + \frac{\mathbf{A}^2}{2!} + \frac{\mathbf{A}^3}{3!} + \dots$$

A suggestion how to approximate this matrix numerically can be found in [Hig05].

b) (Normalization property.) The analytical solution of the extended Jeffery equation possesses the basic properties (2.3) of the ODF ψ, including the normalization property. For example, for 3d the analytical solution (2.14) satisfies

$$\int_{\mathbb{S}^2} \psi(\mathbf{p}, 0)\, \mathrm{d}\mathbf{p} = \int_{\mathbb{S}^2} \frac{\psi_0}{\|\mathbf{p}\|^3}\, \mathrm{d}\mathbf{p} = \psi_0 \int_{\mathbb{S}^2} 1\, \mathrm{d}\mathbf{p} = \frac{1}{4\pi} \cdot 4\pi = 1.$$

As shown in [MSHJS11],

$$\int_{\mathbb{S}^2} \psi(\mathbf{p}, t)\, \mathrm{d}\mathbf{p} = \int_{\mathbb{S}^2} \psi_0(\mathbf{p})\, \mathrm{d}\mathbf{p}.$$

Hence, the result for $t = 0$ transfers to $t > 0$ and satisfies the equality constraint

$$\int_{\mathbb{S}^2} \psi(\mathbf{p}, t)\, \mathrm{d}\mathbf{p} = 1.$$

2.2.3 Diffusion term

Depending on whether we have a dilute, a semi-dilute or a concentrated suspension, different suspension behaviors are observed in experiments [FT84]. In fact, the particle volume fraction ϕ is a characteristic property of a mixture. Let $a = \frac{l}{d}$ be the aspect ratio of the fibers as introduced in Section 2.2.1. Following [FT84, Tuc91], suspensions are then characterized as

$$\begin{cases} \text{dilute} & \text{if } \phi < \frac{1}{a^2}, \\ \text{semi-dilute/semi-concentrated} & \text{if } \phi \in \left[\frac{1}{a^2}, \frac{1}{a}\right], \\ \text{concentrated} & \text{if } \phi > \frac{1}{a}. \end{cases}$$

While interactions between the fibers are rare in the dilute regime, they are common for higher concentrated regimes [FT84]. The intensity of interactions diminishes with (lower) number density and (shorter) length of fibers. The classical Jeffery equation (2.5) is only valid for dilute suspensions [CK02], since it only takes into account the translation of the fibers, but not their rotary motion caused by the fiber-fiber interactions. However, all the commercially relevant composites are concentrated [FT84, Tuc91].

To broaden the scope of the model, a phenomenological term was added to the classical Jeffery equation in [FT84]. An extended version of the formula for $\dot{\mathbf{p}}$ reads

$$\dot{\mathbf{p}} = \mathbf{W}\mathbf{p} + \lambda_e \left[\mathbf{D}\mathbf{p} - \mathbf{D}\mathbf{p}\mathbf{p}\mathbf{p}\right] - \frac{D_r}{\psi}\nabla_{\mathbf{p}}\psi. \tag{2.19}$$

With this enhancement, also semi-concentrated suspensions can be described and concentrated ones can at least be approximated [Tuc22]. Substituting (2.19) into the conservation law (2.9), we obtain the space-independent Fokker-Planck equation including an isotropic rotary diffusion term [Tuc91].

Traditionally, the diffusive term describes Brownian motion [ADS+24]. For suspensions of industrial interest and for our application, however, Brownian motion is neglected due to the particle size [LDHB88, FMAA20]. Nevertheless, the effect of fiber-fiber interactions can be modeled in the same way as Brownian motion. In both cases, particle collisions induce a diffusion-type mixing process [KOM09].

Finally, we discuss some alternatives to using a constant rotary diffusion coefficient D_r in equation (2.19). Folgar and Tucker [FT84] used $D_r = C_I\sqrt{\mathbf{D}:\mathbf{D}}$, where C_I is a dimensionless interaction coefficient. Being part of an empirical model, the values of both D_r and C_I are unknown a priori. Consequently, they have to be determined by fitting them to experimental data. For the application of injection molding, values between 10^{-3} and 10^{-2} have proven to be suitable choices for C_I. The expression $C_I\sqrt{\mathbf{D}:\mathbf{D}}$ is still scalar, but with $\mathbf{D} = \mathbf{D}(\nabla\mathbf{u})$ it introduces a space dependency. However, this does not cause any restrictions, since $\nabla_{\mathbf{p}}$ represents a differential operator with respect to orientation and therefore even $D_r = C_I\sqrt{\mathbf{D}:\mathbf{D}}$ satisfies

$$\nabla_{\mathbf{p}} \cdot (D_r\nabla_{\mathbf{p}}\psi) = \Delta_{\mathbf{p}}(D_r\psi) = D_r\Delta_{\mathbf{p}}\psi.$$

In [PT09], it is suggested to replace the constant C_I by a tensor $\mathbf{C} = \mathbf{C}(\mathbf{D}, \mathbf{A})$. Instead of isotropic diffusion, which is the same in every direction, this model of fiber interactions uses anisotropic diffusion, which varies with the direction. The additional term in (2.19) prevents the fibers from fully aligning [Tuc91], which agrees with experimental observations, whereas experimental data show that the fibers in concentrated suspensions align more slowly than predicted by (2.19) [PT09, Tuc22]. This can be improved by the anisotropic approach. Within this thesis, however, we stick to the isotropic rotary diffusion, since we assume short-fiber composites, while anisotropy is of major interest for long-fiber composites [PT09].

2.3 Folgar-Tucker equation

Solving the Fokker-Planck equation numerically is a difficult and computationally intensive task. A frequently used alternative is the Folgar-Tucker equation. The orientation is no longer described by the ODF $\psi(\mathbf{x}, \mathbf{p}, t)$. The main quantity of interest is the second-order orientation tensor $\mathbf{A}(\mathbf{x}, t)$. The Folgar-Tucker equation reads [Tuc22]

$$\frac{D\mathbf{A}}{Dt} = \mathbf{W}\cdot\mathbf{A} - \mathbf{A}\cdot\mathbf{W} + \lambda_e(\mathbf{D}\cdot\mathbf{A} + \mathbf{A}\cdot\mathbf{D} - 2\mathbb{A} : \mathbf{D}) + 2D_r(\mathbf{I} - d\mathbf{A}). \tag{2.20a}$$

It is a hyperbolic evolution equation. The tensors $\mathbf{A} = (a_{ij})_{1\le i,j\le d}$ and $\mathbb{A} = (a_{ijkl})_{1\le i,j,k,l\le d}$ are defined below as moments of the function ψ. A particular challenge is the fact that $\mathbb{A}$ is unknown and requires further modeling. All the quantities occurring in (2.20a) apart from $\mathbf{A}$ and $\mathbb{A}$ were already introduced in the context of the Fokker-Planck or the Jeffery equation. To clarify the meaning of $\mathbb{A} : \mathbf{D}$, we write the equation in the equivalent componentwise form

$$\begin{aligned}\frac{Da_{ij}}{Dt} \;=\; \sum_k \omega_{ik}a_{kj} - a_{ik}\omega_{kj} \;+\; \lambda\left(\sum_k d_{ik}a_{kj} + a_{ik}d_{kj} - 2\sum_{kl} a_{ijkl}d_{kl}\right) \\ + 2\,D_r(\delta_{ij} - da_{ij}).\end{aligned} \tag{2.20b}$$

2.3.1 Orientation tensors

Let us consider the definition of the orientation tensors and their properties.

Definition 2.8 (Orientation tensor)**.** *A general n^{th}-order orientation tensor in d dimensions is defined by*

$$\mathbf{A}_n := \int_{\mathbb{S}^{d-1}} \underbrace{\mathbf{p}\otimes\cdots\otimes\mathbf{p}}_{n \text{ times}}\, \psi(\mathbf{p})\,\mathrm{d}\mathbf{p}.$$

Writing out the dyadic product explicitly results in

$$\mathbf{A}_n := (a_{i_1\dots i_n})_{1\le i_1,\dots,i_n\le d}, \quad \textit{where } a_{i_1\dots i_n} := \int_{\mathbb{S}^{d-1}} p_{i_1}\dots p_{i_n}\psi(\mathbf{p})\,\mathrm{d}\mathbf{p}.$$

A tensor $\mathbf{A}_n$ consists of d^n elements, $d, n \in \mathbb{N}_{>0}$. The number of subscripts n describes the order of the tensor, whereas d is the space dimension. A second-order tensor can be interpreted as a matrix.
With regard to the Folgar-Tucker equation (2.20), the second- and the fourth-order orientation tensors are most relevant. They are given by

$$\mathbf{A}_2 := \mathbf{A} = (a_{ij})_{1\le i,j\le d}, \qquad a_{ij} := \int_{\mathbb{S}^{d-1}} p_i\,p_j\,\psi(\mathbf{x},\mathbf{p},t)\,\mathrm{d}\mathbf{p}, \tag{2.21a}$$

$$\mathbf{A}_4 := \mathbb{A} = (a_{ijkl})_{1\le i,j,k,l\le d}, \qquad a_{ijkl} := \int_{\mathbb{S}^{d-1}} p_i\,p_j\,p_k\,p_l\,\psi(\mathbf{x},\mathbf{p},t)\,\mathrm{d}\mathbf{p}. \tag{2.21b}$$

The definition of the orientation tensors also establishes a relationship between the tensors and the function ψ. The attempt to reconstruct ψ from a finite number of

orientation tensors is a challenging task without a unique solution. The computation of $\mathbf{A}$ and $\mathbb{A}$ from ψ is unique by definition. Fortunately, this is exactly what we need when solving the FPE-NSE system later.
Because an orientation tensor is defined in terms of $\mathbf{p}$ and $\psi(\mathbf{x}, \mathbf{p}, t)$, it should possess additional basic properties. We summarize them in the following lemma.

Lemma 2.9 (Properties of orientation tensors)**.**

a) Odd-order tensors vanish, that is, $\mathbf{A}_{2n+1} = \mathbf{0}$.

b) Orientation tensors are fully symmetric. For example, the tensors $\mathbf{A}$ *and* $\mathbb{A}$ *in three dimensions satisfy*

i) $a_{ij} = a_{ji}$

ii) $a_{ijkl} = a_{jikl} = a_{ijlk} = a_{lkij}$

c) The tensors possess the normalization property:

i) $\sum_i a_{ii} = 1$

ii) $\sum_k a_{ijkk} = a_{ij}$

d) The tensors are positive semi-definite:

i) $\mathbf{v}^T \mathbf{A} \mathbf{v} \geq 0 \qquad \forall\, \mathbf{v} \in \mathbb{R}^d$

ii) $\mathbf{V} : (\mathbb{A} : \mathbf{V}) \geq 0 \quad \forall\, \mathbf{V} \in \mathbb{R}^{d \times d}$

Proof. a) We use $\mathbb{S}_+^{d-1}$ and $\mathbb{S}_-^{d-1}$ to denote two halves of the sphere $\mathbb{S}^{d-1}$, i.e., $\mathbb{S}^{d-1} = \mathbb{S}_+^{d-1} \cup \mathbb{S}_-^{d-1}$. Consider an orientation tensor of order $2n+1$. For an arbitrary entry, we have

$$\begin{aligned}
a_{i_1 \dots i_{2n+1}} &= \int_{\mathbb{S}^{d-1}} p_{i_1} \dots p_{i_{2n+1}}\, \psi(\mathbf{p})\, \mathrm{d}\mathbf{p} \\
&= \int_{\mathbb{S}_+^{d-1}} p_{i_1} \dots p_{i_{2n+1}}\, \psi(\mathbf{p})\, \mathrm{d}\mathbf{p} \;+\; \int_{\mathbb{S}_-^{d-1}} p_{i_1} \dots p_{i_{2n+1}}\, \psi(\mathbf{p})\, \mathrm{d}\mathbf{p} \\
&= \int_{\mathbb{S}_+^{d-1}} p_{i_1} \dots p_{i_{2n+1}}\, \psi(\mathbf{p})\, \mathrm{d}\mathbf{p} \;+\; \int_{\mathbb{S}_+^{d-1}} (-1)^{2n+1} p_{i_1} \dots p_{i_{2n+1}}\, \psi(-\mathbf{p})\, \mathrm{d}\mathbf{p} \\
&= \int_{\mathbb{S}_+^{d-1}} p_{i_1} \dots p_{i_{2n+1}}\, \psi(\mathbf{p})\, \mathrm{d}\mathbf{p} \;-\; \int_{\mathbb{S}_+^{d-1}} p_{i_1} \dots p_{i_{2n+1}}\, \psi(\mathbf{p})\, \mathrm{d}\mathbf{p} = 0,
\end{aligned}$$

where the integral over the sphere $\mathbb{S}^{d-1}$ was decomposed into those over two arbitrary semispheres and the symmetry property (2.3c) of $\psi(\mathbf{p})$ was exploited.

b) This statement follows directly from the definition.

c) Using the definition of the orientation tensors, the fact that $\|\mathbf{p}\| = 1$ and (2.3b), the normalization property of $\mathbf{A}$ and $\mathbb{A}$ is valid due to

$$\begin{aligned}
\sum_i a_{ii} &= \int_{\mathbb{S}^2} \sum_i p_i^2\, \psi(\mathbf{x}, \mathbf{p}, t)\, \mathrm{d}\mathbf{p} = \int_{\mathbb{S}^2} \|\mathbf{p}\|^2\, \psi(\mathbf{x}, \mathbf{p}, t)\, \mathrm{d}\mathbf{p} = 1, \\
\sum_k a_{ijkk} &= \int_{\mathbb{S}^2} p_i p_j \sum_k p_k^2\, \psi(\mathbf{x}, \mathbf{p}, t)\, \mathrm{d}\mathbf{p} = \int_{\mathbb{S}^2} p_i p_j\, \|\mathbf{p}\|^2\, \psi(\mathbf{x}, \mathbf{p}, t)\, \mathrm{d}\mathbf{p} = a_{ij}.
\end{aligned}$$

d) Writing out the matrix-vector multiplication for i) and the tensor contraction for ii), applying the definition of the orientation tensors, rearranging the components properly, and using the non-negativity of ψ, we find that

$$\begin{aligned}
\mathbf{v}^T\mathbf{A}\mathbf{v} &= \sum_i \left(v_i \sum_j a_{ij} v_j \right) = \sum_i \left(v_i \sum_j \int_{\mathbb{S}^2} p_i p_j \; \psi(\mathbf{x}, \mathbf{p}, t) \, \mathrm{d}\mathbf{p} \, v_j \right) \\
&= \int_{\mathbb{S}^2} \sum_i v_i p_i \sum_j v_j p_j \; \psi(\mathbf{x}, \mathbf{p}, t) \; \mathrm{d}\mathbf{p}, = \int_{\mathbb{S}^2} \left(\sum_i v_i p_i \right)^2 \; \psi(\mathbf{x}, \mathbf{p}, t) \; \mathrm{d}\mathbf{p} \; \geq 0, \\
\mathbf{V} : (\mathbb{A} : \mathbf{V}) &= \sum_{ij} v_{ij} \sum_{kl} a_{ijkl} v_{kl} = \int_{\mathbb{S}^2} \sum_{ijkl} p_i \, p_j \, p_k \, p_l \, v_{ij} \, v_{kl} \, \psi(\mathbf{x}, \mathbf{p}, t) \, \mathrm{d}\mathbf{p} \\
&= \int_{\mathbb{S}^2} \left(\sum_{ij} p_i p_j v_{ij} \right)^2 \psi(\mathbf{x}, \mathbf{p}, t) \, \mathrm{d}\mathbf{p} \; \geq 0.
\end{aligned}$$

□

These mathematical properties are of major importance. The fact that odd-order tensors are zero explains their absence in the Folgar-Tucker equation. Because of the symmetry effectively less entries have to be stored during practical calculations. For $d = 3$ only six entries of $\mathbf{A}$ have to be stored instead of nine entries, namely $a_{11}, a_{22}, a_{33}, a_{12}, a_{13}, a_{23}$. Similarly instead of 81 entries only 15 entries are needed for a complete description of the tensor $\mathbb{A}$, namely a_{1111}, a_{2222}, a_{3333}, a_{1112}, a_{1113}, a_{2221}, a_{2223}, a_{3331}, a_{3332}, a_{1122}, a_{2233}, a_{3311}, a_{1123}, a_{2213} and a_{3312}.
Normalization and positive semi-definiteness provide useful tools to verify that the numerical results are physically meaningful. A second-order tensor $\mathbf{A}$ has the normalization property if its trace is equal to one. It is positive semi-definite if and only if all eigenvalues are non-negative.

Meaning of the orientation tensors. The second-order orientation tensors allow a physical interpretation. In a conventional reference coordinate system, the first, second and third axis correspond to the direction of inflow, cross-flow and thickness. The diagonal entries of the orientation tensor describe the degree of orientation with respect to theses axes, while the off-diagonal entries express the tilt from the coordinate axes [GGOS20].

A random/isotropic distribution is characterized by $a_{11} = a_{22} = a_{33} = \frac{1}{3}$, whereas a biaxial random-in-plane orientation is, for instance, given by $a_{11} = a_{22} = \frac{1}{2}$ and $a_{33} = 0$. If the fibers are perfectly aligned in the direction of the i^{th}-axis, we find that $a_{ii} = 1$, while all other entries are zero. Both extreme cases, full alignment and random distribution, are sketched in Figure 2.1.

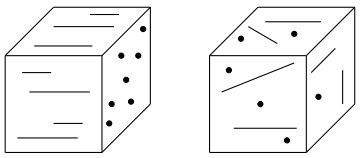


Figure 2.1: *i) fully aligned and ii) random/isotropic distribution*

It is not always possible to determine uniquely, which distribution function ψ belongs to a given orientation tensor. Let us imagine a cube inside a sphere, whose faces are parallel to the coordinate axes. If the number of fibers pointing into the eight

vertices of the cube are equal, we obtain exactly the same orientation tensor as if the fibers were evenly distributed over the entire sphere.

Furthermore, the eigenvalues and eigenvectors are useful to characterize the orientation distribution. The eigenvectors of $\mathbf{A}$ determine the principal directions of the fiber alignment, while the corresponding eigenvalues describe the extent of fiber alignment in that direction [AT87, KOM09].

2.3.2 Comparison of Folgar-Tucker and Fokker-Planck equation

The Folgar-Tucker and Fokker-Planck equations can be used to simulate fiber orientation especially in the framework of deterministic numerical approaches, see [FT84] and [AT87]. The FTE is derived from the FPE and can be understood as an approximation of the FPE. In [LDHB88], the derivation of the FTE is performed by multiplying the original Jeffery equation from left and right with $\mathbf{p}$ and averaging the result. We summarize the connection between the two PDEs in the next lemma and prove it following Lohmann [Loh19].

Lemma 2.10 (Relationship between FTE and FPE)**.** *The Folgar-Tucker equation (2.20) for the fiber orientation tensor $\mathbf{A}(\mathbf{x},t)$ is a moment of the Fokker-Planck equation (2.2) for the orientation distribution function $\psi(\mathbf{x},\mathbf{p},t)$.*

Proof. The basic steps to derive the Folgar-Tucker equation read

$$
\begin{aligned}
\frac{\mathrm{d}a_{ij}}{\mathrm{d}t} &\overset{(1)}{=} \int_{\mathbb{S}^2} p_i p_j \frac{\mathrm{d}\psi}{\mathrm{d}t}\,\mathrm{d}\mathbf{p} \\
&\overset{(2)}{=} -\int_{\mathbb{S}^2} p_i p_j \,\nabla_{\mathbf{p}}\cdot(\dot{\mathbf{p}}\psi)\,\mathrm{d}\mathbf{p} + D_r \int_{\mathbb{S}^2} p_i p_j \Delta_{\mathbf{p}}\psi\,\mathrm{d}\mathbf{p} \\
&\overset{(3)}{=} \int_{\mathbb{S}^2} \nabla_{\mathbf{p}}(p_i p_j)\cdot\dot{\mathbf{p}}\psi\,\mathrm{d}\mathbf{p} + D_r \int_{\mathbb{S}^2} \Delta_{\mathbf{p}}(p_i p_j)\psi\,\mathrm{d}\mathbf{p} \\
&\overset{(4)}{=} \int_{\mathbb{S}^2} \dot{p}_i p_j \psi\,\mathrm{d}\mathbf{p} + \int_{\mathbb{S}^2} p_i \dot{p}_j \psi\,\mathrm{d}\mathbf{p} + 2D_r \int_{\mathbb{S}^2} (\delta_{ij} - d p_i p_j)\,\psi\,\mathrm{d}\mathbf{p} \\
&\overset{(5)}{=} \sum_k (\omega_{ik}a_{kj} - a_{ik}\omega_{kj}) + \lambda\left[\sum_k (d_{ik}a_{kj} + a_{ik}d_{kj}) - 2\sum_{kl} a_{ijkl}d_{kl}\right] \\
&\qquad + 2D_r(\delta_{ij} - d a_{ij}).
\end{aligned}
$$

We justify the validity of each equality in detail. For an evolution equation in terms of $\mathbf{A}$ it makes sense to consider the time derivative of this tensor. For identity (1), only the definition of $\mathbf{A}$ given in (2.21a) is applied. In step (2), the Fokker-Planck equation (2.2) is inserted. For the material derivative defined by (2.10) we obtain

$$
\frac{\mathrm{d}\psi}{\mathrm{d}t} = \frac{\partial\psi}{\partial t} + \mathbf{u}\cdot\nabla_{\mathbf{x}}\psi = -\nabla_{\mathbf{p}}\cdot(\dot{\mathbf{p}}\psi) + D_r\Delta_{\mathbf{p}}\psi.
$$

Equality (3) results from integration by parts. The expression stays relatively simple because $\mathbb{S}^2$ has no boundaries and, consequently, no boundary integral arises. The

greatest effort is hidden behind step (4). The essential formula is $\partial_{p,l} p_k = \delta_{kl} - p_k p_l$, $k, l \in \{1, \dots, d\}$, see Corollary 5.14. The first summand can be rewritten as

$$\begin{aligned}
&\int_{\mathbb{S}^2} \nabla_{\mathbf{p}}(p_i p_j) \cdot \dot{\mathbf{p}}\, \psi \, \mathrm{d}\mathbf{p} \\
&= \int_{\mathbb{S}^2} \left[\sum_k (\partial_{p,k} p_i) p_j \dot{p}_k + p_i (\partial_{p,k} p_j) \dot{p}_k \right] \psi \, \mathrm{d}\mathbf{p} \\
&= \int_{\mathbb{S}^2} \left[\sum_k (\delta_{ik} - p_i p_k) p_j \dot{p}_k + p_i (\delta_{jk} - p_j p_k) \dot{p}_k \right] \psi \, \mathrm{d}\mathbf{p} \\
&= \int_{\mathbb{S}^2} \left[\dot{p}_i p_j + p_i \dot{p}_j \right] \psi \, \mathrm{d}\mathbf{p}.
\end{aligned}$$

We have

$$\begin{aligned}
\Delta_{\mathbf{p}}\, p_i \;&=\; \sum_k \partial^2_{p,k} p_i \;=\; \sum_k \partial_k \left(\delta_{ik} - p_i p_k \right) \;=\; -\sum_k (\delta_{ik} - p_i p_k) p_k \;+\; p_i (1 - p_k^2) \\
&=\; -\sum_k \delta_{ik} p_k \;+\; 2 \sum_k p_i p_k^2 \;-\; \sum_k p_i \;=\; -p_i + 2 p_i - d p_i \;=\; (1 - d)\, p_i.
\end{aligned}$$

Together with another application of Corollary 5.14, this leads to

$$\begin{aligned}
\Delta_{\mathbf{p}} \left(p_i p_j \right) &= \left(\Delta_{\mathbf{p}}\, p_i \right) p_j + p_i (\Delta_{\mathbf{p}} p_j) \;+\; 2\, \nabla_{\mathbf{p}}\, p_i \cdot \nabla_{\mathbf{p}}\, p_j \\
&= 2(1 - d) p_i p_j + 2 \sum_k (\delta_{ik} - p_i p_k)(\delta_{jk} - p_j p_k) \\
&= 2(1 - d) p_i p_j + 2 \sum_k (\delta_{ik}\delta_{jk} - \delta_{ik} p_j p_k - \delta_{jk} p_i p_k + p_i p_j p_k^2) \\
&= 2(1 - d) p_i p_j \;+\; 2 \left(\delta_{ij} - p_i p_j - p_i p_j + p_i p_j \right) \\
&= 2(1 - d) p_i p_j + 2\delta_{ij} - 2 p_i p_j \\
&= 2(\delta_{ij} - d\, p_i p_j)
\end{aligned}$$

In the final step (5), we use Jeffery's equation (2.5a) in the elementwise form

$$\dot{p}_i = \sum_k \omega_{ik}\, p_k + \lambda \left[\sum_k d_{ik}\, p_k - \sum_{kl} d_{kl}\, p_k p_l p_i \right]$$

and the definition (2.21) of the orientation tensors to show that

$$\begin{aligned}
&\int_{\mathbb{S}^2} \left(\dot{p}_i p_j + p_i \dot{p}_j \right) \psi \, \mathrm{d}\mathbf{p} \\
&= \sum_k \omega_{ik} a_{kj} \;+\; \lambda \left[d_{ik} a_{kj} - d_{kl} a_{ijkl} \right] \;+\; \sum_k \omega_{jk} a_{ki} \;+\; \lambda \left[d_{jk} a_{ki} - d_{kl} a_{ijkl} \right] \\
&= \sum_k (\omega_{ik} a_{kj} \;-\; a_{ik} \omega_{kj}) + \lambda \left[\sum_k (d_{ik} a_{kj} + a_{ik} d_{kj}) \;- 2 \sum_{kl} a_{ijkl} d_{kl} \right].
\end{aligned}$$

□

Advantages of the Folgar-Tucker equation over the Fokker-Planck equation. The FPE was and is replaced by the FTE quite often. The FTE is used in commercial applications as well as in existing publications, including recent ones. This involves both solving the FTE itself [Loh19] and using results from the FTE as a reference solution for the FPE [FMAA20].
The main reason to employ the FTE instead of the FPE is that the orientation tensors allow a much more compact representation than the ODF. Let us assume a fixed point $(\mathbf{x}_i, t)$ in space and time. Using the ODF $\psi(\mathbf{x}_i, \mathbf{p}, t)$, on the one hand, the orientation distribution is described by a few hundred points on the sphere $\mathbb{S}^2$, see Sections 5.5.2 and 7.2.2.3. For the FTE in 3d, on the other hand, each tensor $\mathbf{A}(\mathbf{x}_i, t)$ consists of nine elements, which can even be reduced to five elements taking advantage of the symmetry and the normalization property.
Furthermore, already orientation tensors are useful to predict the effect of fibers on material properties [CT95]. Last but not least, the coupling to the NSE makes parts of the FTE relevant for us. Even if we do not consider the FTE, but the FPE in what follows, the NSE do require the orientation tensors $\mathbf{A}$ and $\mathbb{A}$ as input rather than the probability density ψ, see (2.33). Thus, the tensors have to be calculated anyway.

Simplified solution for the Folgar-Tucker equation. If the fibers can be assumed to be parallel to the streamlines of the flow, to be infinitely small, i.e., $\lambda_e = 1$, and not to interact, the so-called aligned fiber approximation reads [LDHB88, Tuc91]

$$\mathbf{A} = (\mathbf{v} \otimes \mathbf{v}) \|\mathbf{v}\|_2^{-2}.$$

Using this expression, the velocity field is responsible not only for the fluid motion but also for the fiber orientation. On the one hand, no property of the orientation can be violated and the computational costs are very low. On the other hand, this approximation has a very limited applicability. It is only justified for simple flow fields. It is unable to predict orientation states, where fibers are not perfectly aligned and it does not suffice for technologically important flows as they arise, e.g., in the process of injection molding [Tuc91]. To obtain a solution in more general cases, the FTE can be solved using numerical methods such as the flux-corrected finite element schemes designed in [Loh19].

2.3.2.1 Closures

So far, no special attention was paid to the unknown fourth-order orientation tensor, which arises in the FTE. It is possible to set up an evolution equation for $\mathbb{A} = \mathbf{A}_4$ but this equation contains $\mathbf{A}_6$, see [AT87]. Accordingly, any equation for a tensor $\mathbf{A}_{2m}$ includes the tensor $\mathbf{A}_{2m+2}$, so that the problem is only shifted. Instead, we try to reconstruct higher order information to approximate the unknown tensor in terms of lower order orientation tensors. Since these approaches 'close' the evolution equation in some sense, they are referred to as closures.

In the literature, plenty of different closures can be found. First, there are relatively simple long-standing approaches like the linear, the quadratic and the hybrid closure.

The reconstruction corresponding to the linear closure reads [Han62]

$$\begin{aligned}&\big(a_{ijkl}\big)_{\text{linear}}\\ &= \alpha(a_{ij}\delta_{kl} + a_{ik}\delta_{jl} + a_{il}\delta_{jk} + a_{jk}\delta_{il} + a_{jl}\delta_{ik} + a_{kl}\delta_{ij}) + \beta(\delta_{ij}\delta_{kl} + \delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk})\\ &= 6\alpha\ \mathbb{S}(a_{ij}\delta_{kl}) + 3\beta\ \mathbb{S}(\delta_{ij}\delta_{kl}),\end{aligned}$$

where $\mathbb{S}$ is the so-called symmetrization operator. The quadratic closure, whose origin is traced back to [HL76, LDHB88], satisfies

$$\big(a_{ijkl}\big)_{\text{quadratic}} = a_{ij}a_{kl}.$$

Finally, the hybrid closure presented in [AT87, AT90] is defined by

$$\big(a_{ijkl}\big)_{\text{hybrid}} = \lambda \cdot \big(a_{ijkl}\big)_{\text{linear}} + (1-\lambda) \cdot \big(a_{ijkl}\big)_{\text{quadratic}}.$$

While the linear closure is exact when the fibers are in a random state, the quadratic closure is exact for perfectly aligned fibers [CT95]. The hybrid closure takes a linear combination of the linear and the quadratic closure to combine their favorable properties. For randomly oriented fibers, $\lambda = 1$ is the right choice, whereas $\lambda = 0$ is appropriate for perfectly aligned fibers. In the literature, we find, for instance, the suggestion to set $\lambda = d^d \det(a_{ij})$ [DV99].
Another problem of the given closures is that they violate some basic properties of orientation tensors given in Lemma 2.9. Using the parameters $\alpha = (d+4)^{-1}$ and $\beta = ((d+2)(d+4))^{-1}$ the linear closure is normalized. However, it does not preserve the positive definiteness. The quadratic closure has the 'symmetry of an elasticity tensor' since $a_{ijkl} = a_{jikl}$, $a_{ijkl} = a_{ijlk}$ and $a_{ijkl} = a_{klij}$, but is not fully symmetric. We look for methods that produce physics-compatible results.

More advanced closures are the natural closures. They are based on the insight that in the absence of fiber-fiber interaction, i.e., for $D_r = 0$, and for specific initial distributions, there is a one-to-one correspondence between the second- and the fourth-order orientation tensor [LDHB88, DV99]. However, in 3d no exact solution is available anymore, so that a numerical solution becomes necessary [VCD94, DV99].

Another family of closures are the orthotropic closures. Their construction is based on the requirement that a closure approximation be orthotropic in the sense that second- and fourth-order orientation tensors have the same principal axes [CT95]. In [CT95] the 'orthotropic smooth closure' (ORS) and an 'orthotropic fitted closure' (ORF) were developed. The former approach is based on linear interpolation between uni-, bi- and triaxial orientation states. For the latter approach, the distribution function ψ was calculated with the FPE for a wide variety of orientation states and data fitting was applied. The ideas of the natural closure and the orthotropic fitted closure are combined by the 'invariant-based optimal fitting approach' (IBOF) [CK02].

The 'exact closures' developed in [MSHJS11] are based on the exact solution of the extended Jeffery equation as presented in Theorem 2.4. An alternative is offered by the interpolatoric closures in [Kuz18], where ideas from natural and orthotropic closures are combined. Technically, the reconstruction is performed for a few characteristic orientation states and the obtained parameters are interpolated to determine

$\mathbb{A}$ for arbitrary states. Finally, the Bingham closure, which is based on the calculation of a Bingham distribution, leads to admissible orientation tensors [FCL98].

Advantages of the Fokker-Planck equation over the Folgar-Tucker equation. Despite significant progress in constructing good closure approximations, closures are always designed for specific flow configurations. Moreover, the closures proposed in [CT95, CK02] are fitted to specific experimental or numerical data of ψ [KOM09]. Hence, an advantage of using the FPE is that we are no longer dependent on the closures associated to the FTE.
In addition, a density distribution ψ contains significantly more information than any number of orientation tensors. For example, ψ might be of special interest in boundary layers, where a tensorial description becomes inaccurate. Moreover, stiffness tensors [Tuc22] and other quantities, which are relevant in engineering, require the knowledge of ψ. Several approaches to reconstruct the probability distribution from orientation tensors exist. A first one, which does not ensure that the reconstruction has the desired physical properties, can be found in [AT87]. More recent positivity-preserving approaches, [Loh16b, BSK19], take advantage of the equivalence between the orientation tensors and the first coefficients of the Fourier/ spherical harmonics series expansion of ψ. However, such reconstructions require a large number of moments to be realistic, so that the effort increases.
Even today, solving the Fokker-Planck equation involves a lot of effort and it is debatable to what extent this is justified. However, when we argue for the FPE, the development of computers must be taken into account. Back in the 1960s, the observation was made that the CPU speed doubles about every 18 month and this empirical relationship became known as Moore's law. Even if this law no longer applies today, the trend is clear. This makes the FPE a promising modeling tool not only for derivation of fitted closures but also for direct numerical simulation of evolving orientation states/PDEs.

2.4 Navier-Stokes equations

2.4.1 Newtonian fluids

The Navier-Stokes equations (NSE) were set up in the first half of the 19^{th} century. Since then, they are used to model a wide range of phenomena where the flow behavior of a fluid, i.e., a gas or a liquid, has to be described. Until this day, the NSE are an area of active mathematical research. The standard NSE consist of two conservation laws, one for momentum and one for mass. Following [Ran17c, ESW14], these PDEs are presented below. The conservation of mass is expressed by the continuity equation

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho \mathbf{u}) = 0, \tag{2.22}$$

where $\rho > 0$ is the density and $\mathbf{u}$ is the velocity field of the fluid. For an incompressible fluid the equation simplifies to

$$\nabla \cdot \mathbf{u} = 0. \tag{2.23}$$

We outline the derivation of the momentum equation. In fluid mechanics, the momentum can be described by the integral over a volume $V = V(t)$ as

$$\boldsymbol{p} = \int_V \rho \mathbf{u} \, \mathrm{d}\mathbf{x}. \tag{2.24}$$

Combining the basic definition of the momentum, $\boldsymbol{p} = \mathrm{m}\mathbf{u}$, with Newton's second law, $\mathbf{F} = \mathrm{m}\mathbf{a}$, where m is the mass of a body, $\mathbf{a}$ its acceleration and $\mathbf{F}$ the force acting on the body, we obtain $\mathbf{F} = \dot{\boldsymbol{p}}$, and therefore

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_V \rho \mathbf{u} \, \mathrm{d}\mathbf{x} = \mathbf{F}. \tag{2.25}$$

Applying Reynolds transport theorem and the divergence theorem, we find that

$$\int_V \left[\frac{\partial}{\partial t}(\rho \mathbf{u}) + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{u}) \right] \, \mathrm{d}\mathbf{x} = \mathbf{F}. \tag{2.26}$$

This force can be split into the sum of a body and a surface force. Thus,

$$\mathbf{F} = \mathbf{F}_1 + \mathbf{F}_2 = \int_V \rho \mathbf{g} \, \mathrm{d}\mathbf{x} + \int_{\partial V} \boldsymbol{\sigma} \mathbf{n} \, \mathrm{d}\sigma, \tag{2.27}$$

where $\mathbf{g}$ represents a gravitational force density, $\mathbf{n}$ is the outer normal of ∂V, and $\boldsymbol{\sigma} \in \mathbb{R}^{d \times d}$ denotes the Cauchy stress tensor. Separation of the total stress into a hydrostatic component associated with pressure p and the additional viscous shear stress tensor $\boldsymbol{\tau} \in \mathbb{R}^{d \times d}$ yields $\boldsymbol{\sigma} = -p\mathbf{I} + \boldsymbol{\tau}$. Making use of this expression and the divergence theorem, we end up with

$$\mathbf{F} = \int_V \left(\rho \mathbf{g} - \nabla p + \nabla \cdot \boldsymbol{\tau} \right) \, \mathrm{d}\mathbf{x}. \tag{2.28}$$

Since the divergence operator applied to tensor rather than a vector, it produces a vector rather than a scalar. In particular,

$$\boldsymbol{\tau} = \begin{pmatrix} \tau_{xx} & \tau_{xy} & \tau_{xz} \\ \tau_{yx} & \tau_{yy} & \tau_{yz} \\ \tau_{zx} & \tau_{zy} & \tau_{zz} \end{pmatrix} \implies \nabla \cdot \boldsymbol{\tau} = \begin{pmatrix} \frac{\partial \tau_{xx}}{\partial x} + \frac{\partial \tau_{xy}}{\partial y} + \frac{\partial \tau_{xz}}{\partial z} \\ \frac{\partial \tau_{yx}}{\partial x} + \frac{\partial \tau_{yy}}{\partial y} + \frac{\partial \tau_{yz}}{\partial z} \\ \frac{\partial \tau_{zx}}{\partial x} + \frac{\partial \tau_{zy}}{\partial y} + \frac{\partial \tau_{zz}}{\partial z} \end{pmatrix}.$$

According to the two expressions (2.26) and (2.28) for the force $\mathbf{F}$ and the assumption that $\nabla \cdot \mathbf{u} = 0$ is constant, a general formulation of the momentum equation reads

$$\rho \left(\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u} \right) = \rho \mathbf{g} - \nabla p + \nabla \cdot \boldsymbol{\tau}. \tag{2.29}$$

For Newtonian fluids, it is usually assumed that

$$\boldsymbol{\tau} = \lambda (\nabla \cdot \mathbf{u}) \mathbf{I} + 2\mu \mathbf{D} \overset{(2.23)}{=} 2\mu \mathbf{D}, \tag{2.30}$$

where λ is the volume viscosity, μ is the dynamic viscosity, and $\mathbf{D}$ is the deformation tensor defined by (2.6). The viscosity variables may also depend on pressure or

temperature [BT92a, Tuc22]. Within this thesis, however, isothermal fluids are assumed. The divergence of our shear stress tensor is given by

$$\nabla \cdot \boldsymbol{\tau} = \mu \Delta \mathbf{u}. \tag{2.31}$$

Finally, we set $P = p/\rho$ and $\nu = \mu/\rho$. Based on the kinematic viscosity ν, we define the dimensionless Reynolds number $\mathrm{Re} := UL/\nu$, where U is the characteristic velocity and L the characteristic length. The value of 'Re' characterizes different types of flows ranging from turbulent to laminar ones.

Summing up, the incompressible NSE for a Newtonian fluid can be written as

$$\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u} = \nu \Delta \mathbf{u} - \nabla P + \mathbf{g}, \tag{2.32a}$$

$$\nabla \cdot \mathbf{u} = 0, \tag{2.32b}$$

that is, as a system for the unknown velocity $\mathbf{u}$ and pressure P.

The momentum equation is nonlinear since the velocity $\mathbf{u}$ is convected by itself [KH15]. In fact, the left hand side of (2.32a) exactly represents the convective derivative, see (2.10). The nonlinear term is often denoted by $\mathbf{u} \cdot \nabla \mathbf{u}$ in the literature, see, e.g., [ESW14]. In 3d, the corresponding scalar product is defined by

$$\mathbf{u} \cdot \nabla \mathbf{u} = (\mathbf{u} \cdot \nabla u_x, \mathbf{u} \cdot \nabla u_y, \mathbf{u} \cdot \nabla u_z)^T.$$

Remark 2.11 (Stokes equations)**.** Considering a stationary problem and assuming a zero Reynolds number limit, the Navier-Stokes equations simplify to the Stokes equations

$$\nabla P = \nu \Delta \mathbf{u} + \mathbf{g},$$
$$\nabla \cdot \mathbf{u} = 0.$$

While the incompressibility constraint remains, in contrast to the Navier-Stokes equations we do not have to take the nonlinearity into account.

To simulate flows of mixtures, the FPE is sometimes coupled with the Stokes equations [LC03, Loz03, CL04a, CL04a, HO06]. In other cases it is coupled to the NSE [Kne06, Kne08, KS09b], or even both options are used [KS09a].

2.4.2 Non-Newtonian fluids

Macroscopically, a fiber suspension behaves as a single fluid, whose nonlinear rheology depends on the local orientation state. Because of the fibers within the fluid, we use a non-Newtonian flow model in this case. The NSE given by (2.32) are valid only in the regions between the fibers. Thus, generalized incompressible NSE are needed to take the impact of fiber-induced stresses into account. In principle, the conservation equations remain the same, but we have to formulate an adapted constitutive law for the stress. The result can then be inserted into the momentum equation (2.29) in the form of the tensor $\boldsymbol{\tau}$.

An expression for $\boldsymbol{\tau}$ taking into account the fibers located in the fluid reads [Tuc91, Tuc22]

$$\boldsymbol{\tau} = 2\mu_I(\mathbf{D} + N_p \mathbb{A} : \mathbf{D} + N_s(\mathbf{DA} + \mathbf{AD})). \tag{2.33a}$$

The constant μ_I is an isotropic viscosity [Tuc22]. The dimensionless 'particle number' N_p and the 'shear number' N_s represent the anisotropic contributions to the viscosity introduced by the fibers. As calculations for specific cases demonstrate, both N_p and N_s depend on the aspect ratio l/d and on the volume fraction ϕ, but not on the orientation state of the fibers [Tuc91]. For $N_p = N_s = 0$, equation (2.33a) breaks down to (2.30), that is, to the Newtonian case. For aspect ratios like $l/d = 10$ and even more for larger aspect ratios, we find that $N_p \gg 1$ but $N_s \ll N_p$ [Tuc91, Tuc22]. Consequently, in the case of slender fibers the particle number N_p dominates and it is acceptable to neglect the term with N_s. The constitutive law then simplifies to

$$\boldsymbol{\tau} = 2\mu_I(\mathbf{D} + N_p \mathbb{A} : \mathbf{D}), \tag{2.33b}$$

so that $\boldsymbol{\tau} = \boldsymbol{\tau}(\mathbf{A}, \mathbb{A})$ is reduced to $\boldsymbol{\tau} = \boldsymbol{\tau}(\mathbb{A})$. Overall, the particle orientation distribution has an effect on the flow and this effect is represented by the non-Newtonian component of the stress tensor.

Clearly, the hydrodynamic behavior of a fiber suspension depends on the volume fraction ϕ of fibers (see Section 2.2.3). In [Tuc22, Sec.6.3.3], the particle number N_p is approximated as a function of ϕ. In our model, we assume that ϕ is constant and specify N_p directly.
In [Tuc91, Fig.1], the parameter N_p is plotted as the ratio of a and ϕ. The values $N_p = 0.1, 1, 10, 100, 1000$ represent the spectrum from dilute to semi-concentrated and concentrated suspensions. Using $N_p = 6$ later in this work, we perform numerical studies for a semi-concentrated suspension.

3 Finite element method

3.1 Numerical methods for PDEs

Partial differential equations (PDEs) are omnipresent in mathematically oriented scientific fields. They model phenomena ranging from the natural sciences to the financial world. Popular numerical approaches to solve a PDE are the finite difference (FD), the finite volume (FV), and the finite element (FE) methods.

Starting point for all these methods is the discretization of the computational domain. An FD method approximates the derivatives of the PDE using Taylor series. The FD methods are the oldest technique to discretize PDEs. Their strengths are simplicity and efficiency, but their applicability is limited to structured grids.

FV and FE schemes support the use of unstructured grids. The traditional strengths of these methods are complementary. The FE methods emerged in the context of elliptic PDEs, whereas FV methods are particularly well suited for hyperbolic transport problems. Since FV schemes are based on the integral conservation laws, they are both globally and locally conservative by construction [KH15]. First- and second-order approximations on structured meshes are particularly straightforward, whereas difficulties arise for higher-order schemes or when numerical differentiation is required for diffusive terms [KH15]. Then FE methods may be preferred, even for convection-dominated transport problems.

The FE methodology is powerful due to its great flexibility and its wide applicability. It can be used both on unstructured meshes and for multidimensional problems. Moreover, in contrast to FD and FV schemes the FE method is backed by rigorous mathematical theory. It is noteworthy that with a suitably chosen notation more similarities than differences can be detected between FV and low-order FE [GS98]. Therefore certain findings, for instance in the framework of limiting, can be transferred from one method to another. Throughout this work, we focus on the FE method. Its procedure can roughly be divided into the three steps:

1.) derive a weak formulation of the PDE,
2.) discretize it using suitable basis functions,
3.) assemble an algebraic system for the discrete unknowns.

These aspects are considered in detail in the following sections and the method is illustrated for the convection-diffusion equation. Moreover, the matrices of the linear system are analyzed and time stepping schemes are presented.
The presentation of the FE method in this chapter, is mainly based on [KH15, Ran17b]. Further standard texts are, e.g., [Eva10, Bra13]. Theoretical aspects are covered in [LT05], while the books [KA13, ESW14] are more application-oriented. Two books including further aspects of this thesis as well are [Loh19, Kuz10].

3.2 Weak formulation

A classical solution of an α^{th}-order PDE, $\alpha \in \mathbb{N}$, belongs to the class C^α, that is, it has to be continuously differentiable α times. For the most PDEs arising in practical applications, we cannot expect a solution with such a high degree of smoothness. Instead we introduce the Sobolev spaces, which are fundamental for the method of finite elements and represent a cornerstone for the solution theory of PDEs. They impose only weak requirements on the regularity of the solution and thereby broaden the space of admissible functions considerably. In order to define them, we need the functional analytic concepts of Lebesgue spaces and weak derivatives.

Definition 3.1 (Lebesgue space $L^p(\Omega)$)**.** *Let $p \in [1, \infty)$, $\Omega \subset \mathbb{R}^d$. Then the Lebesgue space $L^p(\Omega)$ is defined by*

$$L^p(\Omega) := \left\{ u : \Omega \to \mathbb{R} \;\middle|\; \|u\|_{L^p(\Omega)} := \left(\int_\Omega |u|^p \, \mathrm{d}\mathbf{x} \right)^{1/p} < \infty \right\}.$$

Equipped with the norm $\|\cdot\|_{L^p(\Omega)}$ the Lebesgue space $L^p(\Omega)$ is a Banach space. The space of square-integrable functions $L^2(\Omega)$ equipped with the scalar product [Bra13]

$$(u, v)_{L^2(\Omega)} := \int_\Omega uv \, \mathrm{d}\mathbf{x}$$

is a Hilbert space. The next definition is motivated by integration by parts. We use a multiindex notation with $\alpha \in \mathbb{N}_0^d$, where $\partial^\alpha := \partial_1^{\alpha_1} \dots \partial_d^{\alpha_d}$ and $|\alpha| = \sum_{i=1}^d \alpha_i$.

Definition 3.2 (Weak derivative)**.** *[KA13, Def. 2.9]. Let $u \in L^2(\Omega)$. A function $v \in L^2(\Omega)$ is called weak derivative $\partial^\alpha u$ if*

$$\int_\Omega v\varphi \, \mathrm{d}\mathbf{x} \;=\; (-1)^{|\alpha|} \int_\Omega u \, \partial^\alpha \varphi \, \mathrm{d}\mathbf{x} \quad \forall \varphi \in C_c^\infty(\Omega).$$

With these preliminaries the Sobolev space can now be defined.

Definition 3.3 (Sobolev space $W^{k,p}(\Omega)$)**.** *[LT05], [Sch13, Def. 3.16]. Let $p \in [1, \infty)$, $k \in \mathbb{N}_0$. The Sobolev space $W^{k,p}(\Omega)$ is defined by*

$$W^{k,p}(\Omega) := \{ u : \Omega \to \mathbb{R} \mid \partial^\alpha u \in L^p(\Omega) \quad \forall \alpha \in \mathbb{N}_0^n \text{ with } |\alpha| \le k \}$$

and the corresponding norm reads

$$\|u\|_{W^{k,p}(\Omega)} := \left(\sum_{|\alpha| \le k} \|\partial^\alpha u\|_{L^p(\Omega)}^p \right)^{1/p}.$$

Obviously, $W^{0,p}(\Omega) = L^p(\Omega)$. While every Sobolev space is a Banach space, only $H^k(\Omega) := W^{k,2}(\Omega)$ is a Hilbert space as well. For $p = 2$ and $k = 1$ we obtain

$$H^1(\Omega) := \{ u : \Omega \to \mathbb{R} \mid \partial_i u \in L^2(\Omega) \; \forall \, i \in \{1, \dots, d\} \}$$

and the H^1-norm is

$$\|u\|_{H^1(\Omega)} = \left(\|u\|^2_{L^2(\Omega)} + \|\nabla u\|^2_{L^2(\Omega)}\right)^{1/2} = \left(\int_\Omega |u|^2 + |\nabla u|^2 \ \mathrm{d}\mathbf{x}\right)^{1/2}, \tag{3.1a}$$

where $|\nabla u|^2 := (\nabla u)^T(\nabla u)$. Considering only the second part of $\|u\|_{H^1(\Omega)}$, we obtain the H^1-seminorm

$$|u|_{H^1(\Omega)} = \|\nabla u\|_{L^2(\Omega)} = \left(\int_\Omega |\nabla u|^2 \ \mathrm{d}\mathbf{x}\right)^{1/2}. \tag{3.1b}$$

It is only a seminorm, since $|u|_{H^1(\Omega)} = 0$ does not guarantee that $u = 0$. However, in the case of zero boundary values it becomes a norm. In fact, $H^1(\Omega)$ is often restricted to zero boundary values, that is, to [KA13, Eq. (2.19)]

$$H^1_0(\Omega) := \{u \in H^1(\Omega) | u = 0 \text{ on } \partial\Omega\}.$$

In a more abstract setting, the weak formulation of a general PDE reads

$$\text{Find } u \in V \text{ such that } a(u, v) = b(v) \quad \forall\, v \in V, \tag{3.2}$$

where V is a vector space, $a(\cdot, \cdot)$ is a bilinear form and $b(\cdot)$ is a bilinear form. A typical space for V is the Sobolev space $H^1(\Omega)$ as it contains the weak derivatives which are required for the weak formulation of second-order PDEs.

3.3 Discretization

3.3.1 Triangulation

In order to obtain a finite dimensional problem, discretization is necessary. We decompose the domain Ω into polygonal elements K. This decomposition is called triangulation $\mathcal{T}_h$ and the union of the elements is called the mesh/grid. Let us consider a formal definition of a triangulation.

Definition 3.4 (Triangulation)**.** *[KA13]. An admissible triangulation of $\Omega \subset \mathbb{R}^d$ is a finite set $\mathcal{T}_h = \{K_1, \ldots, K_e\}$ of closed polygons/polyhedrons and*

1. *$\overline{\Omega} = \bigcup_{K \in \mathcal{T}_h} K$.*
2. *The elements K, K' have the property that $int(K) \cap int(K') = \emptyset$, where $int(K)$ denotes the interior of the element K.*
3. *If $K \cap K' \neq \emptyset$, then the intersection $K \cap K'$ is a face, an edge, or a node.*

The first property states that the elements have to cover Ω completely, while the second one implies that the elements have to be disjoint. The third criterion makes the triangulation admissible. For instance, it prohibits hanging nodes, see Figure 3.1.

We call the vertices of the elements nodes. The mesh size can be defined by $h := \{\mathrm{diam}(K) \mid K \in \mathcal{T}_h\}$, where $\mathrm{diam}(K) := \sup\{\|\mathbf{x} - \mathbf{y}\| \mid \mathbf{x}, \mathbf{y} \in K\}$ [KA13].

In 1d, the domain is subdivided into intervals, while in 2d triangles or rectangles and in 3d tetrahedra or hexahedra are used. In order to avoid problems in numerical calculations, shape regularity has to be guaranteed. In case of triangles, it can be ensured by a minimum angle condition, whereas an aspect ratio condition is suitable for rectangles [ESW14]. For time-dependent problems, choosing the grid as uniform as possible has a positive effect on the permissible time step size.

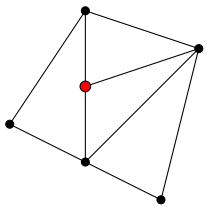
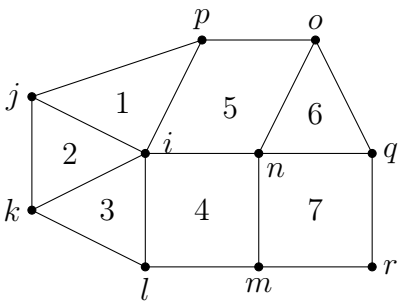


Figure 3.1: *Example of a hanging node*

Notation. We denote the nodes of the triangulation by $\mathbf{x}_1, \dots, \mathbf{x}_N$. Let ε_i be the set of indices of the elements containing the node $\mathbf{x}_i$. The nodal stencil $\mathcal{N}_i$ then is the set of indices of the nodes belonging to these elements, and $\mathcal{N}_i^* := \mathcal{N}_i \backslash \{i\}$. Additionally, the element stencil $\mathcal{N}^e$ defines the set of indices containing all the nodes, which belong to an element K^e, see Figure 3.2. Let us remark that in Figure 3.2 a mixed mesh is applied even though not every FE software is able to handle meshes consisting of both simplices and hypercubes.



$$\begin{aligned} \varepsilon_i &= \{1,2,3,4,5\} \\ \mathcal{N}_i &= \{i,j,k,l,m,n,o,p\} \\ \mathcal{N}_i^* &= \{j,k,l,m,n,o,p\} \\ \mathcal{N}^5 &= \{i,n,o,p\}. \end{aligned}$$

Figure 3.2: *Example of element patch ε_i, nodal stencil $\mathcal{N}_i$ and element stencil $\mathcal{N}^e$.*

In an abstract setting, a finite element is a triple (K, P_K, Σ_K), where [Bra13, KA13]

1.) $K \in \mathcal{T}_h$ is the basic geometric area,
2.) $P_K \subset C(K)$ is a local (polynomial) space,
3.) Σ_K is the set of degrees of freedom.

These components are examined in the next section.

3.3.2 Basis functions

Now that we have chosen a mesh for a finite element discretization in space, the continuous weak formulation (3.2) can be approximated by its discrete counterpart

$$\text{Find } u_h \in V_h \text{ such that} \quad a(u_h, w) = b(w) \quad \forall\, w \in W_h. \tag{3.3}$$

The trial space V_h and the test space W_h are finite dimensional vector spaces. Their exact definition will be given below in the equations (3.7) and (3.10). Throughout this work, we use a Ritz-Galerkin method, that is, $V_h = W_h$. Moreover, we stick to a conforming approximation, that is, $V_h \subseteq V$. Since $\dim V_h =: N < \infty$, a basis $\{\varphi_j\}_{j=1}^N$ exists and $V_h = \operatorname{span}\{\varphi_1, \dots, \varphi_N\}$. The basis function are also called shape

or trial functions here. A finite element function $u_h \in V_h$ that approximates the solution $u \in V$, can be expressed as

$$u_h = \sum_{j=1}^{N} u_j \varphi_j, \tag{3.4a}$$

where the coefficients u_j uniquely determine the finite element function. We assume that the $\{u_1, \ldots, u_N\}$ are not given in advance by boundary values, which are determined, e.g., by $V_h \subset H_0^1(\Omega)$; see [ESW14, Sec. 1.3]. For a space- and time-dependent function $u(\mathbf{x}, t)$, approximation (3.4a) can be written more precisely as

$$u_h(\mathbf{x}, t) = \sum_{j=1}^{N} u_j(t)\, \varphi_j(\mathbf{x}). \tag{3.4b}$$

This approach involves a separation of variables, where the coefficients u_j are time-dependent. The coefficients u_j are also called degrees of freedom. In the most common case, they are simply function values at nodal points. However, the coefficients of a general FE approximation may also represent, e.g., higher-order and normal derivatives or even integral averages over the edge of an element [Bra13].

Substituting the approximate solution u_h defined in (3.4a) into the discrete weak formulation (3.3) and setting $w = \varphi_i$, we find

$$a\left(\sum_{j=1}^{N} u_j \varphi_j,\ \varphi_i\right) \ = \ b(\varphi_i) \quad \forall\, i = 1, \ldots, N. \tag{3.5a}$$

The concrete bilinear form $a(\cdot, \cdot)$ depends on the given PDE. Using the properties of a bilinear form, we can transform expression (3.5a) into

$$\sum_{j=1}^{N} \underbrace{a(\varphi_j, \varphi_i)}_{:= a_{ij}} u_j \ = \ \underbrace{b(\varphi_i)}_{:= b_i} \quad \forall\, i = 1, \ldots, N. \tag{3.5b}$$

The transition from $a(\varphi_j, \varphi_i)$ to a_{ij} and from $b(\varphi_i)$ to b_i in (3.5b) requires a complex assembly process in practice. Some aspects are roughly outlined in Appendix 10.3. The new notation reveals the structure of a matrix-vector multiplication. Hence, the coefficients of u_h satisfy an algebraic system of the form

$$Au = b. \tag{3.5c}$$

Note that the name u is used twice in different contexts. Within the PDE, $u : \mathbb{R}^d \to \mathbb{R}$ is a function, whereas in the corresponding linear system the vector of coefficients $u \in \mathbb{R}^N$ is a discrete quantity.

Example: Convection-diffusion equation. As a model problem for the Fokker-Planck equation we consider a scalar time-dependent convection-diffusion equation, see [Kuz10, Sec. 2.1.1]. On the one hand, it is a generalization the spatial advection

equation (2.4a). On the other hand, it has the structure of the space-independent FPE (2.4b). Our convection-diffusion equation reads

$$\frac{\partial u}{\partial t} + \nabla \cdot (\mathbf{v}u) - \varepsilon \Delta u = 0 \quad \text{in } \Omega \times (0, T], \tag{3.6a}$$

where $u : \mathbb{R}^d \to \mathbb{R}$ is a scalar conserved quantity, $\mathbf{v} \in \mathbb{R}^d$ a constant velocity, and $\varepsilon > 0$ a diffusion coefficient. As a first weak formulation, we consider

$$\int_\Omega \left[\frac{\partial u}{\partial t} + \nabla \cdot (\mathbf{v}u) - \varepsilon \, \Delta u \right] w \, \mathrm{d}\mathbf{x} = 0 \quad \forall \, w \in V. \tag{3.6b}$$

The requirements for the differentiability of the weak solution are reduced by transferring derivatives onto the test functions. Here, we apply the integration by parts only to the diffusive term and not to the convective term, so that

$$\begin{aligned} \int_\Omega \Big[\frac{\partial u}{\partial t} \, w \; + \; \nabla \cdot (\mathbf{v}u) \, w \; + \; \varepsilon \, \nabla u \cdot \nabla w \, \mathrm{d}\mathbf{x} \Big] \\ - \; \varepsilon \int_{\partial\Omega} w \, \nabla u \cdot \mathbf{n} \, \mathrm{d}s \; = \; 0 \quad \forall \, w \in V. \end{aligned} \tag{3.6c}$$

We choose $V = H_0^1(\Omega)$. Because of the zero boundary values, the boundary integral can be omitted. Remark 3.8a) lists further situations, where this simplification is justified. More general approaches are given, for instance, in [KH23].
For a proper discretization of Ω, an approximation to (3.6a) reads

$$\int_\Omega \frac{\partial u_h}{\partial t} \, w_h \; + \; \nabla \cdot (\mathbf{v}u_h) \, w_h \; + \; \varepsilon \, \nabla u_h \cdot \nabla w_h \, \mathrm{d}\mathbf{x} \; = \; 0 \quad \forall \, w_h \in W_h. \tag{3.6d}$$

In a next step, approximation (3.4) is employed for u_h and $w_h := \varphi_i$. Arranging the components in a suitable way leads to

$$\begin{aligned} \sum_j \Big[\left(\int_\Omega \varphi_i \varphi_j \, \mathrm{d}\mathbf{x} \right) \frac{\mathrm{d}u_j}{\mathrm{d}t} + \left(\int_\Omega \nabla \cdot (\mathbf{v}\varphi_j) \varphi_i \, \mathrm{d}\mathbf{x} \right) u_j \\ + \; \varepsilon \left(\int_\Omega \nabla \varphi_i \cdot \nabla \varphi_j \, \mathrm{d}\mathbf{x} \right) u_j \Big] \; = \; 0 \quad \forall \, i \in \{1, \ldots, N\} \end{aligned} \tag{3.6e}$$

This can finally be interpreted as the linear system

$$M_C \dot{u} \; + \; K u \; + \; \varepsilon \, L u \; = \; 0, \tag{3.6f}$$

where M_C is the consistent mass matrix, K is the convection matrix, and L is the stiffness matrix. They are investigated in more detail in Section 3.4.

Differential operator		FE matrix and its entries	
$\partial u / \partial t$	(time derivative)	$M_C = (m_{ij})_{i,j=1}^N$,	$m_{ij} := \int_\Omega \varphi_i \varphi_j \, \mathrm{d}\mathbf{x}$
$-\Delta u$	(diffusive part)	$L = (l_{ij})_{i,j=1}^N$,	$l_{ij} := \int_\Omega \nabla \varphi_i \cdot \nabla \varphi_j \, \mathrm{d}\mathbf{x}$,
$\nabla \cdot (\mathbf{v}u)$	(convective part)	$K = (k_{ij})_{i,j=1}^N$,	$k_{ij} := \int_\Omega \nabla \cdot (\mathbf{v}\varphi_j) \varphi_i \, \mathrm{d}\mathbf{x}$

Table 3.3: *FE matrices for the different terms of the convection-diffusion equation.*

Finite element spaces. The properties of the FE discretization and of the resulting algebraic system (3.5c) are ultimately determined by the choice of the space V_h. The conformity requirement $V_h \subseteq V$ already leads to conditions for the smoothness of the shape function in V_h and the choice of boundary conditions. Under these restrictions, there are still several options for the choice of the shape functions. If the elements K are simplices, we define a general finite element space by

$$V_h^{(\alpha,r)} := \{\varphi \in C^\alpha\left(\overline{\Omega}\right) \mid \varphi|_K \in \mathcal{P}_r(K) \quad \forall\, K \in \mathcal{T}_h\}. \tag{3.7}$$

The parameter α indicates how many times the shape functions are continuously differentiable. For our purposes, it is sufficient to have continuous functions ($\alpha = 0$). The parameter r indicates that the shape functions are piecewise polynomials of degree at most r. For example, $V_h^{(0,1)}$ consists of continuous and piecewise linear shape functions and satisfies $V_h^{(0,1)} \subseteq H^1(\Omega)$ [KA13, Lemma 3.20].

In d space dimensions the space of the polynomials $\varphi : K \to \mathbb{R}$ has the form

$$\mathcal{P}_r := \operatorname{span}\left\{\sum_{i=1}^{d} x_i^{\alpha_i} \;\middle|\; 0 \le \sum_{i=1}^{d} \alpha_i \le r,\ \alpha_i \in \mathbb{N}_0\right\}, \tag{3.8}$$

that is, $\mathcal{P}_1 = \operatorname{span}\{1, x, y\}$ for triangles ($d = 2$) and $\mathcal{P}_1 = \operatorname{span}\{1, x, y, z\}$ for tetrahedra ($d = 3$). If $\mathcal{T}_h$ consists of hypercubes, an appropriate polynomial space is

$$\mathcal{Q}_r := \operatorname{span}\left\{\sum_{i=1}^{d} x_i^{\alpha_i} \;\middle|\; 0 \le \alpha_1, \ldots, \alpha_d \le r,\ \alpha_i \in \mathbb{N}_0\right\} \tag{3.9}$$

Concretely, the spaces for the multilinear polynomials are $\mathcal{Q}_1 = \operatorname{span}\{1, x, y, xy\}$ for rectangles ($d = 2$) and $\mathcal{Q}_1 = \operatorname{span}\{1, x, y, z, xy, xz, yz, xyz\}$ for hexahedra ($d = 3$). In the case of polynomials from $\mathcal{Q}_r$ we define the element space by

$$V_h^{(\alpha,r)} := \{\varphi \in C^\alpha\left(\overline{\Omega}\right) \mid \varphi|_K \circ \tau_K \in \mathcal{Q}_r(\widehat{K}) \quad \forall\, K \in \mathcal{T}_h\}, \tag{3.10}$$

where $\tau_K : \widehat{K} \to K$ is a transformation from a reference element $\widehat{K}$ to a physical element K [QV08, Loh19].

Design principles. We are not restricted to (multi-)linear shape functions. In particular, for smooth data the order of convergence improves with increasing polynomial degree r. However, the choice of higher-order basis functions also has disadvantages. A downside is that the overall effort increases. Assembling the FE matrices and solving the system becomes more expensive because the matrices consist of more entries, which must be both calculated and stored.

A general requirement for the basis functions is their simplicity. A finite element matrix has to have as few non-zero entries as possible. A sparse system matrix is not only advantageous but even necessary to solve large systems in a realistic time. This is ensured by choosing shape functions with a compact support. Such functions disappear for the vast majority of grid points. For $r > 1$ we obtain a stronger coupling, the stencil of each shape function becomes larger, so that the sparsity pattern of the FE matrix changes in an unfavorable way.

Lagrange basis functions. Within every polynomial space there are various ways for the specification of basis functions [Bra13, ESW14]. The nodal basis represents a particularly simple choice for which the degrees of freedom are function values on certain points in K, in the most basic case even function values in the physical nodes, see Figure 3.5a) and b). Moreover, the basis functions are non-zero for exactly one degree of freedom [Bra13]. We call such functions Lagrange elements since they result from the piecewise application of Lagrange interpolation. The basis functions are uniquely defined by the Kronecker delta, that is, by the requirement

$$\varphi_j(\mathbf{x}_i) \ := \delta_{ij} \ := \begin{cases} 1, & i = j \\ 0, & \text{otherwise} \end{cases} \qquad \forall\, i, j \in \{1, \dots, N\}.$$

The piecewise-linear shape functions in 1d are visualized in Figure 3.4.

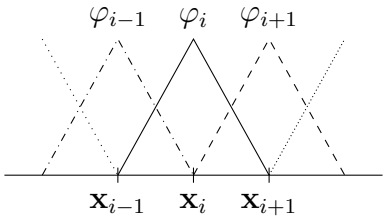
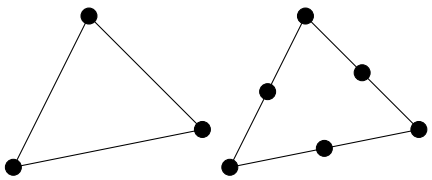
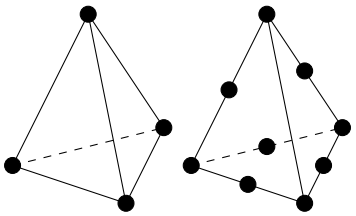


Figure 3.4: *Lagrange basis functions in 1d.*

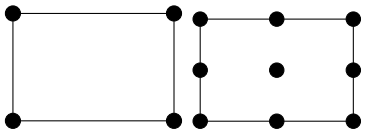
All in all, Lagrange functions are widely used, since they represent a simple conforming approach and their shape functions have a compact support. We denote them by $\mathbf{P}_r$ and $\mathbf{Q}_r$. In Figure 3.5 they are visualized for $r = 1$ and $r = 2$.



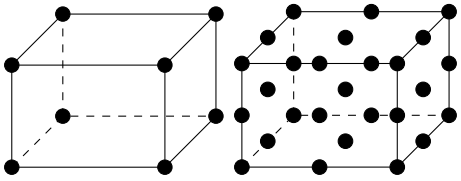
a) $\mathbf{P}_1$ *and* $\mathbf{P}_2$ *for triangles.*



b) $\mathbf{P}_1$ *and* $\mathbf{P}_2$ *for tetrahedra.*



c) $\mathbf{Q}_1$ *and* $\mathbf{Q}_2$ *for rectangles.*



d) $\mathbf{Q}_1$ *and* $\mathbf{Q}_2$ *for hexahedra.*

Figure 3.5: *Visualization of the degrees of freedom for different (bi-, tri-)linear and (bi-, tri-)quadratic Lagrange basis functions.*

Properties of (multi-)linear Lagrange basis functions. The methods that we present in Chapter 4 exploit some essential properties of well-known (multi-)linear nodal basis functions and corresponding FE approximations u_h. Following [Loh19], we summarize those properties as follows:

i)	$\text{supp}(\varphi_i) = \{\cup K_i \mid i \in \varepsilon_i\}$	(compact support)
ii)	$u_h(\mathbf{x}_i) = u_i \quad \forall\, i \in \{1, \dots, N\}$	(nodal values)
iii)	$\varphi_i > 0 \quad \forall\, i \in \{1, \dots, N\}$	(positivity)
iv)	$\sum_{i=1}^{N} \varphi_i = 1$	(partition of unity)
v)	$\sum_{i=1}^{N} \nabla\varphi_i = 0$	(vanishing gradient sum)
vi)	$\min_{j\in\mathcal{N}^e} u_j \leq u_h(\mathbf{x}) \leq \max_{j\in\mathcal{N}^e} u_j \quad \forall\, \mathbf{x} \in K^e$	(discrete maximum principle)

The implications of i)-vi) will be explained below when it comes to proving a particular result.

3.4 Finite element matrices

Mass and stiffness matrix

The FE method is usually introduced with the Laplace or Poisson equation. The resulting stiffness matrix is denoted by L like Laplacian. In our application, the impact of the diffusive term is quite small due to the accompanying coefficient. Moreover, unlike the convective part, on regular meshes the diffusive part causes no trouble and therefore does not require any special treatment.
Extending the Poisson equation to the heat equation, the discretization of the temporal derivative leads to the mass matrix M. This matrix also appears in other contexts, e.g. when discretizing a reactive term u. Stiffness matrix and mass matrix, whose definition can already be found in Table 3.3, share some basic properties that are collected in the following lemma.

Lemma 3.5 (Basic properties of M and L)**.**

a) Both the mass matrix M and the stiffness matrix L are symmetric. Moreover, M is positive definite and L is positive semi-definite.

b) Considering the spectral norm $\kappa_2 := \lambda_{max}/\lambda_{min}$, the condition numbers of the FE matrices with respect to the mesh size h are

$$\kappa_2(M) = \mathcal{O}(1) \quad \textit{and} \quad \kappa_2(L) = \mathcal{O}(h^{-2}).$$

Proof. See [KA13, Ran17b]. □

We recognize that the condition numbers depend on the differential operator but not on the polynomial degree of the basis functions or the space dimension. For the stiffness matrix, the condition also depends on the mesh size h. It behaves like $\mathcal{O}(h^{-2})$. Consequently, a decreasing mesh size $h \to 0$ improves the accuracy but worsens the condition number.

As recorded before, the consistent, i.e., unmodified, mass matrix is defined by

$$M_C = (m_{ij})_{i,j=1}^{N}, \quad m_{ij} := \int_\Omega \varphi_i \varphi_j \, \mathrm{d}\mathbf{x}. \tag{3.11a}$$

From several points of view, it is reasonable or even necessary to use a diagonal mass matrix M_L instead [Han94]. Such a lumped matrix is given by [Kuz10]

$$M_L = \operatorname{diag}(m_i)_{i=1}^N, \quad m_i := \sum_{j=1}^N m_{ij}. \tag{3.11b}$$

In this definition, each row sum of M_C is used as a diagonal entry for the corresponding row of the diagonal matrix M_L. The advantages of M_L over M_C are multifaceted. When we apply an explicit time stepping scheme, solving the linear system is simplified immensely. Furthermore, the use of M_L instead of M_C can increase the stability of the system.
Several properties of the mass matrix, and in particular of M_L, are collected in the following lemma, which owes a lot of content to [Loh19, Sec. 4.3.1.2].

Lemma 3.6 (Properties of M_L)**.** *The lumped mass matrix defined by* (3.11b) *satisfies*

a) $m_i > 0$ *b)* $\sum_i m_i = |\Omega|$ *c)* $\int_\Omega u_h = \sum_i m_i u_i$.

Proof. To prove a) the partition of unity property of the basis functions is central, thus

$$m_i = \sum_j m_{ij} = \sum_j \int_\Omega \varphi_i \varphi_j \, \mathrm{d}\mathbf{x} = \int_\Omega \varphi_i \underbrace{\sum_j \varphi_j}_{=1} \, \mathrm{d}\mathbf{x} = \int_\Omega \varphi_i \, \mathrm{d}\mathbf{x} > 0 \tag{3.12}$$

The fact that $m_i = \int_\Omega \varphi_i \, \mathrm{d}\mathbf{x}$ is also used to prove b):

$$\sum_{i,j} m_{ij} = \sum_i m_i = \sum_i \int_\Omega \varphi_i \, \mathrm{d}\mathbf{x} = \int_\Omega \underbrace{\sum_i \varphi_i}_{=1} \, \mathrm{d}\mathbf{x} = \left|\Omega\right| \tag{3.13}$$

Reusing the proof for a) one more time yields c):

$$\int_\Omega u_h \, \mathrm{d}\mathbf{x} = \int_\Omega \sum_i u_i \varphi_i(\mathbf{x}) \, \mathrm{d}\mathbf{x} \; = \; \sum_i u_i \int_\Omega \varphi_i(\mathbf{x}) \, \mathrm{d}\mathbf{x} \; = \sum_i m_i u_i \tag{3.14}$$

□

Convection matrix

Let us take a look at the convection matrix. While there is just one typical approach to define the entries of the stiffness matrix, different (quadrature-based) approximation may be used for the convection matrix. Following the approach in Section 3.3.2, the individual entries of the convection matrix K are defined by

$$k_{ij} := \int_\Omega \varphi_i \nabla \cdot (\mathbf{v}\varphi_j) \, \mathrm{d}\mathbf{x}. \tag{3.15}$$

This formulation uses the consistent Galerkin approximation to the convective term $\nabla \cdot (\mathbf{v}u)$. No further modifications are applied within the weak formulation, and the velocity field $\mathbf{v}$ is not discretized here.

Lemma 3.7 (Basic properties of K)**.** *Let the velocity field* $\mathbf{v}$ *be solenoidal, that is,* $\nabla \cdot \mathbf{v} = 0$*. Then*

a) $\sum_{j=1}^{N} k_{ij} = 0 \quad \forall\, i \in \{1, \ldots, N\}$ *(zero row sum).*

In addition, assume that the boundary integral resulting from the integration by parts of k_{ij} *vanishes, see Remark 3.8. Then*

b) $k_{ij} = -k_{ji} \quad \forall\, i, j \in \{1, \ldots, N\}$ *(skew-symmetry).*

Remark. The skew-symmetry implies that the diagonal elements of K are zero.

Proof. (of Lemma 3.7). We use formulation (3.15) to prove both the zero row sum property and the skew-symmetry of a convection matrix.

a) To show the zero row sum property, we change the order of summation and differentiation. This is allowed because of the linearity of the integral and of the divergence operator. Since the basis functions form a partition-of-unity and $\mathbf{v}$ is solenoidal, we have

$$\sum_{j=1}^{N} k_{ij} = \int_\Omega \varphi_i \nabla \cdot \big(\mathbf{v} \underbrace{\sum_{j=1}^{N} \varphi_j}_{=1}\big)\, \mathrm{d}\mathbf{x} = \int_\Omega \varphi_i \underbrace{(\nabla \cdot \mathbf{v})}_{=0}\, \mathrm{d}\mathbf{x} = 0.$$

b) Next the skew-symmetry has to be proven. First we apply elementwise integration by parts to k_{ij}, and use the assumption that the boundary integral disappears. Then we make use of the product rule. The resulting expression simplifies since $\mathbf{v}$ is solenoidal. In this way, we obtain

$$\begin{aligned} k_{ij} = & \int_\Omega \nabla \cdot (\mathbf{v}\varphi_j)\, \varphi_i \, \mathrm{d}\mathbf{x} \\ = & -\int_\Omega (\mathbf{v} \cdot \nabla \varphi_i)\, \varphi_j \, \mathrm{d}\mathbf{x} + \underbrace{\int_{\partial\Omega} (\mathbf{v} \cdot \mathbf{n})\, \varphi_i \varphi_j \, \mathrm{d}S}_{=0} \\ = & -\int_\Omega \big[\nabla \cdot (\mathbf{v}\varphi_i) - \underbrace{(\nabla \cdot \mathbf{v})}_{=0} \varphi_i\big] \varphi_j \, \mathrm{d}\mathbf{x} \\ = & -\int_\Omega \nabla \cdot (\mathbf{v}\varphi_i)\, \varphi_j \, \mathrm{d}\mathbf{x} \\ = & -k_{ji} \end{aligned}$$

□

Remarks 3.8 (Further properties of K)**.**

a) (Boundary integrals/skew-symmetry.) Boundary integrals such as in the proof of Lemma 3.7b) arise due to integration by parts. For the internal nodes the integrals cancel out due to continuity. Consequently, the skew-symmetry always holds true for these nodes.

Furthermore, the assumption of vanishing boundary integrals is justified if periodic boundary conditions are prescribed, if the solution values at the boundary are zero (homogeneous Dirichlet boundaries), or if there are no boundaries such as in the case of a PDE to be solved on the sphere.

b) (Spectral condition.) We did not find any statements about the condition number for the convection matrix in the literature. In any case, a convection matrix is not symmetric, so that no real eigenvalues can be assumed and the calculation of the spectral norm becomes difficult.

c) (Different formulations.) Using integration by parts, as in the proof of Lemma 3.7, the formula for an entry of the matrix K reads

$$k_{ij} := -\int_\Omega (\mathbf{v}\cdot\nabla\varphi_i)\,\varphi_j \,\mathrm{d}\mathbf{x} + \int_{\partial\Omega} (\mathbf{v}\cdot\mathbf{n})\,\varphi_i\varphi_j \,\mathrm{ds}. \tag{3.16a}$$

An alternative transformation using the product rule $\nabla\cdot(\mathbf{v}u) = (\nabla\cdot\mathbf{v})u + \mathbf{v}\cdot\nabla u$ yields We find that

$$k_{ij} := \int_\Omega (\nabla\cdot\mathbf{v})\varphi_i\varphi_j + \int_\Omega (\mathbf{v}\cdot\nabla\varphi_j)\,\varphi_i \,\mathrm{d}\mathbf{x}. \tag{3.16b}$$

Omitting the boundary integral in (3.16a) might introduce an error for the boundary nodes. Omitting the first term in (3.16b) is only permitted if $\mathbf{v}$ is solenoidal. Otherwise, the mass is not conserved anymore.

Since the sum of the gradients of the basis functions vanishes, see page 43, and since the dot product is linear, the matrix K has zero column sums, i.e., $\sum_i k_{ij} = 0$, in (3.16a) and zero row sums, i.e., $\sum_j k_{ij} = 0$ in (3.16b).

3.5 Temporal discretization

For stationary problems, the spatial discretization of the PDE yields a fully discrete system. For time-dependent problems, spatial discretization leads only to a semi-discrete formulation and an additional temporal discretization is required to obtain a fully discrete system.
One possibility for the temporal discretization are space-time Galerkin methods, where time is treated as an additional space dimension [DH03, Haj22]. Another concept is the method of lines, where space and time are discretized one after another. We distinguish between horizontal and vertical method of lines [KA13]. The former, also known as Rothe-method, first discretizes in time and then in space, whereas the frequently used vertical method of lines does it the other way around.
Applying a finite element discretization in space to a general initial value problem, we obtain

$$M\dot{u}(t) = f(u(t),t),\ t\in(0,T] \quad \text{and} \quad u(0) = u^0, \tag{3.17}$$

where M is the mass matrix, $\dot{u}(t)$ the temporal derivative of the time-dependent solution vector $u(t)$, and the right hand side f summarizes all the remaining components. The solution $u(t)$ is approximated by discrete vectors $u(t^0), u(t^1), \ldots, u(t^n)$, where $0 =: t^0 < t^1 < \ldots < t^n := T$ is the discretization in time. For equidistant time steps $\Delta t := t^{j+1} - t^j$, $j \in \{0,\ldots,n-1\}$, the discrete time instants are $t^l := l\Delta t$, $l\in\{0,\ldots,n\}$.

A variety of methods have been developed to handle time-dependent equations numerically. We consider the class of θ-methods as well as different Runge-Kutta schemes, including strong stability preserving ones.

Theta-schemes. A set of basic methods approximate $\dot{u}$ using finite differences. The forward difference leads to the explicit *forward Euler* method

$$M\frac{u^{n+1}-u^n}{\Delta t} = f(u^n) \quad \Longleftrightarrow \quad Mu^{n+1} = Mu^n + \Delta t f(u^n), \tag{3.18a}$$

The backward difference yields the implicit *backward Euler* method

$$M\frac{u^{n+1}-u^n}{\Delta t} = f(u^{n+1}) \quad \Longleftrightarrow \quad Mu^{n+1} = Mu^n + \Delta t f(u^{n+1}), \tag{3.18b}$$

Finally, the central difference results in the *Crank-Nicolson* method

$$\begin{aligned} M\frac{u^{n+1}-u^n}{\Delta t} &= \frac{1}{2}\big(f(u^n)+f(u^{n+1})\big) \\ &\Longleftrightarrow \quad Mu^{n+1} = Mu^n + \frac{\Delta t}{2}\big(f(u^n)+f(u^{n+1})\big). \end{aligned} \tag{3.18c}$$

While the forward and backward finite differences are first order accurate in time, the central difference is of second order. This transfers directly to the time stepping methods. The three above approaches can be combined within the θ-scheme

$$Mu^{n+1} = Mu^n + \Delta t\big[(1-\theta)f(u^n) + \theta f(u^{n+1})\big], \quad \theta \in [0,1]. \tag{3.19}$$

Obviously, $\theta = 0$ corresponds to the forward Euler, $\theta = 1$ to the backward Euler and $\theta = 0.5$ to the Crank-Nicolson method which can be understood as an averaging of the forward and backward Euler scheme. The degree of implicitness is determined by the choice of θ. The fully discrete scheme is explicit if $\theta = 0$ and M is a diagonal matrix. Practical examples for the different time stepping methods including observations on stability can be found in Section 4.4.2.

Runge-Kutta schemes. Like the θ-schemes, the Runge-Kutta schemes are a family of single-step methods, i.e., they use only information from the last time step to determine the value for the new time step. Moreover, most of the Runge-Kutta routines are defined for different stages, i.e., they use intermediate steps to reach higher order. Applied to (3.17), the family of explicit R-stage Runge-Kutta methods can be written as [Ran17a]

$$\begin{aligned} Mu^{n+1} = Mu^n + \Delta t \sum_{r=1}^{R} c_r k_r, \\ k_1 &= f(t_n, u^n), \\ k_2 &= f(t_n + a_2\Delta t, u^n + \Delta t\, b_{21} k_1), \\ k_3 &= f(t_n + a_3\Delta t, u^n + \Delta t\,(b_{31}k_1 + b_{32}k_2)), \ \ldots, \\ k_r &= f(t_n n + a_r\Delta t, u^n + \Delta t \sum_{s=1}^{r-1} b_{rs}k_s). \end{aligned}$$

Implicit methods would run the sum over s not only up to $r-1$, but up to r. After choosing R, the number of stages, the coefficients c_r, a_r and b_{rs} must be determined. Usually, there is more than one meaningful option. Requirements for the

scheme to be consistent and of a certain order provide equations for the coefficients. Specifically, a Taylor expansion is performed and used to compare it with the coefficients [Ran17a]. For $R = 1$, the Runge-Kutta scheme reduces to the forward Euler method. For $R = 2$, we obtain $c_1 + c_2 = 1$ and $c_2 a_2 = c_2 b_{21} = 0.5$, which is, for example, fulfilled by $c_1 = c_2 = 0.5$ and $a_2 = b_{21} = 1$. The resulting scheme is called Heun's method and reads

$$Mu^{n+1} = Mu^n + \frac{\Delta t}{2}\big(f(t_n, u^n) + f(t_{n+1}, u^n + \Delta t f(t_n, u^n))\big). \tag{3.20a}$$

In the case of an autonomous differential equation, which does not explicitly depend on time t, it reduces to

$$Mu^{n+1} = Mu^n + \frac{\Delta t}{2}\big(f(u^n) + f(u^n + \Delta t f(u^n))\big). \tag{3.20b}$$

The case $R = 4$ has a special significance because no explicit RK scheme can achieve order R with R stages if $R > 4$. The classical 4^{th}-order scheme is

$$\begin{aligned} Mu^{n+1} = Mu^n + \frac{\Delta t}{6}(k_1 + 2k_2 + 2k_3 + k_4), \\ k_1 = f(t_n, u^n), \quad k_2 = f(t_{n+\frac{1}{2}}, u^n + \frac{\Delta t}{2} k_1), \\ k_3 = f(t_{n+\frac{1}{2}}, u^n + \frac{\Delta t}{2} k_2), \quad k_4 = f(t_{n+1}, u^n + \Delta t k_3). \end{aligned}$$

Conditions for stability and other desirable properties of the above schemes are investigated in [Kuz10]. We now turn to the method that will be used for large parts of the numerical examples in this work, the so-called SSP-RK2 method.

Strong stability preserving Runge-Kutta methods. The strong stability preserving Runge-Kutta (SSP-RK) methods were developed over the past decades [GKS11]. A beginning is marked by [SO88], where the approaches are still called total variation diminishing (TVD) schemes. The development was mainly motivated by the discontinuous solutions of hyperbolic PDEs and the corresponding ODEs resulting from the vertical method of lines.

The forward Euler time stepping method preserves stability if it is combined with a suitable spatial discretization. Using the semi-discrete formulation and a certain norm $\|\cdot\|$, for all numerical solutions u we assume that

$$\|Mu^{n+1}\| = \|Mu^n + \Delta t f(u^n)\| \leq \|Mu^n\| \qquad \text{if } \Delta t \in [0, \Delta t_{FE}]. \tag{3.21}$$

The drawback of the forward Euler scheme is that it is only first-order accurate. Therefore SSP methods use convex combinations of Euler steps to construct higher order discretizations, which maintain stability. We call a method strong stability preserving with SSP-coefficient $\mathcal{C}$ if $\|u^{n+1}\| \leq \|u^n\|$ is satisfied and the time step restriction $\Delta t \leq \mathcal{C}\Delta t_{FE}$ holds [GKS11].

A wide variety of such methods has been constructed. Let us consider the combination of two forward Euler steps for the general formulation (3.17):

1. Forward Euler for $u^n \to \widetilde{u}^{n+1}$:

$$M\,\frac{\widetilde{u}^{n+1}-u^n}{\Delta t} = f(u^n) \iff M\widetilde{u}^{n+1} = Mu^n + \Delta t\, f(u^n).$$

2. Forward Euler for $\widetilde{u}^{n+1} \to \widetilde{u}^{n+2}$:

$$M\,\frac{\widetilde{u}^{n+2}-\widetilde{u}^{n+1}}{\Delta t} = f(\widetilde{u}^{n+1}) \iff M\widetilde{u}^{n+2} = M\widetilde{u}^{n+1} + \Delta t\, f(\widetilde{u}^{n+1}).$$

3. Averaging:

$$\begin{aligned} Mu^{n+1} &= \frac{1}{2}\left(Mu^n + M\widetilde{u}^{n+2}\right) = \frac{1}{2}M(u^n + \widetilde{u}^{n+1}) + \frac{\Delta t}{2}\, f(\widetilde{u}^{n+1}) \\ &= Mu^n + \frac{\Delta t}{2}\big(f(u^n) + f(u^n + \Delta t f(u^n))\big). \end{aligned} \tag{3.22}$$

We recognize the basic procedure here: Each intermediate step of an explicit SSP-RK scheme is a forward Euler step. Finally, a convex combination of the intermediate results is formed, which remains in the convex set. During the development of SSP-RK methods, it was realized that certain standard Runge-Kutta methods can be written as convex combination of Euler steps, and therefore automatically have the desired stability properties.
To obtain (3.22), we inserted the expression for $\widetilde{u}^{n+1}$. For practical calculations, this has no benefits, since $\widetilde{u}^{n+1}$ is stored as interim value anyway. However, with this reformulation we recognize exactly Heun's method (3.20b). In fact, Heun's method represents a SSP-RK2 scheme. This also implies that it is second order. Since we are only going to have second-order discretizations in space in this thesis, this is sufficient for our purposes. It can be proved that $\mathcal{C} = 1$ is the optimal constant for Heun's method [GKS11].

Summing up, our SSP-RK2 scheme combines three advantages: It is explicit, so that no nonlinear system has to be solved, it is of second order, and it is stable under the time step restriction of the forward Euler method.

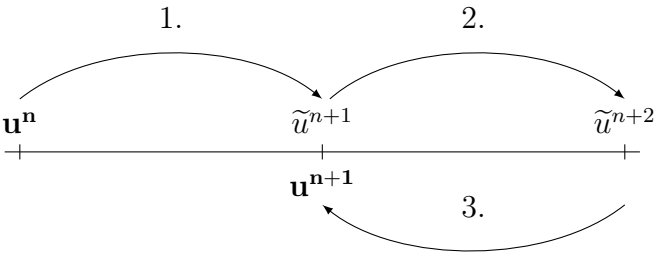


Figure 3.6: *Visualization of the SSP-RK2 method.*

4 Limiting

4.1 Motivation and state of art

Motivation. The previous chapter discussed the baseline Galerkin method. Unfortunately, practical experience has taught that this approach often leads to poor and unphysical results like spurious oscillations and false negative values. In particular, the problems arise for hyperbolic PDEs with nonsmooth solutions.

To obtain good results, we need stabilization/limiting techniques with respect to the space discretization and, in the case of a time-dependent problem, a suitable time stepping method. The two words stabilization and limiting describe a similar goal but take different perspectives. Stabilization is the more general term. It aims to achieve a stable and robust algorithm. Limiting emphasizes more the intention to keep the numerical results within a certain domain, i.e., to preserve physical bounds. For example, any volume fraction and any concentration must stay within the interval $[0, 1]$ [KH15]. The probability density function ψ, which is the central quantity in this work, is not limited upwards since it might, for example, be a delta distribution. Its invariant domain is the set of non-negative states over which integration yields the value one, see (2.3).

Stabilization requires at least L^2-stability, while limiting ensures L^∞-stability, i.e., existing minimums and maximums are not exceeded. Further considerations can be found in [KH23, Sec. 7.3]. If only bounds had to be preserved, clipping unphysical values would be enough, while conservation of mass is already ensured by the baseline Galerkin method. Designing a methodology that guarantees boundedness in a mass-conservative manner is the challenge.

State of art. In recent decades, the beginnings go back to the 1980s, a wide variety of stabilization methods have been developed. The traditional approach are variational stabilization techniques, where modifications such as the addition of further terms, are already carried out on the level of the variational formulation. The streamline upwind/Petrov-Galerkin (SUPG) algorithm, [BH82, Don18], and its numerous extensions, like the Galerkin/Least-Squares (GLS) stabilization methods, are prototypical. A disadvantage of this class of methods is the presence of a free parameter. Furthermore, non-physical oscillations are damped, but strict criteria such as the maximum principles are not guaranteed [LKSM17].

Two important families of high-resolution schemes are the flux corrected transport (FCT) and the total variation diminishing (TVD) schemes. The FCT methodology was introduced in the early 1970s [BB73]. A further development was given by Zalesak's limiter [Zal79] that applies finite volume and finite different schemes on structured meshes. Using FCT, we first compute a low-order solution, which is

then corrected by adding a limited antidiffusive flux in order to satisfy local maximum principles. Thus, an FCT scheme is a fractional step method consisting of a low-order predictor and a bound-preserving high-order corrector.
An advantage of FCT is the high flexibility in the choice of the numerical methods. A disadvantage is the splitting error depending on Δt. In addition, Δt generally has to be very small for accuracy [KH23]. Another problem becomes apparent when solving steady-state problems with FCT-like limiters. It is then a common approach to introduce pseudo time steps that serve as relaxation parameters for the iterative solver. This might cause severe convergence problems, since the supposedly stationary solution does not remain unchanged by the additional pseudo-time step [KH23]. This problem is a strong motivation to favor monolithic approaches, as we do in this thesis.

Another classical approach is the family of TVD schemes. Their origin can be traced back to [Har84]. TVD schemes are monolithic and work in an algebraic way. While FCT schemes are always nonlinear, TVD schemes such as the first-order upwind scheme can be linear. Examples of second-order TVD limiters that can be applied to nonlinear problems are the so-called minmod, van Leer or superbee limiter [KH23]. While FCT-like predictor-corrector approaches can suffer from a lack of monotonicity and therefore are not necessarily non-oscillatory [Kuz20], each TVD method is monotonicity preserving [KH23]. The TVD approaches are restricted to one dimension first of all. A simple extension of existing schemes to more dimensions runs the risk to show unexpected behavior [Kuz10].

By combining the best features of the FCT and the TVD methods, a unified algebraic flux correction (AFC) paradigm has been developed. The AFC framework is presented in Section 4.2. Before that, we introduce some basic terms and concepts.

4.1.1 Basic properties

In what follows, we define a number of features for a good limiting algorithm. A scheme is local extremum preserving if no new extrema arise.

Definition 4.1 (LED). *[Kuz10, Def. 3.17] A semi-discrete scheme is called local extremum diminishing (LED) if a maximum $u_i = \max_j u_j$ cannot increase, i.e.,*

$$u_i \geq u_j \quad \forall\, j \in \mathcal{N}_i \quad \Longrightarrow \quad \frac{\mathrm{d}u_i}{\mathrm{d}t} \leq 0,$$

and, in the same way, a minimum $u_i = \min_j u_j$ cannot decrease, i.e.,

$$u_i \leq u_j \quad \forall\, j \in \mathcal{N}_i \quad \Longrightarrow \quad \frac{\mathrm{d}u_i}{\mathrm{d}t} \geq 0.$$

A weaker requirement is given by the next definition.

Definition 4.2 (Positivity preservation). *[Kuz10, Def. 3.18] We call a semi-discrete scheme positive or positivity preserving if*

$$u_i(0) \geq 0 \quad \forall\, i \quad \Longrightarrow \quad u_i(t) \geq 0 \quad \forall\, i,\, \forall\, t > 0.$$

Closely related to positivity preservation is the notion of invariant domain preservation (IDP).

Definition 4.3 (IDP). *[KH23] Let $\mathcal{G}$ be a convex invariant set such that for the initial value it holds $u(\mathbf{x}, 0) = u_0(\mathbf{x}) \in \mathcal{G}\ \forall\, \mathbf{x} \in \mathcal{G}$. If the numerical approximation u_h stays in $\mathcal{G}$, i.e.,*

$$u_h(\mathbf{x}, t) \in \mathcal{G} \quad \forall\, \mathbf{x} \in \Omega,\, t > 0,$$

the respective scheme is called invariant domain preserving.

If IDP schemes are combined with non-negativity constraints, the terms positivity preserving and invariant domain preserving are equivalent. A whole family of methods introduced in the previous paragraph is named after its feature of being total variation diminishing (TVD). The definition reads:

Definition 4.4 (TVD). *[KH23] A semi-discrete scheme is called total variation diminishing if the total variation of the numerical solution u_h, i.e.,*

$$TV(u_h(t)) := \sum_i |u_{i+1}(t) - u_i(t)|$$

remains the same or decreases, i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}t} TV(u_h(t)) \geq 0 \quad \forall\, t > 0.$$

Any TVD method is monotonicity preserving [LeV92, Th. 15.3], i.e., if $u(0)$ is monotone, $u(\cdot, t)$ is monotone. All of these requirements for a discrete solution only make sense if they are also met for the analytical solution.

4.1.2 Maximum principles

Maximum principles both characterize the analytical solution of a PDE and establish a theoretical background to realize desired properties of the numerical solution. We start on the continuous level.

Definition 4.5 (Continuous maximum principle). *Let Σ be the set of points, where initial and boundary conditions of a PDE with solution u are prescribed. Then a general maximum principle states that*

$$\min_{\Sigma} u \leq u \leq \max_{\Sigma} u.$$

This weak maximum principle says that the extremum of a solution u is reached on Σ, while a strong maximum principle implies that the extremum can only be reached at points beyond Σ if the function u is constant. Maximum principles are a powerful tool. Even if the exact solution is unknown, far-reaching statements can be made about its properties, such as uniqueness. Moreover, a priori bounds obtained by maximum principles provide physical constraints [KH15].

We now switch to a discrete maximum principle that gives useful design criteria for the development of numerical methods.

Theorem 4.6 (Discrete maximum principle (DMP))**.** *Consider a semi-discrete scheme of the form* $M\frac{\mathrm{d}u}{\mathrm{d}t} = Qu$*. Suppose that*

$$i)\ m_{ii} > 0 \qquad ii)\ m_{ij} = 0\ \forall\, i \neq j \qquad iii)\ q_{ij} \geq 0\ \forall\, i \neq j \qquad iv)\ \sum_j q_{ij} = 0.$$

Then the scheme is LED.

Proof. Since $\sum_j q_{ij} = 0$ with the same reasoning as in (4.8c), we can write

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j \in \mathcal{N}_i^*} q_{ij}\,(u_j - u_i)$$

and with the assumptions i)-iii) it follows that

$$\frac{\mathrm{d}u_i}{\mathrm{d}t} = \underbrace{\frac{1}{m_i}}_{>0} \sum_{j \in \mathcal{N}_i^*} \underbrace{q_{ij}}_{\geq 0} (u_j - u_i) \quad \begin{cases} \geq 0 & \text{if } u_i \text{ is a local minimum,} \\ \leq 0 & \text{if } u_i \text{ is a local maximum.} \end{cases}$$

□

Remark (Numerous variants of maximum principles)**.** The setting in Theorem 4.6 is perfect for the linear advection equation in the next section. However, there exists a wide range of maximum principles. A variety of continuous maximum principles for elliptic and parabolic equations can be found in [Eva10]. In [Kuz10] and [Loh19], maximum principles are extended to first-order hyperbolic PDEs. Other more specific discrete maximum principles are given in [Kuz12, Loh19, KH23]. For example, [Kuz10, Th. 3.24] adds a reactive term to the semi-discrete equation given in Theorem 4.6. A further distinction is made between local and global DMPs. A local DMP, on the one hand, gives a range for the numerical solution at single node and this range depends on the locally neighboring nodes. A global DMP, on the other hand, gives global lower and upper bounds for all nodes and these bounds depend on the PDE's boundary values.

4.2 Algebraic flux correction

Using algebraic flux correction (AFC), first the spatial discretization is modified to a low-order physics-compatible low-order approximation. Then limited fluxes are added to recover a higher-order discretization without loosing the desired properties.

While traditional stabilization approaches modify or extend the bilinear form, the AFC methodology derives all necessary information directly from the finite element matrices. All modifications are made at this level, that is, AFC works in a 'black-box manner' [KH23]. This is a great strength, since it simplifies the implementation in existing code. Furthermore, because of its algebraic nature, AFC can be applied regardless of the mesh and, hence, regardless of the dimension of the problem.

The term 'algebraic flux correction' was first introduced in [KLT05]. Continuations, also dedicated to the analysis of AFC, can be found, e.g., in [Kuz12, BJK16, BJKR18, Loh19, Haj22, KH23].

At least for the baseline AFC methodology, the restriction to (multi-)linear Lagrange basis functions is necessary. The associated properties are collected in Section 3.3.2.

According to Theorem 4.6i), it is essential that $m_i > 0$. While this is true for (multi-)linear Lagrange functions, it might already be violated for quadratic basis functions, which no longer satisfy the positivity property [Haj22, Rem. 3.9].
Furthermore, the degrees of freedom of the (multi-)linear basis functions are their nodal values. The DMP in Section 3.3.2 states that the continuous FE interpolant u_h is bounded by these nodal values. Consequently, limiting the nodal values transfers to the whole solution, so that overshoots and undershoots do not emerge at all [LKSM17, Loh19].

In what follows, we introduce the individual components of AFC one after the other.

4.2.1 Low-order method

We consider a linear homogeneous scalar advection equation as a simple and prototypical hyperbolic PDE. Its continuous form reads

$$\frac{\partial u}{\partial t} + \nabla \cdot (\mathbf{v}u) \;=\; 0 \quad \text{in } \Omega \times (0,T].$$

Discretizing with the standard Galerkin approach leads to the semi-discrete equation

$$M_C\dot{u} + Ku = 0 \quad \Longleftrightarrow \quad M_C\dot{u} = -Ku, \tag{4.1}$$

where M_C is the consistent mass matrix and K is the convection matrix, see Section 3.3.2. Solving equation (4.1) numerically raises stability problems. To prevent this, we make sure that the discrete maximum principle given by Theorem 4.6 applies by using the LED and positive low-order approach

$$M_L\dot{u} = (D - K)u. \tag{4.2}$$

The consistent mass matrix M_C is replaced by the lumped matrix M_L and K is supplemented by the artificial diffusion matrix D that is defined by

$$D = (d_{ij})_{i,j=1}^N, \quad d_{ij} := \begin{cases} \max(k_{ij}, 0, k_{ji}), & i \neq j, \\ -\sum_{k\neq i} d_{ik}, & i = j. \end{cases} \tag{4.3}$$

The first two requirements of Theorem 4.6, i.e., $m_{ii} > 0$ and $m_{ij} = 0$, are fulfilled for the lumped mass matrix M_L by definition (3.11b) and due to Lemma 3.6a).

The other two requirements are satisfied by adding D to the matrix K. We first consider the non-diagonal elements. For $k_{ij} \leq 0$, $i \neq j$, Theorem 4.6iii) is satisfied in advance, so that the corresponding d_{ij} are simply set to zero. For $k_{ij} > 0$ we only change as much as necessary by setting d_{ij} to k_{ij}. Hence, $d_{ij} = \max(0, k_{ij})$. In the same way, we find that $d_{ji} = \max(k_{ji}, 0)$. Combining these demands to $d_{ij} = \max(k_{ij}, 0, k_{ji}) = d_{ji}$, matrix D is symmetric.
The last requirement is that the row sums of the matrix $D - K$ are zero. For matrix K this is always true, so that we can simply define the diagonal elements of matrix D as the negative sum of the other elements in the respective row.

Remarks 4.7 (Artificial diffusion operator)**.**

a) (Sparsity pattern.) Matrix D adopts the sparsity pattern of the FE matrices M_C and K, since its elements satisfy that $d_{ij} = 0 \quad \forall\, j \notin \mathcal{N}_i$.

b) (Alternative definition.) The non-diagonal elements of D can also be defined by $d_{ij} = \max(|k_{ij}|, |k_{ji}|)$ [KH23, Sec. 3.1.6]. This is equivalent to definition (4.3) if matrix K is skew-symmetric.

c) (Optimality.) We change only as much of the convection matrix K as necessary by adding the matrix D. In fact, the matrix D has the smallest possible Frobenius norm $\|\cdot\|_F$ for the desired properties [Loh19].

d) (Symmetry not obligatory.) Our artificial diffusion matrix D is a diffusion operator due to its symmetry and the zero row sum. However, a non-symmetric matrix D can produce less diffusive bound-preserving schemes and it is admissible as well [Loh19].

4.2.2 High-order schemes

The low-order method (4.2) offers a remedy for unphysical numerical results but at the same time it is extremely diffusive, that is, strong smearing effects are observed. We illustrate this with some numerical studies in Section 4.4. The diffusive behavior can be cured and the order of convergence can be increased again by adding the antidiffusive term

$$f(u) = (M_L - M_C)\dot{u} - Du \tag{4.4}$$

to the right hand side, so that the approach reads

$$M_L\dot{u} = (D - K)u + f(u). \tag{4.5}$$

From a purely algebraic point of view, the baseline Galerkin method (4.1) and the unlimited high-order method (4.5) are equivalent. Differences between the two schemes arise because an approximation to $\dot{u}$ is needed to calculate $f(u)$. A high-order approximation for the time derivative $\dot{u}$ is defined by

$$M_C\dot{u}^H = -Ku.$$

Inverting M_C to calculate $\dot{u}^H$ numerically may cause instabilities. Therefore, in this thesis we prefer the low-order approximation

$$M_L\dot{u}^L = (D - K)u.$$

It has been proven to be sufficiently accurate [Loh19].

Overall, the high-order method (4.5) is appropriate in regions where the solution is smooth, while the low-order scheme (4.2) is required for other data. The challenge is to switch between these two approaches in an adequate way. This can be achieved by modifying (4.5), so that only an adapted amount $f^*(u)$ of compensating antidiffusion is added. The resulting limited high-order method

$$M_L\dot{u} = (D - K)u + f^*(u), \tag{4.6}$$

combines the advantages of a high-order method with these of a stable bound preserving scheme. Table 4.1 gives an overview of the basic schemes presented so far.

baseline Galerkin	$M_C\dot{u} = \quad - K\ u$	(4.1)
low-order	$M_L\dot{u} = (D-K)u$	(4.2)
unlimited high-order	$M_L\dot{u} = (D-K)u + f(u)$	(4.5) $\approx$ (4.1)
limited high-order	$M_L\dot{u} = (D-K)u + f^*(u)$	(4.6)

Table 4.1: *There are differently advanced levels of limiting. We distinguish between three basic stages, where the limited high-order method is the most sophisticated one.*

The crucial point is how to choose $f^*(u)$. While the low-order method and the unlimited high-order method arise in a relatively obvious way, there are numerous ways to realize a limited high-order scheme. Several schemes are designed based on a high-order discretization, which is constrained when necessary, and where the constraints can be reversed when they are no longer needed. In Section 4.3 we will present the monolithic convex limiting (MCL) that was introduced in [Kuz20] as one option to pursue AFC.

4.2.3 Flux decomposition and antidiffusive fluxes

Sparsity pattern and notation. In Section 3.3.2, we discussed the choice of different basis functions, which determine the adjacency of the mesh nodes and consequently the sparsity pattern of the FE matrices. Since the chosen Lagrangian basis functions have a compact support, our FE matrices are sparse. Specifically, an FE matrix $A = (a_{ij})_{i,j=1}^N$ satisfies

$$a_{ij} = 0 \quad \text{if } j \notin \mathcal{N}_i. \tag{4.7a}$$

In what follows, we write the i^{th} equation, $i \in \{1, \dots, N\}$, of the each algebraic system in a componentwise manner. Using (4.7a), the sums in a (semi-)discrete equation associated with a node i reduce to

$$\sum_{j=1}^{N} a_{ij} = \sum_{j \in \mathcal{N}_i} a_{ij}. \tag{4.7b}$$

The sparsity pattern is essential both from a theoretical point of view and for the performance of the code. In practice, we take advantage of the sparse structure by using suitable data structures and storage techniques, see Appendix 10.3.

Matrix-vector multiplication for sparse FE matrices. For the convection matrix K, for instance, we have

$$\begin{aligned}(Ku)_i &= \sum_{j \in \mathcal{N}_i} k_{ij} u_j \\ &= \sum_{j \in \mathcal{N}_i^*} k_{ij} u_j + k_{ii} u_i\end{aligned}$$

$$= \sum_{j \in \mathcal{N}_i^*} k_{ij}(u_j - u_i) + \sum_{j \in \mathcal{N}_i^*} k_{ij} u_i + k_{ii} u_i$$
$$= \sum_{j \in \mathcal{N}_i^*} k_{ij}(u_j - u_i) + u_i \sum_{j \in \mathcal{N}_i} k_{ij}. \tag{4.8a}$$

Once we have a solenoidal velocity field, by the Lemma 3.7a) we find that $\sum_{j \in \mathcal{N}_i} k_{ij} = 0$. Then, (4.8a) reduces to

$$(Ku)_i = \sum_{j \in \mathcal{N}_i^*} k_{ij}(u_j - u_i). \tag{4.8b}$$

Since the artificial diffusion matrix D has zero row sum by construction, in analogy to (4.8b) it applies that

$$(Du)_i = \sum_{j \in \mathcal{N}_i^*} d_{ij}(u_j - u_i). \tag{4.8c}$$

Flux decomposition. The mass conservation property can be reflected in a conservative flux decomposition [Kuz10]. The exchange of mass between the individual nodes then can be expressed in terms of internodal fluxes f_{ij}, where

$$f_{ji} = -f_{ij} \quad \text{and} \quad f_i = \sum_{j \in \mathcal{N}_i^*} f_{ij}.$$

Rewriting the antidiffusive term $f(u)$ given in (4.4), employing that

$$(M_L u - M_C u)_i = m_i u_i - \sum_{j \in \mathcal{N}_i} m_{ij} u_j = \sum_{j \in \mathcal{N}_i^*} m_{ij}(u_i - u_j)$$

and applying (4.8c), one obtains the raw antidiffusive flux

$$f_{ij} = \left(m_{ij} \frac{\mathrm{d}}{\mathrm{d}t} + d_{ij} \right) (u_i - u_j).$$

At this point the importance of a compact support of the basis functions is worth mentioning again. Without a compact support the numerical results would become non-physical due to fluxes between nodes that are no direct neighbors.

Limited antidiffusive fluxes. The typical approach for the limited flux f_{ij}^* reads

$$f_{ij}^* = \alpha_{ij} f_{ij}, \tag{4.9}$$

where the correction factors have to be chosen as $\alpha_{ij} \in [0, 1]$. Setting $\alpha_{ij} = 0$ we find that $f_{ij}^* = 0$, so that the low-order method (4.2) is reproduced, whereas $\alpha_{ij} = 1$ satisfies $f_{ij}^* = f_{ij}$, which corresponds to the unlimited high-order method (4.5). To obtain the best compromise the coefficients should be as large as possible as long as the solution is bound-preserving.

For reasons of mass conservation, that is, to keep $f_{ji}^* = -f_{ij}^*$, the symmetry $\alpha_{ij} = \alpha_{ji}$ is always postulated. This symmetry requirement is also necessary for existence proofs and a priori estimates given in [BJK16, Loh19]. The choice of an appropriate correction factor α_{ij} determines the method. The algorithm to calculate the α_{ij} is called flux limiter. In what follows, we derive a possible set of α_{ij} using the framework of MCL.

4.3 Monolithic convex limiting

Many modern AFC schemes are realized as MCL schemes. These schemes express the discrete solution as convex combinations of intermediate states and constrain the states to be admissible [Haj22]. The methodology was designed for nonlinear hyperbolic problems and can also be applied to systems, see [Haj22]. Basic ideas of the convex limiting methodology were introduced in [GP16] in the context of FCT algorithms. We chose the MCL approach designed in [Kuz20]. A summary can be found in [KH23, Sec. 3.2.7].

One reason why the design of property-preserving schemes is still an active field of research today is Godunov's order barrier theorem. It states that a linear, monotonicity preserving method can be at most first-order accurate [LeV92, Th. 16.1]. As a consequence, our MCL scheme is always nonlinear, even though we only apply it to linear advection equations.

The approach is designed to be invariant domain preserving and thus it is bound-preserving. For a scalar conservation laws, the invariant domain with respect to the global bounds $u^{\min}$ and $u^{\max}$ is defined by the interval

$$\mathcal{G} = [u^{\min}, u^{\max}].$$

In practice, it is not sufficient to require that the numerical solution stays in $\mathcal{G}$, because spurious oscillations can still arise. Instead, we consider the intervals

$$\mathcal{G}_i = [u_i^{\min}, u_i^{\max}],$$

where the local bounds are defined by

$$u_i^{\min} := \min_{j \in \mathcal{N}_i} u_j \qquad \text{and} \qquad u_i^{\max} := \max_{j \in \mathcal{N}_i} u_j. \tag{4.10}$$

In what follows, we first consider the low-order scheme and introduce low-order bar states $\bar{u}_{ij}$. These auxiliary states are used to argue via convexity that the scheme is IDP, assuming that a CFL-like condition holds true. In a second step, the concept is transferred to high-order schemes.

Bar states. We use the semi-discrete low-order scheme (4.2) as starting point. Applying (4.8b) and (4.8c), we obtain

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j \in \mathcal{N}_i} (d_{ij} - k_{ij}) u_j = \sum_{j \in \mathcal{N}_i^*} (d_{ij} - k_{ij})(u_j - u_i). \tag{4.11}$$

It is possible to write the equation in the bar state form already at this point, see [Haj22]. We, however, derive the expression for the low-order bar states again.

The MCL approach is combined with SSP-RK2 as time stepping scheme throughout this thesis. If its two forward Euler steps are bound preserving, then the entire method is, as discussed in Section 3.5. By applying a forward Euler step to (4.11), we switch to a fully discrete formulation. Then, we gradually rewrite this generic

forward Euler step into the bar state form. An intermediate step u_i^{SSP} for the low-order method reads [Kuz20]

$$\begin{aligned} u_i^{\text{SSP}} &= u_i + \frac{\Delta t}{m_i} \sum_{j \in \mathcal{N}_i^*} \left[d_{ij} \left(u_j - u_i\right) - k_{ij} \left(u_j - u_i\right) \right] \\ &= \left(1 - \frac{2\Delta t}{m_i} \sum_{j \in \mathcal{N}_i^*} d_{ij}\right) u_i + \frac{\Delta t}{m_i} \sum_{j \in \mathcal{N}_i^*} \left[d_{ij}(u_j + u_i) - k_{ij}(u_j + u_i) \right] \\ &= \left(1 - \frac{2\Delta t}{m_i} \sum_{j \in \mathcal{N}_i^*} d_{ij}\right) u_i + \frac{2\Delta t}{m_i} \sum_{j \in \mathcal{N}_i^*} d_{ij} \left[\frac{u_j + u_i}{2} - \frac{k_{ij}(u_j - u_i)}{2d_{ij}} \right] \\ &= \left(1 - \frac{2\Delta t}{m_i} \sum_{j \in \mathcal{N}_i^*} d_{ij}\right) u_i + \frac{2\Delta t}{m_i} \sum_{j \in \mathcal{N}_i^*} d_{ij}\, \bar{u}_{ij}, \end{aligned} \tag{4.12}$$

where the expression

$$\bar{u}_{ij} := \frac{u_j + u_i}{2} - \frac{k_{ij}(u_j - u_i)}{2d_{ij}} \tag{4.13}$$

is called low-order bar state. The low-order bar state is symmetric, that is, $\bar{u}_{ij} = \bar{u}_{ji}$, if and only if the convection matrix K is skew-symmetric. This is the case under the conditions listed in Remark 3.8a). Furthermore, using the definition of the non-diagonal elements of D, see (4.3), for the low-order bar state we find that

$$\bar{u}_{ij} = \begin{cases} u_i & \text{if } d_{ij} = k_{ij}, \\ u_j & \text{if } d_{ij} = k_{ji}, \end{cases}$$

which implies that $\bar{u}_{ij} \in \mathcal{G}$ if $u_i, u_j \in \mathcal{G}$. The next question is what can be said about the intermediate step u_i^{SSP}. The linear combination (4.12) is a convex combination if the sum of the coefficients equals one, which is given, and if additionally

$$\frac{2\Delta t}{m_i} \sum_{j \in \mathcal{N}_i^*} d_{ij} \in [0, 1] \qquad \forall\, i \in \{1, \dots, N\}. \tag{4.14a}$$

The components are all non-negative. So is their sum. Thus, the key requirement is

$$\frac{2\Delta t}{m_i} \sum_{j \in \mathcal{N}_i^*} d_{ij} \leq 1 \qquad \forall\, i \in \{1, \dots, N\}. \tag{4.14b}$$

Because of the convexity we can then conclude that $u_i^{\text{SSP}} \in \mathcal{G}$ if $u_i, \bar{u}_{ij} \in \mathcal{G}$. We emphasize that the second sum over $j \in \mathcal{N}_i^*$ in (4.12) refers to both d_{ij} and $\bar{u}_{ij}$, so that the convex combination in general consists of more than two terms. Consequently, the result can be be formulated as

$$u_i^{\text{SSP}} \in [u_i, \bar{u}_{ij}] \quad \text{considering all } \bar{u}_{ij}. \tag{4.15}$$

We express (4.14b) as an implementable CFL-like condition for Δt:

Corollary 4.8 (CFL-condition)**.** *Applying an explicit SSP-RK method to* (4.11)*, the time step restriction*

$$\Delta t = \min_{i\in\{1,\dots,N\}} \frac{m_i}{2\sum_{j\in\mathcal{N}_i^*} d_{ij}} \tag{4.16}$$

guarantees a correct functionality of the scheme. Specifically, this means that $u_i^{SSP} \in \mathcal{G}$ *if* $u_j \in \mathcal{G}\ \ \forall\, j \in \mathcal{N}_i$*.*

Proof. Formula (4.16) follows immediately from the requirement (4.14b). The summarizing statement on the IDP property combines the two above results for u_i, u_j and $\bar{u}_{ij}$, on the one hand, and (4.15) for $u_i, \bar{u}_{ij}$ and u_i^{SSP}, on the other hand. □

Choice of correction factor. To transfer the convexity property to the limited high-order method, we consider the semi-discrete formulation (4.6), that is,

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} = \sum_{j\in\mathcal{N}_i^*} [\, d_{ij}(u_j - u_i) - k_{ij}(u_j - u_i) + f_{ij}^* \,]. \tag{4.17}$$

In analogy to the forward Euler step (4.12) for the low-order scheme, the bar state form derived from (4.17) reads

$$u_i^{\text{SSP}} = \left(1 - \frac{2\Delta t}{m_i}\sum_{j\in\mathcal{N}_i^*} d_{ij}\right) u_i + \frac{2\Delta t}{m_i}\sum_{j\in\mathcal{N}_i^*} d_{ij}\left(\bar{u}_{ij} + \frac{f_{ij}^*}{2d_{ij}}\right), \tag{4.18}$$

so that the high-order bar state is defined by

$$\bar{u}_{ij}^* := \bar{u}_{ij} + \frac{f_{ij}^*}{2d_{ij}}. \tag{4.19}$$

The forward Euler step described by (4.18) provably provides the IDP property if the corresponding high-order bar states $\bar{u}_{ij}^*$ stay in a convex invariant set $\mathcal{G}$. Therefore, our task is to ensure that $\bar{u}_{ij}^* \in \mathcal{G}$ if $u_j \in \mathcal{G}\ \ \forall j \in \mathcal{N}_i$. Specifically, the validity of the local maximum principle must be assured, that is,

$$u_i^{\min} \le \bar{u}_{ij}^* \le u_i^{\max}. \tag{4.20}$$

Taking into account that $f_{ji} = -f_{ij}$, four inequality constraints are available:

$$\begin{aligned} u_i^{\min} &\overset{\text{a)}}{\le} \bar{u}_{ij} + \alpha_{ij}\frac{f_{ij}}{2d_{ij}} \overset{\text{b)}}{\le} u_i^{\max}, \\ u_j^{\min} &\overset{\text{c)}}{\le} \bar{u}_{ji} - \alpha_{ij}\frac{f_{ij}}{2d_{ij}} \overset{\text{d)}}{\le} u_j^{\max}. \end{aligned}$$

Case-by-case analysis leads to

$$\text{a)} \iff \alpha_{ij} f_{ij} \ge 2d_{ij}(u_i^{\min} - \bar{u}_{ij}) \implies \begin{cases} \alpha_{ij} \ge \frac{2d_{ij}(u_i^{\min} - \bar{u}_{ij})}{f_{ij}}, & \text{if } f_{ij} > 0, \\ \alpha_{ij} \le \quad \% \quad , & \text{if } f_{ij} < 0, \ (\star) \end{cases}$$

$$\text{b)} \quad \Longleftrightarrow \quad \alpha_{ij} f_{ij} \le 2d_{ij}(u_i^{\max} - \bar{u}_{ij}) \quad \Longrightarrow \begin{cases} \alpha_{ij} \le \frac{2d_{ij}(u_i^{\max} - \bar{u}_{ij})}{f_{ij}}, & \text{if } f_{ij} > 0, \ (\star) \\ \alpha_{ij} \ge \quad \% \quad , & \text{if } f_{ij} < 0, \end{cases}$$

$$\text{c)} \quad \Longleftrightarrow \quad \alpha_{ij} f_{ij} \le 2d_{ij}(\bar{u}_{ij} - u_j^{\min}) \quad \Longrightarrow \begin{cases} \alpha_{ij} \le \frac{2d_{ij}(\bar{u}_{ij} - u_j^{\min})}{f_{ij}}, & \text{if } f_{ij} > 0, \ (\star) \\ \alpha_{ij} \ge \quad \% \quad , & \text{if } f_{ij} < 0, \end{cases}$$

$$\text{d)} \quad \Longleftrightarrow \quad \alpha_{ij} f_{ij} \ge 2d_{ij}((\bar{u}_{ij} - u_j^{\max})) \Longrightarrow \begin{cases} \alpha_{ij} \ge \frac{2d_{ij}(\bar{u}_{ij} - u_j^{\max})}{f_{ij}}, & \text{if } f_{ij} > 0, \\ \alpha_{ij} \le \quad \% \quad , & \text{if } f_{ij} < 0. \ (\star) \end{cases}$$

For $f_{ij} = 0$, the correction factor α_{ij} is arbitrary. Otherwise, the α_{ij} are always assumed to be positive. We choose them as close as possible to one without violating the DMP constraints. The four non-trivial inequalities, marked with $(\star)$, are combined to

$$\widetilde{\alpha}_{ij} = \begin{cases} \min\left\{\frac{2d_{ij}(u_i^{\max} - \bar{u}_{ij})}{f_{ij}}, \frac{2d_{ij}(\bar{u}_{ji} - u_j^{\min})}{f_{ij}}\right\}, & \text{if } f_{ij} > 0, \\ \min\left\{\frac{2d_{ij}(u_i^{\min} - \bar{u}_{ij})}{f_{ij}}, \frac{2d_{ij}(\bar{u}_{ji} - u_j^{\max})}{f_{ij}}\right\}, & \text{if } f_{ij} < 0. \end{cases}$$

The notation $\widetilde{\alpha}_{ij}$ indicates that a further requirement is necessary to ensure the upper limit for α_{ij} as well, that is,

$$\alpha_{ij} = \min\{1, \widetilde{\alpha}_{ij}\}.$$

Summing up, for the limited antidiffusive flux $f_{ij}^* = \alpha_{ij} f_{ij}$ we end up with

$$\begin{aligned} f_{ij}^* &= \begin{cases} \min\{f_{ij},\ \min\{2d_{ij}(u_i^{\max} - \bar{u}_{ij}),\ 2d_{ij}(\bar{u}_{ji} - u_j^{\min})\}\}, & \text{if } f_{ij} > 0, \\ \max\{f_{ij},\ \max\{2d_{ij}(u_i^{\min} - \bar{u}_{ij}),\ 2d_{ij}(\bar{u}_{ji} - u_j^{\max})\}\}, & \text{if } f_{ij} < 0, \end{cases} \\ &= \begin{cases} \min\{f_{ij},\ \min\{2d_{ij}u_i^{\max} - \bar{\omega}_{ij},\ \bar{\omega}_{ji} - 2d_{ij}u_j^{\min}\}\}, & \text{if } f_{ij} > 0, \\ \max\{f_{ij},\ \max\{2d_{ij}u_i^{\min} - \bar{\omega}_{ij},\ \bar{\omega}_{ji} - 2d_{ij}u_j^{\max}\}\}, & \text{if } f_{ij} < 0, \end{cases} \end{aligned} \tag{4.21}$$

where the original bar state is reformulated with the scaled bar state

$$\bar{\omega}_{ij} = 2d_{ij}\bar{u}_{ij} = d_{ij}(u_j + u_i) - k_{ij}(u_j - u_i). \tag{4.22}$$

Since the local bounds and the correction factors depend on the discrete solution, the MCL scheme is obviously nonlinear.

Practical implementation. The limited antidiffusive fluxes f_{ij}^* are implemented using (4.21). The correction factors α_{ij} are just a tool to derive the formula. Making use of the scaled bar states $\bar{\omega}_{ij}$, the division by d_{ij} is avoided, so that a possible source of numerical instabilities is eliminated.
The quantities $f_{ij}^*, f_{ij}, \bar{\omega}_{ij}$ and $\bar{\omega}_{ji}$ have the same sparsity pattern as the FE matrices, that is, the ratio of non-zero elements is quite small. Hence, we determine only the non-zero values as individual scalars. Finally, the sums $\sum_{j \in \mathcal{N}_i^*} f_{ij}^*$ have to be calculated, see (4.17).

4.3.1 Extension to convection-diffusion equations

The MCL approach presented above was originally designed for purely hyperbolic problems. The question arises how to handle convection-diffusion equations such as the space-independent Fokker-Planck equation. We examine whether and to which extent MCL can be transferred.[1]

Straightforward extension. Let us consider a linear scalar convection-diffusion equation with isotropic diffusion, that is,

$$\frac{\partial u}{\partial t} + \nabla \cdot (\mathbf{v}u) - \varepsilon \Delta u = 0 \qquad \text{in } \Omega \times (0, T].$$

For the semi-discrete equation, the low-order approximation is

$$M_L \dot{u} + (K - D)u + \varepsilon L u = 0,$$

where $D = (d_{ij})_{i,j=1}^N$ represents the artificial diffusion and $L = (l_{ij})_{i,j=1}^N$ the physical diffusion. We combine the artificial and the physical diffusion in one expression

$$d_{ij}^{(\text{comb})} := d_{ij} - \varepsilon l_{ij},$$

since the resulting matrix still satisfies

$$\begin{aligned} d_{ij}^{(\text{comb})} &= d_{ji}^{(\text{comb})} && \text{(symmetry)} \\ \text{and} \quad \sum_{j \in \mathcal{N}_i} d_{ij}^{(\text{comb})} &= 0 && \text{(zero row sum)}. \end{aligned}$$

The low-order bar state then reads

$$\bar{u}_{ij}^{(\text{comb})} = \frac{u_j + u_i}{2} - \frac{k_{ij}(u_j - u_i)}{2 d_{ij}^{(\text{comb})}}.$$

If we assume that matrix L is a Z-matrix, that is,

$$l_{ij} \leq 0 \quad \forall\, i \in \{1, \dots, N\},\ j \in \mathcal{N}_i^*,$$

see [Loh19, Def. 4.1], then

$$\left| \frac{k_{ij}}{d_{ij}^{(\text{comb})}} \right| \leq 1$$

and the extended bar state $\bar{u}_{ij}^{(\text{comb})}$ is a convex combination of u_i and u_j. The Z-matrix property could be confirmed for our stiffness matrix on the sphere. It is a feasible condition.

However, the adapted CFL-condition similar to (4.14b) reads

$$\frac{2\Delta t}{m_i} \sum_{j \in \mathcal{N}_i^*} d_{ij}^{(\text{comb})} \leq 1 \qquad \forall\, i \in \{1, \dots, N\}.$$

[1] Thank you to Dr. Hennes Hajduk and Paul Moujaes for discussing this issue with me.

Since this condition does not only depend on the artificial diffusion introduced for limiting purposes but also on the physical diffusion, which is part of the original PDE, the adapted CFL-condition becomes extremely restrictive and leads to prohibitively small time steps Δt if the diffusion coefficient becomes larger. Since in our application the diffusion coefficient is small, however, we stick to this approach for this thesis.

Strang splitting. An alternative is to split the equation into its hyperbolic and its parabolic part, for example, with Strang splitting, see (6.9). Using Strang splitting, we first solve the advection equation for $\mathcal{L}_1 = \nabla \cdot (\mathbf{v}u)$, then the heat equation with $\mathcal{L}_2 = -\varepsilon \Delta u$ and finally the advection equation with operator $\mathcal{L}_1$ again.
Further suggestions for maximum principle preserving discretizations of scalar convection-diffusion equations can be found, e.g., in [JA21, QdLK22].

4.4 Numerical studies

4.4.1 Setting

We compare numerical results for various time stepping and limiting schemes for the case of pure advection, that is,

$$\frac{\partial u}{\partial t} + \nabla \cdot (\mathbf{v}u) = 0 \quad \text{on } \Omega \times (0, T].$$

The square two-dimensional domain $\Omega = [0, 1]^2$ is used as computational area and discretized with $2^l \times 2^l$ squares on level l.

Choice of velocity field. For the homogeneous advection equation we specify a velocity field $\mathbf{v}$ to construct a test case, where an exact reference solution u is available. A simple constant velocity field $\mathbf{v}_t = (c_1, c_2)^T$ describes a translation. Since it is easy to follow the movement, an exact solution can be found easily at any time. This still applies for the solenoidal space-dependent field $\mathbf{v}_r = (0.5 - y, x - 0.5)^T$, which we choose in what follows. The field $\mathbf{v}_r$ describes a counterclockwise rotation around the center of the square, so that the initial state is expected to be restored after one complete revolution.

Rotational velocity fields describing a circular convection might be written in a more systematic way as $\mathbf{v}_r = \omega(y_0 - y, x - x_0)^T$, where $(x_0 \,|\, y_0)$ is the axis of rotation, and ω is the (constant) angular velocity in rad/sec [Zal79]. Consequently, for $\omega = 1$ a full rotation cycle is completed at $T = 2\pi$, so that $u(t) = u(t + 2\pi)$ if the velocity at the inflow boundary is zero, see Figure 4.2.

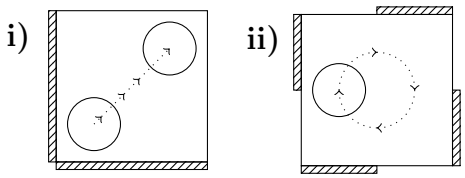


Figure 4.2: *i) Translation and ii) rotation, where the shaded areas mark the inflow boundaries.*

The $\mathbf{v}_r$ describes a so-called solid body rotation as it was introduced in [Zal79] and extended in [LeV96]. Solid body movements do not change the shape of the

transported geometry. Consequently, with our setup not only the general quality of the simulation results can be quantified, but additionally the methods ability to preserve the shape of the moved geometry is checked.
Time-dependent flow fields, which deform the geometry during the transport are considered in Section 5.6.3. For such fields it has to be ensured that the initial configuration is restored after a certain period of time.

Initial configuration. As configuration to be rotated we consider a smooth as well as a discontinuous configuration. The smooth configuration we choose is a 'cosine bell', see, e.g., [LSPT12]. It is defined by

$$u_0(r) := \begin{cases} \frac{h_{\max}}{2}\left(1+\cos(\frac{\pi r}{R})\right), & \text{if } r \le R, \\ 0\,, & \text{otherwise,} \end{cases}$$

where

$$r = r(x,y) := \sqrt{(x-m_1)^2+(y-m_2)^2}.$$

Parameter $h_{\max}$ describes the maximal height of the cosine bell reached in its midpoint (m_1, m_2), so that $u_0 \in [0, h_{\max}]$. The geometry has radius R.
For our overall discontinuous initial condition we select a configuration consisting of three solid bodies representing different degrees of smoothness: a smooth hump constructed as a cosine bell as well, a sharp cone and finally a slotted cylinder, see, e.g., [Kuz10]. Within the square domain Ω, these three geometries are each defined in a circle of radius $R = 0.15$. For $i \in \{SlCy, cone, hump\}$ we obtain

$$u_0(r_i) := \begin{cases} 1\,, & \text{if } r_{SlCy} \le R; \|x - m_1^{cone}\| \ge 0.025 \text{ or } y \ge 0.85, \\ 1 - r_{cone}\,, & \text{if } r_{cone} \le R, \\ \frac{h_{\max}}{2}\left(1+\cos(\frac{\pi r_{hump}}{R})\right), & \text{if } r_{hump} \le R, \\ 0\,, & \text{otherwise.} \end{cases}$$

The values r_i are not constant but space-dependent, that is,

$$r_i = r_i(x,y) := \sqrt{(x-m_1^i)^2+(y-m_2^i)^2}, \quad i \in \{SlCy, cone, hump\},$$

where the individual midpoints are defined by

$$(m_1^i, m_2^i) = \begin{cases} (0.5, 0.75) & \text{if } i = SlCy, \\ (0.5, 0.25) & \text{if } i = cone, \\ (0.25, 0.5) & \text{if } i = hump. \end{cases}$$

The remaining parameter $h_{\max}$ is set to 0.5. Thus, overall the initial values lie in the interval $[0, 1]$. With an adequate choice of the parameters we have to ensure that the distance of the configuration to the boundary of Ω stays big enough, so that neither a geometry collides with a boundary during the movement nor a diffusive behavior causes any trouble. Actually, for hyperbolic problems, boundary conditions have

to be specified at the inflow [KH15, p.13]. We handle this implicitly by setting all boundary values to zero.

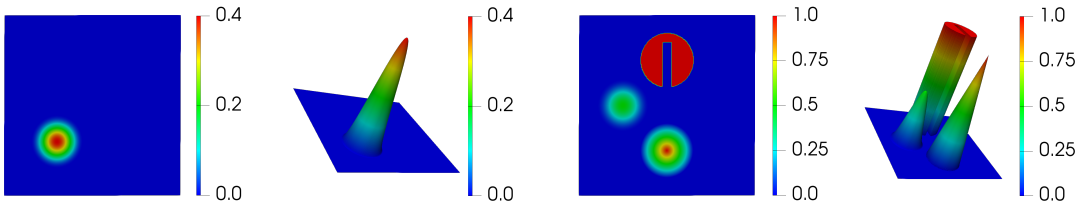


Figure 4.3: *Smooth initial configuration (left) and mixed initial configuration (right).*

Experimental order of convergence. Here and subsequently, the experimental order of convergence (EOC) for the level with mesh size $\frac{h}{2}$ is calculated by

$$\mathrm{EOC} = \frac{1}{\log(2)} \log\left(\frac{e_h}{e_{h/2}}\right) = \log_2\left(\frac{e_h}{e_{h/2}}\right),$$

where $e_h = \|u_h - u\|_{L^2(\Omega)}$ and $e_{h/2} = \|u_{h/2} - u\|_{L^2(\Omega)}$ are errors for two consecutive levels measuring the L^2-error between the reference solution u and the numerical solutions u_h and $u_{h/2}$, respectively.

4.4.2 Numerical results

For the spatial discretization bilinear Lagrange basis functions are used. We apply seven test cases to both initial configurations. First, the baseline Galerkin method (4.1) is combined with four different time stepping schemes, namely

i) forward Euler (3.18a),
ii) backward Euler (3.18b),
iii) Crank-Nicolson (3.18c),
iv) SSP-RK2 (3.22).

Then, we fix SSP-RK2 as time stepping scheme and use it in combination with

v) low-order method (4.2),
vi) unlimited high-order method (4.5),
vii) limited high-order method (4.6) realized as MCL-approach.

For the first four test cases linear system have to be solved. We apply the SPAI ('sparse approximate inverse') preconditioner, and use BiCGStab ('stabilized biconjugate gradients method'), an extension of the CG-method, as solver. A detailed description of these solvers can be found in [Mei15].

Smooth configuration. Our results for the pure cosine bell configuration are summarized in Table 4.4. Using the parameters $(m_1, m_2, h_{\max}, R) = (0.3, 0.3, 0.4, 0.15)$ it should apply that $u(x,t) \in [0, 0.4]$. In fact, however, overshoots and undershoots arise for the baseline Galerkin approach and the unlimited high-order method, even though we consider a smooth geometry. This vividly illustrates the need for a limiting strategy. Compared to an unsteady configuration, see Table 4.5, the violation

of the boundaries is relatively small. In some sense, there are different 'degrees of smoothness'. Replacing the parameters $(m_1, m_2, h_{\max}, R) = (0.3, 0.3, 0.4, 0.15)$ by $(m_1, m_2, h_{\max}, R) = (0.25, 0.25, 1, 0.1)$ the overshoots were, on average, one power of ten higher.

It is worth mentioning that the numerical results deteriorate as the simulation progresses. On the one hand, this is reflected in the fact that the lower boundary of the final data range is quite similar to the global minimum, i.e., to the minimum across all space and time points at the corresponding level. On the other hand, the EOC of the low-order method, for example, reads 0.24, 0.39 and 0.56 after half a rotation ($t = \pi$), which is better than the results we observe for the final time in Table 4.4v).

The EOCs for the backward Euler in Table 4.4 are relatively low, but they converge towards one. All in all, the order of convergence seems to be limited by the time stepping scheme, which is of lower order than the space discretization.

Discontinuous configuration. The configuration comprises three geometries. The convergence results and the overshoots and undershoots of this 'mixed configuration' reflect the worst case. The results are presented in Table 4.5 and Figure 4.6. Again, we see the effects of different time stepping schemes. The problems of the forward Euler method are striking. For diffusion equations it is known that the Crank-Nicolson method is unconditionally stable but may exhibit oscillations, whereas the lower-order backward Euler method is both stable and immune to oscillations. This tendency can be observed in Figure 4.6 as well.

It is worth mentioning that for the backward Euler and the Crank-Nicolson method, already starting with 150 steps on level 5 leads to reasonable results. The choice of the time steps for our SSP-RK2 scheme in combination with MCL is based on the CFL-condition (4.16). Combining the baseline Galerkin approach with SSP-RK2 is not very fruitful. If we start with 500 steps on level 5, that is even with more steps than used in combination with MCL, the solution values explode on level 7 and 8. At the same time, starting with more than 1000 steps does not significantly improve anything.

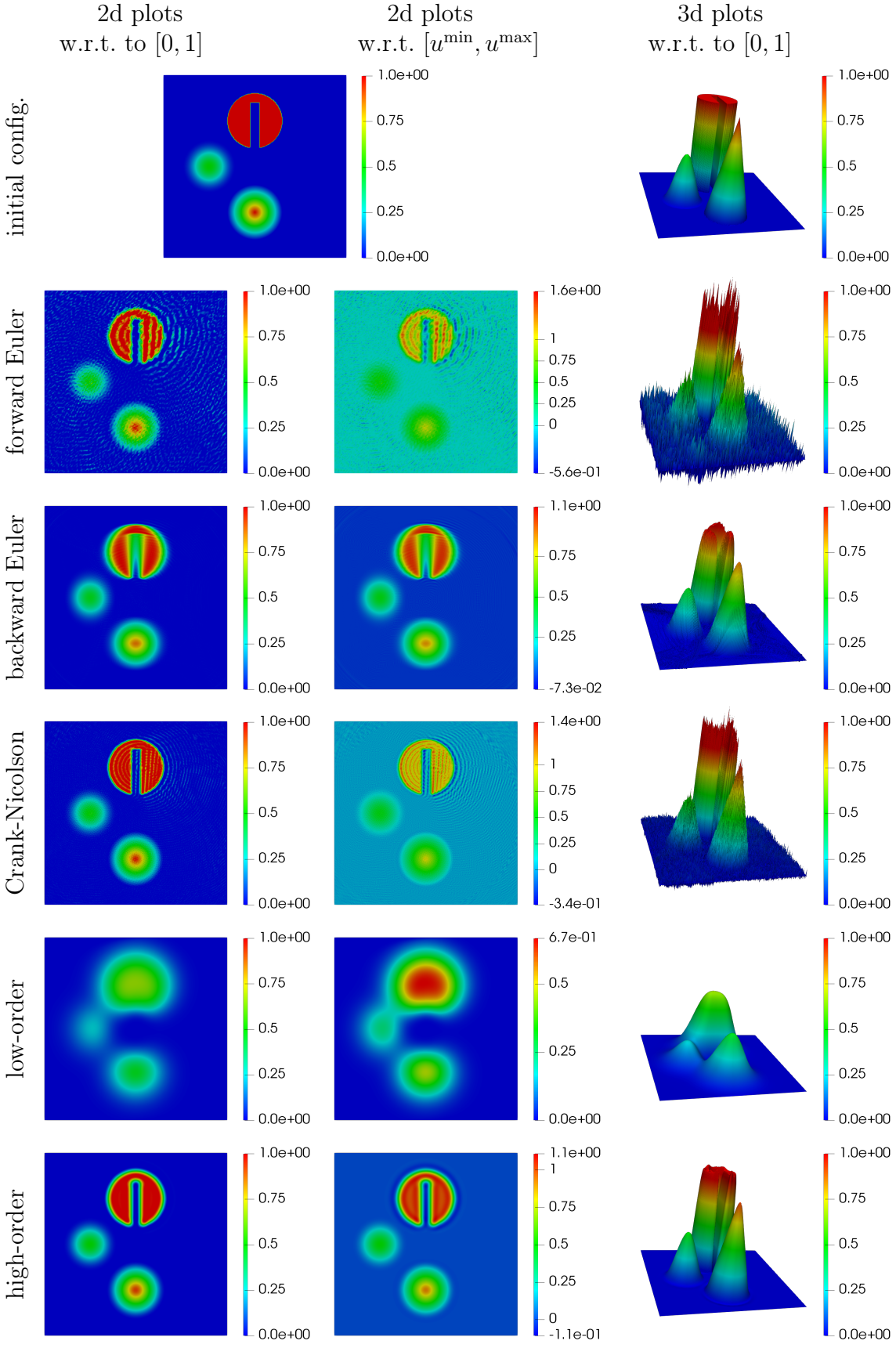
With respect to the different stages of limiting, both the diffusivity of the low-order method and the violation of boundaries in the case of the unlimited high-order method are illustrated. The results of the limited high-order method carried out using MCL and SSP-RK2, however, are satisfactory in every way. As the most sophisticated method, it is the standard approach for the later benchmarks in this thesis.

level	steps	global minimum	final data range	L^2-err	EOC
i) Galerkin (Q_1), forward Euler					
5	20000	-1.92e-2	[-1.6e-2, 0.404]	3.52e-3	-
6	40000	-6.08e-3	[-5.5e-3, 0.404]	7.53e-4	2.23
7	80000	-2.20e-3	[-2.0e-3, 0.400]	2.48e-4	1.60
8	160000	-6.16e-3	[-5.8e-3, 0.401]	2.46e-4	0.01
ii) Galerkin (Q_1), backward Euler					
5	500	-4.06e-3	[-1.83e-3, 0.243]	1.84e-2	-
6	1000	-1.14e-3	[-5.72e-4, 0.298]	1.22e-2	0.60
7	2000	-3.15e-4	[-1.75e-4, 0.341]	7.34e-3	0.73
8	4000	-9.29e-5	[-5.39e-5, 0.368]	4.13e-3	0.83
iii) Galerkin (Q_1), Crank-Nicolson					
5	500	-1.79e-2	[-1.61e-2, 0.398]	3.87e-3	-
6	1000	-5.31e-3	[-5.13e-3, 0.399]	8.22e-4	2.24
7	2000	-1.77e-3	[-1.74e-3, 0.399]	2.00e-4	2.04
8	4000	-6.15e-4	[-6.03e-4, 0.400]	5.09e-5	1.97
v) Low-order method, SSP-RK2					
5	400	0	[0, 0.032]	3.98e-2	-
6	800	0	[0, 0.062]	3.59e-2	0.15
7	1600	0	[0, 0.109]	3.03e-2	0.24
8	3200	0	[0, 0.175]	2.32e-2	0.39
vi) Unlimited high-order method, SSP-RK2					
5	400	-1.20e-2	[-1.19e-2, 0.229]	1.84e-2	-
6	800	-6.86e-3	[-6.84e-3, 0.357]	5.61e-3	1.71
7	1600	-3.19e-3	[-3.19e-3, 0.394]	1.21e-3	2.22
8	3200	-1.39e-3	[-1.38e-3, 0.399]	2.72e-4	2.15
vii) MCL, SSP-RK2					
5	400	-3.82e-19	[-1.62e-19, 0.176]	2.14e-2	-
6	800	-3.75e-19	[-3.34e-19, 0.316]	6.08e-3	1.81
7	1600	-2.40e-19	[-2.29e-19, 0.375]	1.03e-3	2.56
8	3200	-1.39e-19	[-1.30e-19, 0.392]	1.81e-4	2.51

Table 4.4: *Quantitative results for the smooth cosine bell configuration with the parameters* $(m_1, m_2, h_{\max}, R) = (0.3, 0.3, 0.4, 0.15)$ *at the final time* $T = 2\pi$.

level	steps	global minimum	final data range	L^2-err	EOC
i) Galerkin (Q_1), forward Euler					
5	25000	-0.67	[-0.47, 1.49]	1.38e-1	-
6	50000	-0.56	[-0.47, 1.58]	1.03e-1	0.42
7	100000	-0.63	[-0.56, 1.56]	9.03e-2	0.18
8	200000	-1.84	[-1.47, 1.81]	1.27e-1	-0.49
ii) Galerkin (Q_1), backward Euler					
5	500	-0.44	[-0.19, 0.85]	1.52e-1	-
6	1000	-0.42	[-0.09, 0.90]	1.21e-1	0.33
7	2000	-0.43	[-0.07, 1.00]	9.96e-2	0.28
8	4000	-0.45	[-0.07, 1.06]	8.22e-2	0.30
iii) Galerkin (Q_1), Crank-Nicolson					
5	500	-0.66	[-0.47, 1.43]	1.31e-1	-
6	1000	-0.49	[-0.44, 1.54]	9.84e-2	0.47
7	2000	-0.53	[-0.40, 1.40]	7.64e-2	0.37
8	4000	-0.49	[-0.34, 1.41]	5.66e-2	0.43
iv) Galerkin (Q_1), SSP-RK2					
5	1000	-0.65	[-0.45, 1.46]	1.34e-1	-
6	2000	-0.49	[-0.39, 1.46]	9.89e-2	0.50
7	4000	-0.52	[-0.39, 1.46]	7.60e-2	0.32
8	8000	-0.49	[-0.34, 1.43]	5.55e-2	0.45
v) Low-order method, SSP-RK2					
5	400	0	[0, 0.26]	2.25e-1	-
6	800	0	[0, 0.40]	2.06e-1	0.13
7	1600	0	[0, 0.55]	1.85e-1	0.16
8	3200	0	[0, 0.67]	1.63e-1	0.19
vi) Unlimited high-order method, SSP-RK2					
5	400	-0.14	[-0.07, 0.99]	1.59e-1	-
6	800	-0.18	[-0.06, 1.00]	1.14e-1	0.48
7	1600	-0.20	[-0.06, 1.12]	8.51e-2	0.42
8	3200	-0.20	[-0.11, 1.11]	6.25e-2	0.45
vii) MCL, SSP-RK2					
5	400	-3.93e-18	[-4.67e-25, 0.79]	1.58e-1	-
6	800	-5.47e-18	[-2.78e-19, 0.82]	1.23e-1	0.37
7	1600	-5.84e-18	[-5.60e-19, 1.00]	8.87e-1	0.47
8	3200	-7.42e-18	[-2.59e-18, 1.00]	6.29e-2	0.50

Table 4.5: *Quantitative results for the discontinuous slotted cylinder - smooth cone - sharp hump - configuration at the final time $T = 2\pi$.*


2d plots
w.r.t. to [0, 1]
2d plots
w.r.t. [u^min, u^max]
3d plots
w.r.t. to [0, 1]
initial config.
forward Euler
backward Euler
Crank-Nicolson
low-order
high-order
1.0e+00
0.75
0.5
0.25
0.0e+00
1.6e+00
1
0
-5.6e-01
1.1e+00
-7.3e-02
1.4e+00
-3.4e-01
6.7e-01
-1.1e-01

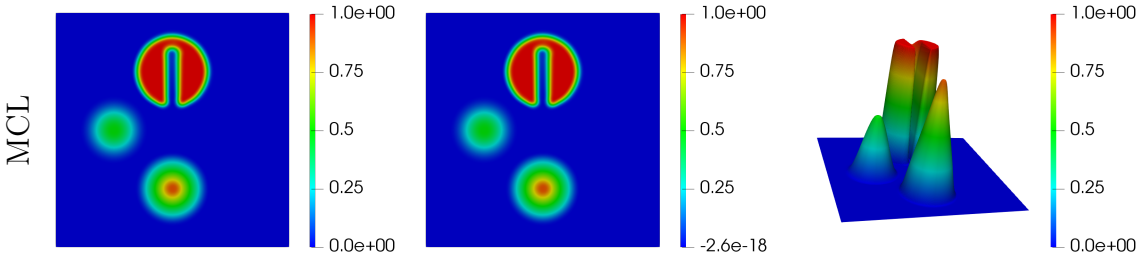


Figure 4.6: *Visual results for the discontinuous configuration after one full rotation at $T = 2\pi$. As a reference, the initial configuration is provided. The outer columns contain our 2d and 3d results plotted in the interval $[0, 1]$, while the middle column is scaled to the actual values. The results are documented for forward Euler (on level 7), backward Euler, Crank-Nicolson, low-order method, high-order method and MCL (all on level 8).*

5 PDEs on surfaces

5.1 Introduction

Applications. Numerous applications, whether in fluid dynamics, engineering or biology, require PDEs on surfaces. These surfaces and their geometry vary in complexity. They also range from those that are stationary to those that are evolving [DE13]. Atmospheric flows over the topology of the earth can be modeled with appropriate PDEs [PBR13]. In medical imagining, the human brain [Hin16] is considered. In physiology, the surfactants in the lung are considered. Chemotaxis is a biochemical phenomenon, which describes the movement of an organism due to a chemical stimulus [Ali16]. Convection and diffusion on biomembranes, which are often modeled as interfaces in multiphase flows [ES10], are of further interest.

Solution strategies. A first attempt to solve the Laplace equation on a surface was made in [Dzi88]. This was later extended to parabolic equations [DE07]. In [DE08], parabolic problems are solved using an implicit representation of the surface.

Overall, it has to be distinguished between parametric and indirect representations of curved surfaces [Kam13]. In this thesis, we use a parametric representation as direct approach [LV10] by applying a geometric mapping from a reference domain to segments of the surface. An advantage is that the surface can be interpreted as a lower dimensional manifold embedded in the surrounding space $\mathbb{R}^d$.
Indirect methods describe the surface implicitly, i.e., all the geometry is encoded in a level set function $\phi(x)$ [OF01, SS03, DE13]. The level set function defines an interface, and makes it easy to determine whether a point x lies inside ($\phi(x) < 0$) or outside ($\phi(x) > 0$) that interface. Methods based on level set functions are suitable for complex morphology, for example when the geometry splits into more than one object or has holes. They are efficient when topology changes or surfaces intersect [Kam13]. Another advantage is that special quantities such as the mean curvature can simply be calculated by taking derivatives of $\phi(x)$ [Ali16]. Conservation of topology and volume is not automatically guaranteed, however, e.g., [Kuz14] deals with solutions for this.
A disadvantage of level set methods in contrast to a parametric approach is that the implicit function is defined in a space which is one dimension higher than it has to be to describe the surface [Kam13]. This is worthwhile if we benefit from advantages listed above but wasted otherwise. In this thesis, we investigate a Fokker-Planck equation, where the geometric object of interest is the surface of the unit sphere in three-dimensional space. Thus, we have a stationary and simple geometry, which justifies our choice to apply parametric representation. Furthermore, that approach allows us to reuse a lot of the standard finite element code.

Structure of this chapter. Surfaces can be embedded in the larger concept of manifolds. Thus, we introduce the necessary mathematical fundamentals in this general framework. We consider manifolds of arbitrary dimension. However, we do not consider manifolds with boundaries, since the unit sphere does not have any. To apply the FE method to PDEs on surfaces, the respective differential operators as well as the integration on manifolds are discussed.
The theoretical part of this chapter is based on [Gri09, Bär10, For12] and [For13]. Moreover, [DE13] offers a more in-depth analysis of PDEs on surfaces. Numerical tests for different types of PDEs and continuous as well as discontinuous setups for PDEs on the unit sphere form the second part of this chapter.

5.2 Basics of differential geometry

Notation and setting. We consider an open subset $\Omega \subset \mathbb{R}^k$, a subset $M \subset \mathbb{R}^n$ with $k \leq n$, and a mapping $\tau : \Omega \to M$, see Figure 5.1.

The clou is that, although M is a subset of $\mathbb{R}^n$, it can be considered locally as an object in $\mathbb{R}^k$.
Functions living in Ω such as $f : \Omega \to \mathbb{R}$ are called local functions and will be denoted by lower-case letters. So-called global functions living in M, such as $F : M \to \mathbb{R}$, will be denoted by capital letters [Hin16].

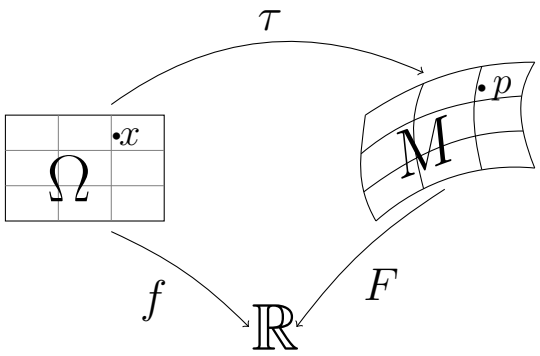


Figure 5.1: *Formal geometric setting.*

Let $x \in \Omega$, $p \in M$, and $\tau(x) = p$. The relationship between f and F is given by

$$f(x) = F(\tau(x)) \quad \text{or, in short,} \quad f = F \circ \tau.$$

Submanifold. To define a submanifold we first provide the definitions of a homeomorphism and an immersion.

Definition 5.1 (Homeomorphism)**.** *[For13, Ch. 2]. Let X and Y be topological spaces. A bijective function $f : X \to Y$ is called homeomorphism if it is continuous and the inverse function $f^{-1} : Y \to X$ is continuous as well.*

Definition 5.2 (Immersion)**.** *[For13, Ch. 9]. Let $\Omega \subset \mathbb{R}^k$ be an open subset. We call a C^1- function*

$$\tau : \ \Omega \to \mathbb{R}^n, \ (x_1, \dots, x_k) \mapsto \tau(x_1, \dots, x_k)$$

an immersion if the rank of the Jacobian matrix

$$D\tau = \begin{pmatrix} \frac{\partial \tau}{\partial x_1} \dots \frac{\partial \tau}{\partial x_k} \end{pmatrix} = \begin{pmatrix} \frac{\partial \tau_1}{\partial x_1} \dots \frac{\partial \tau_1}{\partial x_k} \\ \vdots \qquad \vdots \\ \frac{\partial \tau_n}{\partial x_1} \dots \frac{\partial \tau_n}{\partial x_k} \end{pmatrix} = \begin{pmatrix} (\nabla \tau_1)^T \\ \vdots \\ (\nabla \tau_n)^T \end{pmatrix} \in \mathbb{R}^{n \times k} \tag{5.1}$$

is maximal, that is, we have rank k, in every point $x \in \Omega$.

Definition 5.3 (Submanifold I)**.** *[For13, Ch. 9]. A subset $M \subset \mathbb{R}^n$ is called k-dimensional submanifold of $\mathbb{R}^n$ if for every point $p \in M$ there are an open neighborhood $U \subset \mathbb{R}^n$, an open subset $\Omega \subset \mathbb{R}^k$ and a mapping $\tau : \Omega \to \mathbb{R}^n$, where*

i) τ is an immersion,
ii) τ is a homeomorphism for $\tau(\Omega) = M \cap U$.

We call τ parametrization or local coordinate chart.

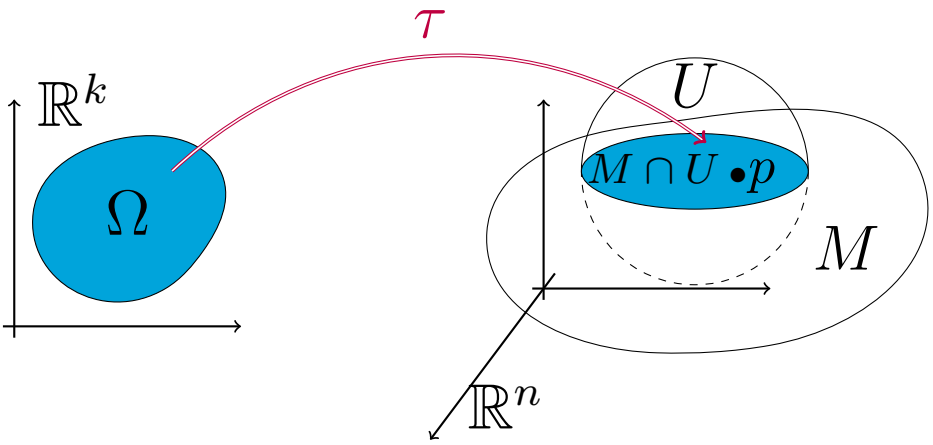


Figure 5.2: *Illustration of Definition 5.3 for a submanifold M [For13, Ch. 9].*

Remarks 5.4.

i) A curve, i.e., a continuous function $\gamma : I \subset \mathbb{R} \to \mathbb{R}^n$ can be parameterized as a whole. In contrast to this, for $k > 1$ only a certain neighborhood of a point $p \in M$ can be parameterized. This clarifies why τ is called *local* coordinate chart.

ii) Strictly speaking, unlike a submanifold, a manifold has not to be part of $\mathbb{R}^n$. However, even if M is simply called manifold in this thesis, we always assume that $M \subset \mathbb{R}^n$.

iii) To do calculus, differentiable manifolds are necessary. We usually assume C^1-manifolds [For12]. However, it is no problem to define a C^α-manifold with $\alpha > 1$. E.g, when defining the Gram matrix later, the parametrization τ has to be twice differentiable. We always assume that the manifold is sufficiently smooth.

Often it is no trivial task to find a parameterization. We present another useful option to define a submanifold.

Definition 5.5 (Submanifold II)**.** *[For13, For12]. A subset $M \subset \mathbb{R}^n$ is called k-dimensional submanifold of $\mathbb{R}^n$ if for every point $p \in M$ there are an open neighborhood $U \subset \mathbb{R}^n$ and $n - k$ continuously differentiable functions*

$$f_1, \ldots, f_{n-k} : U \to \mathbb{R},$$

so that

$$M \cap U = \{f_1(x) = \cdots = f_{n-k}(x) = 0\}$$

and for $f = (f_1, \ldots, f_{n-k})$

$$rank\big(Df(x)\big) = rank \begin{pmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_{n-k}}{\partial x_1} & \cdots & \frac{\partial f_{n-k}}{\partial x_n} \end{pmatrix} = n - k \quad \forall\, x \in M \cap U.$$

The Definitions 5.3 and 5.5 are equivalent, see [For13]. In Appendix 10.2, we present various parameterizations of the sphere $\mathbb{S}^2$ to illustrate that it is a submanifold.

Tangent plane. For a continuously differentiable scalar function $f : I \to \mathbb{R}$, $I \subset \mathbb{R}$ an interval, the idea of a tangent is well-known. The tangent in a given point $p \in I$ is the straight line that touches the function in p and has the slope $f'(p)$. With the concept of a tangent plane the idea can be generalized.

Definition 5.6 (Tangent plane)**.** *[Bär10, Def. 3.2.1]. Given a submanifold* $M \subset \mathbb{R}^n$ *and a point* $p \in M$*. The tangent plane of* M *in* p *is defined by*

$$T_pM := \{v \in \mathbb{R}^n \mid \exists \varepsilon > 0 \text{ and a continuously differentiable curve } \gamma : (-\varepsilon, \varepsilon) \to M, \\ \text{where } \gamma(0) = p \text{ and } \gamma'(0) = v\}.$$

The elements $v \in T_pM$ *are called tangent vectors.*

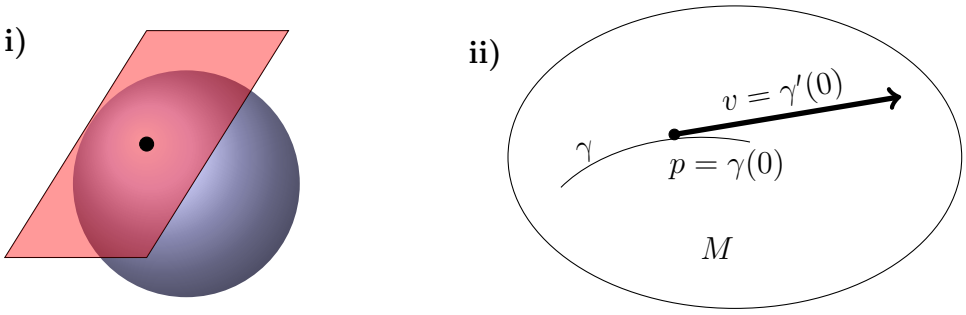


Figure 5.3: *Illustration of i) a tangent plane on the sphere and ii) a tangent vector* v*.*

To work with T_pM, it is necessary to describe it with a basis.

Lemma 5.7 (Basis of T_pM)**.** *Let* τ *be a parametrization for* M *as introduced in Definition 5.3. Then a basis of* T_pM *reads*

$$B = \left\{ \frac{\partial \tau}{\partial x_1}(x), \ldots, \frac{\partial \tau}{\partial x_k}(x) \right\},$$

Consequently, a vector $v \in T_pM$ *can be written as*

$$v = \sum_{i=1}^{k} v_i \frac{\partial \tau}{\partial x_i}(x).$$

Proof. (following [For13]). We prove that $B \subset T_pM$ by checking the characteristics of a tangent plane T_pM given in Definition 5.6. Consider a curve

$$\gamma : (-\varepsilon, \varepsilon) \to M, s \mapsto \gamma(x_1 + sv_1, \ldots, x_k + sv_k) = \tau(x + sv).$$

Obviously, $\gamma(0) = \tau(x) = p$. Furthermore, with the chain rule we obtain

$$\gamma'(s) = v_1 \frac{\partial \tau}{\partial x_1}(x + sv) + \ldots + v_k \frac{\partial \tau}{\partial x_k}(x + sv).$$

Hence, $\gamma'(0) = \sum_{i=1}^{k} v_i \frac{\partial \tau}{\partial x_1}(x) = v$, and thus $B \subset T_pM$.
The tangent plane T_pM has dimension k. Moreover, the vectors $\frac{\partial \tau}{\partial x_i}(x)$, $i = 1, \ldots, k$, correspond to the columns of $D\tau$, which are linearly independent since $D\tau$ has rank k, see Definition 5.2. Therefore, B is indeed a basis. □

Gram matrix. We upgrade our differentiable manifold to a Riemannian manifold by equipping it with a metric. A metric is a mapping that allows us to describe properties such as length, area, angle or curvature. In what follows, it is necessary to define differential operators as well as the integration over a manifold. A metric in $\mathbb{R}^n$ is typically induced by the canonical dot product

$$\langle \cdot, \cdot \rangle : \quad \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}, \quad \langle x, y \rangle = \left\langle \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} \right\rangle \mapsto \sum_{i=1}^{n} x_i y_i.$$

The metric on the Riemannian manifold should be consistent with that of the surrounding Cartesian space $\mathbb{R}^n$. This is ensured by restricting the canonical dot product to the tangent plane in order to define a metric on a manifold [Bär10]:

$$\begin{aligned} \langle \cdot, \cdot \rangle_g &:= \langle \cdot, \cdot \rangle|_{T_pM \times T_pM} \to \mathbb{R}, \\ \langle U, V \rangle_g &= \left\langle \sum_{i=1}^{k} U_i \frac{\partial \tau}{\partial x_i}(x), \sum_{j=1}^{k} V_j \frac{\partial \tau}{\partial x_j}(x) \right\rangle_g \\ &= \sum_{i,j=1}^{k} U_i V_j \underbrace{\left\langle \frac{\partial \tau}{\partial x_i}(x), \frac{\partial \tau}{\partial x_j}(x) \right\rangle}_{:= g_{ij}} = \sum_{i,j=1}^{k} U_i V_j \, g_{ij} \end{aligned} \tag{5.2}$$

Especially, for a two-dimensional submanifold in $\mathbb{R}^3$ this mapping is known as first fundamental form [Bär10]. The matrix $G := (g_{ij})_{i,j=1,\ldots k}$ is called metric tensor or

Gram matrix. It can be rewritten as

$$
\begin{aligned}
G &= (g_{ij})_{i,j=1,\dots k} \\
&= \begin{pmatrix} \frac{\partial\tau}{\partial x_1}\cdot\frac{\partial\tau}{\partial x_1} & \cdots & \frac{\partial\tau}{\partial x_1}\cdot\frac{\partial\tau}{\partial x_k} \\ \vdots & & \vdots \\ \frac{\partial\tau}{\partial x_k}\cdot\frac{\partial\tau}{\partial x_1} & \cdots & \frac{\partial\tau}{\partial x_k}\cdot\frac{\partial\tau}{\partial x_k} \end{pmatrix} = \left(\begin{array}{c} \left(\frac{\partial\tau}{\partial x_1}\right)^T \\ \hline \cdots \\ \hline \left(\frac{\partial\tau}{\partial x_k}\right)^T \end{array}\right) \left(\begin{array}{c|c|c} & & \\ \frac{\partial\tau}{\partial x_1} & \vdots & \frac{\partial\tau}{\partial x_k} \\ & & \end{array}\right) \\
&= (D\tau)^T D\tau \in \mathbb{R}^{k\times k}
\end{aligned}
$$

Unlike $D\tau$, the Gram matrix G is quadratic. Obviously, G is symmetric. Its positive definiteness follows by

$$\langle X, X\rangle_g = X^T G X = X^T D\tau^T D\tau X = (D\tau X)^T (D\tau X) := \|D\tau X\|^2 > 0 \;\forall X \neq 0.$$

As a symmetric positive definite matrix, G is invertible. We indicate that the elements from the inverse matrix are used by shifting the indices, that is,

$$G^{-1} := (g^{ij})_{i,j=1,\dots,k}\,.$$

To shorten notation, let

$$g := \det(G).$$

Being positive definite, G has a positive determinant. Thus, taking $\sqrt{g}$ does not pose a problem. Finally, it is worth mentioning that G and g are not constant. Hence, to be formally correct, we had to write $G(x)$ and $g(x)$.

5.3 Differential operators

Gradient, divergence and Laplacian are well-known as differential operators in the Cartesian space.

Definition 5.8 (Cartesian operators)**.** *[For13]. Let $U \subset \mathbb{R}^n$ be an open set.*

*a) For a C^1-function $f : U \to \mathbb{R}$ the **gradient** is the vector field*

$$\nabla f : U \to \mathbb{R}^n,\ (x_1,\dots,x_n) \mapsto \left(\frac{\partial f}{\partial x_1},\dots,\frac{\partial f}{\partial x_n}\right)^T. \tag{5.3a}$$

The gradient expands to a Jacobian matrix if the operator ∇ is applied to a vector field instead of a scalar field, see (5.1).

*b) The **divergence** of a C^1-vector field $\mathbf{v} : U \to \mathbb{R}^n$ is defined by*

$$\nabla\cdot\mathbf{v} : U \to \mathbb{R},\ (x_1,\dots,x_n) \mapsto \sum_{i=1}^{n} \frac{\partial v_i}{\partial x_i}. \tag{5.3b}$$

c) *Let $f : U \to \mathbb{R}$ be a C^2- function. Then, the* ***Laplace operator****, which combines divergence and gradient, is defined by*

$$\Delta f := \nabla \cdot \nabla f : \ U \to \mathbb{R}, \ (x_1, \ldots, x_n) \mapsto \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}. \tag{5.3c}$$

While these explicit formulas are suitable for concrete calculations, their extension to manifolds is not straightforward. Thus, we introduce an alternative representation of the gradient.

Lemma 5.9 (Relationship between the directional derivative and the gradient)**.** *[For13, Ch. 6]. Let $Df(x)v$ be the directional derivative of a function f in the direction v, where $v \in \mathbb{R}^n$ is a normed vector, i.e., $\|v\|_2 = 1$. Then, the gradient $\nabla f(x)$ is uniquely defined by*

$$Df(x)v = \langle \nabla f(x), v \rangle \quad \forall\, v \in \mathbb{R}^n. \tag{5.4}$$

Formulation (5.4) offers a geometric interpretation of the gradient. Let θ be the angle between the vectors $\nabla f(x) \neq 0$ and v. Since $\|v\| = 1$, the dot product can be rewritten as

$$\langle \nabla f(x), v \rangle = \|\nabla f(x)\| \cos(\theta).$$

The expression becomes maximal if $\theta = 0$, that is, if $\nabla f(x)$ and v are oriented in the same direction. Consequently, $\nabla f(x)$ points in the direction of the steepest ascent and $\|\nabla f(x)\|$ describes the magnitude of this ascent.

Surface gradient

A gradient with respect to a Riemannian submanifold is often called surface gradient. Its geometric interpretation is the same as in the Cartesian space: the gradient $\nabla_M F(p)$ points to the direction of the steepest ascent of manifold M in point p. Hence, $\nabla_M F$ must not have any component orthogonal to M, but $\nabla_M F(p) \in T_p M$. This is why another name for it is tangential gradient.

We present several definitions of the surface gradient, which are ultimately equivalent, but take different perspectives. Using the scalar product restricted to $T_p M$, the surface gradient can be defined in full analogy to (5.4).

Definition 5.10 (Surface gradient I)**.** *[Gri09]. Let $DF(p)v$ be the directional derivative of a function F in the direction v, where $v \in T_p M$ is a normed vector. Then, the gradient $\nabla_M F(p)$ is uniquely defined by*

$$DF(p)v = \langle \nabla_M F(p), v \rangle_g \quad \forall\, v \in T_p M. \tag{5.5}$$

Remark (Notation)**.** Numerous notations exist for differential operators on manifolds. By writing ∇_M, we express that we are on a manifold M. If the manifold is the unit sphere, $\nabla_{\mathbb{S}^2}$ is more precise. This can be shortened to $\nabla_{\mathbf{p}}$, where $\mathbf{p} \in \mathbb{S}^2$.

For practical purposes, we need an explicit formula for the gradient.

Lemma 5.11 (Surface gradient II)**.** *[Gri09]. Let $F : M \to \mathbb{R}, M \subset \mathbb{R}^n$ open, be a C^1-function. An expression for its gradient $\nabla_M F(p) : M \to T_pM$ is given by*

$$\nabla_M F = \sum_{i,j=1}^{k} \frac{\partial \tau}{\partial x_i} g^{ij} \frac{\partial f}{\partial x_j}. \tag{5.6}$$

Proof. [Gri09]. Since $\nabla_M F(p) \in T_pM$, it can be expressed as linear combination of the basic vectors of T_pM. Following Lemma 5.7, this means

$$\nabla_M F(p) = \sum_{i=1}^{k} \beta_i \frac{\partial \tau}{\partial x_i}(x).$$

The task is to find the coefficients β_i. Using the chain rule, the partial derivatives of f can be expressed as

$$\frac{\partial f}{\partial x_i}(x) = \frac{\partial (F \circ \tau)}{\partial x_i}(x) = DF\left(\frac{\partial \tau}{\partial x_i}(x)\right). \tag{5.7}$$

We proceed by employing formulation (5.5) for the surface gradient. As left hand side we obtain

$$DF(p)v \;=\; DF(p)\left(\sum_{i=1}^{k} v_i \frac{\partial \tau}{\partial x_i}(x)\right) \;=\; \sum_{i=1}^{k} v_i \; DF\left(\frac{\partial \tau}{\partial x_i}(x)\right) \overset{(5.7)}{=} \sum_{i=1}^{k} v_i \; \frac{\partial f}{\partial x_i}(x),$$

where we used in the first step that the directional vector v is an element of T_pM. Furthermore, with the first fundamental form as defined by (5.2) the right hand side of (5.5) can be written as

$$\langle \nabla_M F, v \rangle_g \;=\; \sum_{i,j=1}^{k} v_i \, g_{ij} \, \beta_j.$$

Combining the results for the right and left hand side, we obtain

$$\sum_{i=1}^{k} v_i \; \frac{\partial f}{\partial x_i}(x) \;=\; \sum_{i,j=1}^{k} v_i \, g_{ij} \, \beta_j \quad \forall \, v_i,$$

and thus,

$$\frac{\partial f}{\partial x_i}(x) \;=\; \sum_{j=1}^{k} g_{ij} \beta_j \quad \forall \, i.$$

By taking the inverse of $(g_{ij})_{ij}$ this matrix-vector multiplication can finally be reformulated to

$$\beta_i = \sum_{j=1}^{k} g^{ij} \frac{\partial f}{\partial x_j}, \tag{5.8}$$

which is the desired conclusion. □

Corollary 5.12. *Expressed with matrices the surface gradient reads*

$$\nabla_M F = (D\tau)G^{-1}\nabla f. \tag{5.9}$$

Proof. This follows immediately from the elementwise formulation (5.6). □

Formulation (5.9) is suitable to define and assemble the finite element matrices on the sphere later on. Another formulation for the surface gradient is worth mentioning if only because it provides a tool to derive the Folgar-Tucker equation from the Fokker-Planck equation, see the proof of Lemma 2.10.

Lemma 5.13 (Surface Gradient III)**.** *[DE13, Def. 2.3]. Let $\mathbf{n} \in \mathbb{R}^n$ be a unit normal vector with respect to M, that is, $\mathbf{n} \perp T_pM$ and $\|\mathbf{n}\|_2 = 1$. An expression for the surface gradient then reads*

$$\nabla_M F = \nabla F - (\nabla F \cdot \mathbf{n})\mathbf{n} = \mathcal{P}\,\nabla F, \quad \text{where } \mathcal{P}_{ij} := \delta_{ij} - n_i n_j.$$

Here, $\mathcal{P}$ is the orthogonal projection of the Euclidean gradient ∇F to the tangent plane. Function F is defined in a neighborhood of M. Its gradient $\nabla_M F$ only depends on the values of F restricted to M.

Proof. The surface gradient $\nabla_M F$ is always tangential to the manifold M. Since $\mathbf{n} \perp T_pM$, this can be expressed by the orthogonality relation

$$\nabla_M F \cdot \mathbf{n} = 0.$$

To obtain a meaningful surface gradient, we remove the normal component from ∇F and keep only its tangential part. Given that

$$\nabla_M F = \nabla F - \mu\mathbf{n},$$

the parameter $\mu \in \mathbb{R}$ has to be determined. Substituting the second into the first equation, we end up with $(\nabla F - \mu\mathbf{n}) \cdot \mathbf{n} = 0$, so that $\mu = \nabla F \cdot \mathbf{n}$. □

Corollary 5.14. *Let $\mathbf{p} \in \mathbb{R}^d$ be an orientation vector as it was already introduced at the very beginning of this work. Then the single entries of the Jacobian matrix $\nabla_M \mathbf{p}$ are given by*

$$\partial_{p,j} p_i = \delta_{ij} - p_i p_j, \qquad i, j \in \{1, \dots, d\}. \tag{5.10}$$

Proof. A single entry of the Jacobian $\nabla_M \mathbf{p} \in \mathbb{R}^{d\times d}$ is denoted by $(\nabla_M \mathbf{p})_{ij} = \partial_{p,j} p_i$. On the sphere we have $\mathbf{n} = \mathbf{p}$ and therefore the projection operator can be rewritten as $\mathcal{P}_{ij} := \delta_{ij} - p_i p_j$. □

Example 5.15 (Application of Corollary 5.14 for $d = 3$)**.** *A concrete expression for $\mathbf{p} \in \mathbb{R}^3$ is given by the spherical coordinates, which are defined in Appendix 10.1 by*

$$\mathbf{p} \;=\; (p_1,\; p_2,\; p_3)^T \;=\; (\sin\theta\cos\varphi,\; \sin\theta\sin\varphi,\; \cos\theta)^T.$$

Applying Corollary 5.14, on the one hand, we find that

$$\nabla_M \mathbf{p} = \begin{pmatrix} 1-p_1^2 & -p_1p_2 & -p_1p_3 \\ -p_1p_2 & 1-p_2^2 & -p_2p_3 \\ -p_1p_3 & -p_2p_3 & 1-p_3^2 \end{pmatrix}$$

$$= \begin{pmatrix} 1-\sin^2\theta\cos^2\varphi & -\sin^2\theta\sin\varphi\cos\varphi & -\sin\theta\cos\theta\cos\varphi \\ * & 1-\sin^2\theta\sin^2\varphi & -\sin\theta\cos\theta\sin\varphi \\ * & * & 1-\cos^2\theta \end{pmatrix}$$

On the other hand, the formula for the gradient in spherical coordinates, see (10.1), can be applied to p_1, p_2 and p_3 to obtain the rows of the symmetric Jacobian as

$$\nabla_M \mathbf{p} = \begin{pmatrix} \partial_1 p_1 & \partial_2 p_1 & \partial_3 p_1 \\ \partial_1 p_2 & \partial_2 p_2 & \partial_3 p_2 \\ \partial_1 p_3 & \partial_2 p_3 & \partial_3 p_3 \end{pmatrix}$$
$$= \begin{pmatrix} \cos^2\theta\cos^2\varphi + \sin^2\varphi & (\cos^2\theta - 1)\sin\varphi\cos\varphi & -\sin\theta\cos\theta\cos\varphi \\ * & \cos^2\theta\sin^2\varphi + \cos^2\varphi & -\sin\theta\cos\theta\sin\varphi \\ * & * & \sin^2\theta \end{pmatrix}$$

Comparing the six different entries, and using the Pythagorean trigonometric identity if necessary, shows that both approaches give the same result.

Surface divergence

Physically, the divergence defines the amount of change in a flow. A positive divergence describes a source, a negative divergence a sink. For a vanishing divergence, the corresponding velocity field is called solenoidal.
To describe the convection on a surface, the divergence has to be applied to a velocity field which is tangential to the surface. As in the Euclidean space, the surface divergence applied to a vector field V satisfies $\nabla_M \cdot V = \operatorname{tr}(\nabla_M V)$. An explicit formula for the surface divergence is given by the following definition:

Definition 5.16 (Surface divergence)**.** *Let $V = \sum_{i=1}^{k} V_i \frac{\partial \tau}{\partial x_i}(x)$ be a function living in the tangent plane T_pM. Then, its surface divergence is defined by [Ali16]*

$$\nabla_M \cdot V = \frac{1}{\sqrt{g}} \sum_{i=1}^{k} \frac{\partial}{\partial x_i}(\sqrt{g}\, V_i). \tag{5.11}$$

Remark (Different definition of the tangential field)**.** Let us consider the case of spherical coordinates. On the one hand, the field based on the basis of the tangent plane, see Lemma 5.7, reads

$$V = V_\theta \frac{\partial \tau}{\partial \theta} + V_\varphi \frac{\partial \tau}{\partial \varphi} = V_\theta\, \mathbf{e}_\theta + V_\varphi \sin\theta\, \mathbf{e}_\varphi.$$

On the other hand, the usual convention with normed basis vectors is

$$V = V_\theta\, \mathbf{e}_\theta + V_\varphi\, \mathbf{e}_\varphi.$$

We must be aware that different formulas for the surface divergence result from this, see equation (10.2) in the Appendix.

Laplace-Beltrami operator

The generalization of the Laplace operator to a Riemannian manifold is known as Laplace-Beltrami operator. Like its Cartesian analog, the Laplace-Beltrami operator takes the divergence of the gradient. This leads to the following definition:

Definition 5.17 (Laplace-Beltrami operator)**.** *[DE13, Eq. (2.2)]. For a twice differentiable function F, the Laplace-Beltrami operator is defined by*

$$\Delta_M F := \nabla_M \cdot \nabla_M F = \frac{1}{\sqrt{g}} \sum_{i,j=1}^{k} \frac{\partial}{\partial x_i} \left(\sqrt{g}\, g^{ij} \frac{\partial f}{\partial x_j} \right). \tag{5.12}$$

This definition is consistent with the previous considerations: The surface divergence can be applied to functions living in the tangent plane, in particular to $\nabla_M F$. Replacing the coefficients V_i from the general definition of the surface divergence, see (5.11), with the specific coefficients from $\nabla_M F$, see (5.8), we obtain

$$\Delta_M F = \frac{1}{\sqrt{g}} \sum_{i=1}^{k} \frac{\partial}{\partial x_i} \left(\sqrt{g} \sum_{j=1}^{k} g^{ij} \frac{\partial f}{\partial x_j} \right) = \frac{1}{\sqrt{g}} \sum_{i,j=1}^{k} \frac{\partial}{\partial x_i} \left(\sqrt{g}\, g^{ij} \frac{\partial f}{\partial x_j} \right).$$

Lemma 5.18 (Relationship between spherical and Euclidean Laplace operator)**.** *The Laplace-Beltrami operator for $M = \mathbb{S}^2$ is related to the Laplacian in $\mathbb{R}^3$ by*

$$(\Delta_{\mathbb{S}^2} f)(\mathbf{p}) = (\Delta_{\mathbb{R}^3} \overline{f})(\mathbf{p}) \quad \text{for } \mathbf{p} \in \mathbb{S}^2, \tag{5.13}$$

where function f is living on the sphere and

$$\overline{f} : \mathbb{R}^3 \backslash \{0\} \to \mathbb{R},\ \overline{f}(\mathbf{p}) = f\left(\frac{\mathbf{p}}{\|\mathbf{p}\|} \right)$$

is its extension to the Euclidean space $\mathbb{R}^3$.

Proof. [1] The Laplace operator written in spherical coordinates reads [Qua93, A.4.3]

$$\begin{aligned} \Delta_{\mathbb{R}^3} &= \frac{1}{r^2} \frac{\partial}{\partial r} \left(r^2 \frac{\partial}{\partial r} \right) + \frac{1}{r^2} \left(\frac{1}{\sin\theta} \frac{\partial}{\partial \theta} \left(\sin\theta \frac{\partial}{\partial \theta} \right) + \frac{1}{\sin^2\theta} \frac{\partial^2}{\partial \varphi^2} \right) \\ &= \frac{2}{r} \frac{\partial}{\partial r} + \frac{\partial^2}{\partial r^2} \quad + \frac{1}{r^2} \left(\frac{\cos\theta}{\sin\theta} \frac{\partial}{\partial \theta} + \frac{\partial^2}{\partial \theta^2} + \frac{1}{\sin^2\theta} \frac{\partial^2}{\partial \varphi^2} \right). \end{aligned}$$

The function $\overline{f}$ is independent of the radius r, i.e., $\overline{f}(r, \theta, \varphi) = f(\theta, \varphi)$. Consequently, $\frac{\partial}{\partial r} \overline{f} = 0$, i.e., the radial part of the Laplacian can be omitted. In the remaining spherical part, considering the unit sphere, we set $r = 1$ and obtain

$$(\Delta_{\mathbb{R}^3} \overline{f})|_{\mathbb{S}^2} = \frac{\cos\theta}{\sin\theta} \frac{\partial \overline{f}}{\partial \theta} + \frac{\partial^2 \overline{f}}{\partial^2 \theta^2} + \frac{1}{\sin^2\theta} \frac{\partial^2 \overline{f}}{\partial \varphi^2} = \Delta_{\mathbb{S}^2} f.$$

Since the last equality will be shown in Example 5.20ii), this finishes the proof. □

Lemma 5.18 states that applying the spherical Laplace-Beltrami operator to a function is equivalent to employing the Cartesian Laplace operator to that function with normalized arguments. Two examples in Appendix 10.3 illustrate this.
A statement related to that in Lemma 5.18 can be found in [Sch13, Lemma 8.17].

[1] Thanks to Dr. Günter Skoruppa, who gave the idea for this proof.

Application of differential operators. In what follows, the formulas for the surface gradient, the surface divergence and the Laplace-Beltrami operator are evaluated for $M = \mathbb{R}^n$ and for $M = \mathbb{S}^2$.

Lemma 5.19 (Differential operators for $M = \mathbb{R}^n$)**.** *In the Cartesian space, the formulas* (5.6)*,* (5.11) *and* (5.12) *for the gradient, the divergence and the Laplacian on a manifold reduce to the basic formulas* (5.3a)-(5.3c)*.*

Proof. For $M = \mathbb{R}^n$, we find that $\tau : \mathbb{R}^n \to \mathbb{R}^n$, $id : x \mapsto x$ and consequently $D\tau = G = G^{-1} = \mathbf{I}_n$. Moreover, we have $F = f$. For the example of the gradient we write it down in detail:

$$\begin{aligned}\nabla_M F &\overset{(5.6)}{=} \sum_{i,j=1}^{k} \frac{\partial \tau}{\partial x_i}\, g^{ij}\, \frac{\partial f}{\partial x_j} = \sum_{i,j=1}^{n} \frac{\partial \tau}{\partial x_i}\, \delta_{ij}\, \frac{\partial f}{\partial x_j} \\ &= \sum_{i=1}^{n} \frac{\partial \tau}{\partial x_i}\, \frac{\partial f}{\partial x_i} = \sum_{i=1}^{n} \mathbf{e}_i \frac{\partial f}{\partial x_i} = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right)^T \overset{(5.3a)}{=} \nabla f.\end{aligned}$$

For the divergence and the Laplace operator the relationship follows in a similar way. □

When assembling finite element matrices, the matrix formulation of the surface gradient given by Corollary 5.12 is necessary. Spherical coordinates are not needed. This changes when test problems are constructed.
For a given reference solution, the corresponding right hand side of the PDE has to be calculated. Furthermore, the surface divergence is necessary to check whether a given velocity field is solenoidal. Thus, determining the differential operators in spherical coordinates is not only a natural application of the previous formulas, but it is also of practical use for this work.

Example 5.20 (Differential operators in spherical coordinates)**.** *In contrast to the Cartesian space, the coefficients of the derivatives expressed in spherical coordinates are no longer constant. In the appendix, the spherical coordinates are given by*

$$\tau(\theta, \varphi) \;=\; (\sin\theta\cos\varphi,\ \sin\theta\sin\varphi,\ \cos\theta)^T.$$

Consequently,

$$D\tau = \begin{pmatrix} \cos\theta\cos\varphi & -\sin\theta\sin\varphi \\ \cos\theta\sin\varphi & \sin\theta\cos\varphi \\ -\sin\theta & 0 \end{pmatrix}, \quad G = \begin{pmatrix} 1 & 0 \\ 0 & \sin^2\theta \end{pmatrix}, \quad G^{-1} = \begin{pmatrix} 1 & 0 \\ 0 & 1/\sin^2\theta \end{pmatrix},$$

and $g = \det(G) = \sin^2\theta$*. We obtain*

i) for the tangential gradient

$$\begin{aligned}\nabla_{\mathbf{p}} &\overset{(5.9)}{=} (D\tau)\, G^{-1}\, \nabla \\ &= \begin{pmatrix} \cos\theta\cos\varphi & -\sin\theta\sin\varphi \\ \cos\theta\sin\varphi & \sin\theta\cos\varphi \\ -\sin\theta & 0 \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & 1/\sin^2\theta \end{pmatrix} \begin{pmatrix} \frac{\partial}{\partial\theta} \\ \frac{\partial}{\partial\varphi} \end{pmatrix}\end{aligned}$$

$$= \begin{pmatrix} \cos\theta\cos\varphi \\ \cos\theta\sin\varphi \\ -\sin\theta \end{pmatrix} \frac{\partial}{\partial\theta} + \begin{pmatrix} -\sin\varphi \\ \cos\varphi \\ 0 \end{pmatrix} \frac{1}{\sin\theta}\frac{\partial}{\partial\varphi}$$
$$= \mathbf{e}_\theta \frac{\partial}{\partial\theta} + \mathbf{e}_\varphi \frac{1}{\sin\theta}\frac{\partial}{\partial\varphi}, \tag{5.14}$$

ii) for the surface divergence, where $V = V_\theta \frac{\partial\tau}{\partial\theta} + V_\varphi \frac{\partial\tau}{\partial\varphi}$

$$\nabla_{\mathbf{p}} \cdot V \overset{(5.11)}{=} \frac{1}{\sin\theta}\left(\frac{\partial}{\partial\theta}(\sin\theta\, V_\theta) + \frac{\partial}{\partial\varphi}(\sin\theta V_\varphi)\right) = \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta\, V_\theta) + \frac{\partial V_\varphi}{\partial\varphi}.$$

iii) and for the Laplace-Beltrami operator

$$\begin{aligned}
\Delta_{\mathbf{p}} &\overset{(5.12)}{:=} \frac{1}{\sqrt{g}}\sum_{i,j=1}^{2}\frac{\partial}{\partial x_i}\left(\sqrt{g}\, g^{ij}\frac{\partial}{\partial x_j}\right) \\
&= \frac{1}{\sin\theta}\left(\frac{\partial}{\partial\theta}\left(\sin\theta\cdot 1\cdot\frac{\partial}{\partial\theta}\right) + 0 + 0 + \frac{\partial}{\partial\varphi}\left(\sin\theta\cdot\frac{1}{\sin^2\theta}\cdot\frac{\partial}{\partial\varphi}\right)\right) \\
&= \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}\left(\sin\theta\,\frac{\partial}{\partial\theta}\right) + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\varphi^2} \\
&= \frac{\partial^2}{\partial\theta^2} + \frac{\cos\theta}{\sin\theta}\frac{\partial}{\partial\theta} + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\varphi^2}.
\end{aligned} \tag{5.15}$$

All three expressions correspond to the formulas, which can be found in the literature, e.g., [Qua93, A.4.2+3], and which are listed in the appendix as well.

5.4 Integration on a manifold

Finally, we cover the aspect of integration for submanifolds. In particular, given a mapping $\tau : \mathbb{R}^n \to \mathbb{R}^n$, $\widehat{T} \to T$, the substitution rule is

$$\int_T f(y)\,\mathrm{d}y = \int_{\widehat{T}} f(\tau(x))\,|\det D\tau(x)|\,\mathrm{d}x. \tag{5.16}$$

Obviously, this formula cannot be applied to a mapping $\tau : \mathbb{R}^k \to \mathbb{R}^n$, $\Omega \to M$, where $k \neq n$, because $D\tau(\mathbf{x}) \in \mathbb{R}^{n\times k}$, but the determinant is only defined for quadratic matrices. Fortunately, a generalized substitution rule can be used.

Definition 5.21 (Integration on manifolds)**.** *[Bär10],[For12, Ch. 14]. Given the known setting, the integral over a manifold M is defined by*

$$\int_M F(p)\,\mathrm{d}p = \int_\Omega F(\tau(x))\sqrt{g(x)}\,\mathrm{d}x. \tag{5.17}$$

This definition is consistent with formulas we already know. We check this for the cases $k = 1$ and $k = n$.

- $k = 1$. Considering $\tau : \mathbb{R} \supset (a, b) \to \mathbb{R}^n$, we obtain

$$
\begin{aligned}
& & D\tau &= \begin{pmatrix} \frac{\partial \tau_1}{\partial x} & \dots & \frac{\partial \tau_n}{\partial x} \end{pmatrix}^T \\
\implies & & G &= (D\tau)^T D\tau = \sum_{i=1}^{n} \left(\frac{\partial \tau_i}{\partial x} \right)^2 =: \|\tau'(x)\|^2 \\
\implies & & g &= \det G = G, \quad \text{since } G \in \mathbb{R} \\
\overset{(5.17)}{\implies} & \int_M F(p)\,\mathrm{d}p &&= \int_a^b F(\tau(x))\,\|\tau'(x)\|\,\mathrm{d}x
\end{aligned}
$$

  To indicate that for $k = 1$ the manifold is a curve, we write Γ instead of M. The volume or rather the length of $\Gamma = \tau((a, b))$ is computed by

$$\int_\Gamma 1\,\mathrm{d}p = \int_a^b \|\tau'(x)\|\,\mathrm{d}x,$$

  which is a well-known formula for the length of a curve [Bär10, Def. 2.1.15].

- $k = n$. The mapping τ describes a reparametrization. For $\tau : \mathbb{R}^n \to \mathbb{R}^n$ we have $D\tau \in \mathbb{R}^{n\times n}$, that is, the matrix $D\tau$ is quadratic. This is why

$$g = \det G = \det(D\tau^T D\tau) = \det(D\tau^T)\det(D\tau) = [\det(D\tau)]^2.$$

  Applying this to (5.17), we obtain

$$\int_{\tau(\Omega)} F(p)\,\mathrm{d}p = \int_\Omega F(\tau(x))\,|\det(D\tau(x))|\,\mathrm{d}x,$$

  which is exactly the basic formula (5.16) for integration by substitution.

Having seen that Definition 5.21 leads to meaningful results, it remains to prove that the definition of the integral is unambiguous, i.e., it depends only on M and F, but not on τ.

Lemma 5.22 (Well-definedness of the integral)**.** *[For12, Ch. 14]. The definition of the integral $\int_M F$ given by (5.17) is independent of the parametrization τ.*

Proof. The proof is straightforward. We assume that there are two parametrizations $\tau : \Omega \to M$ and $\widetilde{\tau} : \widetilde{\Omega} \to M$. Let $T : \widetilde{\Omega} \to \Omega$ describe the transition map. Then we have $\widetilde{\tau} := \tau \circ T$ and

$$
\begin{aligned}
& & D\widetilde{\tau} &= D(\tau \circ T) = D\tau\, DT \\
\implies & & \widetilde{G} &= (DT^T)\,\underbrace{(D\tau)^T D\tau}_{=G}\, DT \\
\implies & \det(\widetilde{G}) &&= [\det(DT)]^2 \det(G), \quad \text{since } DT \text{ is quadratic} \\
\implies & \sqrt{\widetilde{g}} &&= |\det(DT)|\sqrt{g}.
\end{aligned}
$$

Applying the formula for integration by substitution, we end up with

$$\int_{\widetilde{\Omega}} F \circ \tau\, \sqrt{\widetilde{g}}\,\mathrm{d}\widetilde{p} = \int_{\widetilde{\Omega}} F \circ \tau \circ T\, |\det DT|\, \sqrt{g}\,\mathrm{d}\widetilde{p} = \int_\Omega F \circ \tau\, \sqrt{g}\,\mathrm{d}p,$$

which is the desired conclusion. □

Remark 5.23 (Integration for spherical coordinates)**.** [For12, Ch. 14]. For spherical coordinates with $r = 1$ integration is realized by

$$\int_0^{2\pi} \int_0^{\pi} \dots \sin\theta \, \mathrm{d}\theta \mathrm{d}\varphi.$$

5.5 Transition from theory to practice

5.5.1 Bilinear forms

In this section, we apply the finite element method to manifolds. In Section 3.3.2, a weak formulation was derived for a convection-diffusion equation in space, i.e.,

$$\frac{\partial u}{\partial t} + \nabla_{\mathbf{x}} \cdot (\mathbf{v}u) \;=\; \varepsilon \Delta_{\mathbf{x}} u \quad \text{in } \Omega \times (0, T]. \tag{3.6a}$$

This section is dedicated to a specific convection-diffusion equation, the space-independent Fokker-Planck equation on the unit sphere, i.e.,

$$\frac{\partial \psi}{\partial t} + \nabla_{\mathbf{p}} \cdot (\dot{\mathbf{p}}\psi) \;=\; D_r \Delta_{\mathbf{p}} \psi \quad \text{on } \mathbb{S}^2 \times (0, T]. \tag{2.4b}$$

The velocity field is given by $\mathbf{v} = \dot{\mathbf{p}}$, and the diffusion coefficient is set to $\varepsilon = D_r$. In particular, the Euclidean space $\Omega \subset \mathbb{R}^3$ is replaced by the unit sphere $\mathbb{S}^2 \subset \mathbb{R}^3$. The fact that $\mathbb{S}^2$ has no boundaries simplifies the problem. The changed differential operators for the diffusive and the convective term introduce new challenges [LV10].

Parametric finite elements. For the implementation of parametric finite elements we apply a transformation $\tau : \widehat{K} \to K$ from a reference element $\widehat{K}$ to a physical element K, see, e.g., [Joh16, Ran17b]. While the physical elements result from the discretization of the computational area and their appearance may vary, a reference element has a fixed size and a regular shape. In the case of rectangles and hexahedra it could read $\widehat{K} = [0, 1]^d$ or $\widehat{K} = [-1, 1]^d, d \in \{2, 3\}$.
A huge benefit of the parametric approach is that data such as basis functions and quadrature rules only have to be calculated and implemented once on the reference element. Moreover, quadrature rules are usually defined for these standardized geometries. By and large, only for 1d toy code it is realistic to work without a transformation.

Different dimensions. When the computational area is a submanifold, a transformation is needed as well. In Section 5.2, the transformation was already introduced as mapping $\tau : \Omega \to M$, even though it was called parametrization or local coordinate chart there. Usually physical and reference element have the same dimension, namely $K, \widehat{K} \subset \mathbb{R}^d$. This changes for the sphere $\mathbb{S}^2$, where a segment of the 3d space can be described by a segment of the 2d space, that is, $K \subset M \subset \mathbb{R}^3$ and $\widehat{K} \subset \Omega \subset \mathbb{R}^2$. For this reason, the known calculus cannot be transferred one-to-one to this task. We have reached the point, where all the preparatory work of this chapter comes together. The following two bilinear forms are most important for practical implementation.

Theorem 5.24 (Bilinear forms)**.** *The bilinear forms in the planar domain Ω that arise from a PDE formulated for a manifold in the three-dimensional space are*

i) $a_{diff}(\varphi,\psi) = \int_\Omega (\nabla\varphi)^T\, G^{-1}\, \nabla\psi\, \sqrt{g}\ \mathrm{d}\mathbf{x}$ *for the diffusive term* $-\Delta_{\mathbf{p}} u$ *and*

ii) $a_{conv}(\varphi,\psi) = -\int_\Omega \mathbf{v}\cdot((D\tau)G^{-1}\nabla\psi)\,\varphi\,\sqrt{g}\ \mathrm{d}\mathbf{x}$ *for the convective term* $\nabla_{\mathbf{p}}\cdot(\mathbf{v}u)$,

where $\tau : \mathbb{R}^2 \to \mathbb{R}^3, \Omega \to M$, is the mapping from the planar domain to the manifold, and G is the Gram matrix with determinant g as introduced in Section 5.3.

Proof. First, we write down the basic procedure in formulas to explain it afterwards. For the diffusive operator $-\Delta_{\mathbf{p}} u$ we obtain

$$\begin{aligned}
\int_M -\Delta_{\mathbf{p}}\Phi\,\Psi\,\mathrm{d}\mathbf{p} &\overset{\text{(i)}}{=} \int_M \nabla_{\mathbf{p}}\Phi\cdot\nabla_{\mathbf{p}}\Psi\,\mathrm{d}\mathbf{p} \\
&\overset{\text{(ii)}}{=} \int_\Omega ((D\tau)G^{-1}\nabla\varphi)^T((D\tau)G^{-1}\nabla\psi)\,\sqrt{g}\ \mathrm{d}\mathbf{x} \\
&= \int_\Omega (\nabla\varphi)^T G^{-T}\,\underbrace{(D\tau)^T D\tau}_{=G}\,G^{-1}\nabla\psi\,\sqrt{g}\ \mathrm{d}\mathbf{x} \\
&= \int_\Omega (\nabla\varphi)^T\, G^{-1}\,\nabla\psi\,\sqrt{g}\ \mathrm{d}\mathbf{x},
\end{aligned}$$

and analogously for the convective operator $\nabla_{\mathbf{p}}\cdot(\mathbf{v}u)$ it holds that

$$\begin{aligned}
\int_M \nabla_{\mathbf{p}}\cdot(\mathbf{v}\Phi)\,\Psi\,\mathrm{d}\mathbf{p} &\overset{\text{(i)}}{=} -\int_M (\mathbf{v}\Phi)\ \cdot\ \nabla_{\mathbf{p}}\Psi\,\mathrm{d}\mathbf{p} \\
&= -\int_M (\mathbf{v}\cdot\ \nabla_{\mathbf{p}}\Psi)\,\Phi\,\mathrm{d}\mathbf{p} \\
&\overset{\text{(ii)}}{=} -\int_\Omega \mathbf{v}\cdot((D\tau)G^{-1}\nabla\psi)\,\varphi\,\sqrt{g}\ \mathrm{d}\mathbf{x}.
\end{aligned}$$

Here Φ and Ψ are the trial and test functions on the manifold M, whereas φ and ψ are trial and test functions in the domain Ω. We first derive the weak formulation on the manifold. As usual we obtain it by multiplying the terms of the PDE by a test function, and then integrating this expression over the whole domain M. Afterwards integration by parts might be performed, see the equations labeled with (i). The capital letters for trial and test functions point out that until now the functions are living on the manifold.

Now the approach has to be reformulated as an integral over a 2d domain. At the descriptive level a piece of the sphere can locally be considered as an object in the 2d space. From the technical perspective we apply the transformations labeled with (ii). First, integration on manifolds as presented in Definition 5.21 is used; that is, the standard substitution rule is replaced by the generalized one which yields the term $\sqrt{g}$. Moreover, we need formula (5.9) to substitute the surface gradient. The diffusive term can be further simplified by rewriting the integrand and applying the definition of the Gram matrix. □

Remarks. i) (Avoidance of coordinate charts.) It is worth mentioning that the Jacobian $D\tau$ arises in the formulas, but the parametrization $\tau : \Omega \to M$ is not explicitly required for the assembly of the finite element matrices. The geometry is only needed through the knowledge of the vertices of the triangulation [DE13].

ii) (Number of transformations.) The question might arise if there are not actually two transformations, first the transformation between the 3d and the 2d space, and secondly the common transformation between the reference and the physical element. The answer is that within the finite element software these two transformations are merged into one.

iii) (Generalized integration by parts.) In Theorem 5.24 integration by parts, and thus the weak formulation, apparently transfer one-to-one from Euclidean space. This is due to the special situation on the sphere, where, among other things, no boundary terms arise. In general, the corresponding formulas for manifolds are to be adapted [DE13, Th. 2.10+11+18].

Linear vs. quadratic transformations. Another aspect is the general choice of the transformation. Both the linear and the quadratic transformation are visualized in Figure 5.4. A quadratic transformation in the Euclidean space calculates the midpoints of the edges/faces from the coordinates of the vertices. For the unit sphere this has to be supplemented by a normalization of the midpoints to ensure that these new points are located on the sphere as well.

In the context of our limiting strategy we are restricted to (multi-)linear elements, so that usually a linear standard transformation is enough. Furthermore, a quadratic transformation requires more effort than a quadratic one.
However, in Section 5.6 there are also the numerical examples, which use quadratic elements, so that a quadratic transformation is necessary to obtain the optimal order of convergence. Moreover, even in the case of (multi-)linear elements, a quadratic transformation increases the accuracy.

When transforming between geometric objects of the same dimension, with a suitably chosen transformation, any point is mapped exactly from one domain to the other. This is no longer true when a curvilinear area, in our case the surface of a sphere, is involved. Consequently, the transformation introduces an error and this error can even lead to the violation of the maximum principles. The usage of a quadratic transformation reduces this error as much as possible.

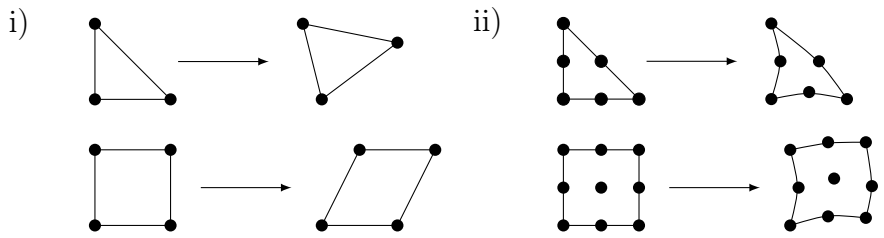


Figure 5.4: *i) Linear and ii) quadratic transformation from the reference to a physical element on the unit sphere, visualized for a triangle as well as for a rectangle.*

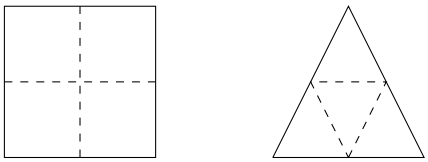
5.5.2 Meshing

We need a discretization of the sphere to employ the finite element method. For example, when the sphere models the globe to simulate the weather on earth, the geographic coordinate system with latitude and longitude comes to mind. However, the quadrilaterals, which arise in the pole regions, are contrary to every desired property of a mesh described in Section 3.3.1. An improved mesh is the cubed sphere grid [PBR13]. It is based on six cube faces. A so-called gnomic/central transformation is applied and different positions are described by trigonometric relations [NTL05]. In practice, however, this did not prove to be the best option. Instead, we take 3d geometries, consisting of either quadrilateral or triangular areas, and define them as initial meshes on level 0. A cube or an icosahedron are suitable. The vertices are chosen as the points on the unit sphere that has its center in the origin. The eight vertices of the cubes then are described by

$$(\pm a, \pm a, \pm a)\,, \quad \text{where } a = \frac{1}{\sqrt{3}},$$

while the twelve vertices of the icosahedron are given by

$$(0, \pm b, \pm c)\,,\ (\pm b, \pm c, 0)\,,\ (\pm c, 0, \pm b)\,, \quad \text{where } b = \sqrt{\frac{2}{5+\sqrt{5}}},\ c = \frac{1+\sqrt{5}}{\sqrt{10+2\sqrt{5}}}.$$



To obtain the coordinates of suitable grid points, in every refinement step a square is divided into four small squares and a triangle is divided into four small triangles, see Figure 5.5

Figure 5.5: *Refinement rules.*

Then every new node is normalized. This results in a regular mesh depicted in Figure 5.6.

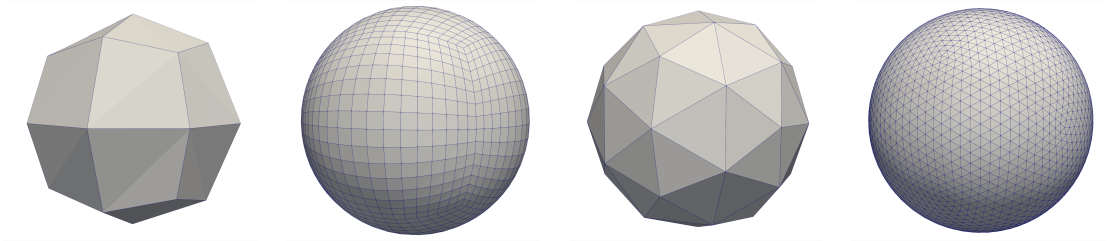


Figure 5.6: *Spherical meshes on level 1 and 4, for both quadrilateral elements (left) and triangular elements (right).*

In order to estimate the cost of simulation, on the one hand, and accuracy of the simulation, on the other hand, it is relevant of how many nodes the spherical grid consists. Table 5.7 summarizes the specific number of mesh nodes. This number can be described with some small recursive formulas, see Lemma 5.25.

level	number of nodes quadrilateral mesh	triangular mesh
0	8	12
1	26	42
2	98	162
3	386	642
4	1 538	2 562
5	6 146	10 242
6	24 578	40 962
7	98 306	163 842
8	393 218	655 362
9	1 572 866	2 621 442

Table 5.7: *Number of spherical mesh points starting from a cube or an icosahedron as geometry.*

Lemma 5.25 (Number of nodes)**.** *Let v_l describe the number of vertices on level l.*

a) For the quadrilateral mesh we have

$$v_0 := 8 \quad \textit{and} \quad v_l = v_{l-1} + 18 \cdot 4^{l-1} \ \textit{for}\ l \geq 1.$$

b) For the triangular mesh we have

$$v_0 := 12 \quad \textit{and} \quad v_l = v_{l-1} + 30 \cdot 4^{l-1} \ \textit{for}\ l \geq 1.$$

Proof. The above formulas follow from geometric considerations. We start with the quadrilateral mesh. At level 0, i.e., for the cube, the number of vertices, edges and faces is given by $v_0 = 8, e_0 = 12, f_0 = 6$. For the next higher level we obtain

$$\begin{aligned} v_l &= v_{l-1} + e_{l-1} + f_{l-1}, \\ e_l &= 2e_{l-1} + 4f_{l-1}, \\ f_l &= 4f_{l-1}. \end{aligned}$$

It holds true that $e_0 = 2f_0$ and combining the second and the third equation it can be shown inductively that $e_l = 2f_l$ for $l > 0$ as well. Moreover, the third equation can be reformulated as $f_l = 4^l f_0 = 6 \cdot 4^l$. Inserting these two formulas in the first equation leads to the desired conclusion.
The procedure for the icosahedron is similar. For the original geometry $v_0 = 12$, $e_0 = 30$, $f_0 = 20$, whereas for $l \geq 1$ the relationship reads

$$\begin{aligned} v_l &= v_{l-1} + e_{l-1}, \\ e_l &= 2e_{l-1} + 3f_{l-1}, \\ f_l &= 4f_{l-1}. \end{aligned}$$

Combining the two equations $2e_l = 3f_l$ and $f_l = 4^l f_0 = 20 \cdot 4^l$ with the first equation establishes our formula. □

5.6 Numerical studies on sphere

In this section we consider different subproblems of the space-independent FPE and solve them numerically. We start with elliptic PDEs on the sphere, extend this to the spherical heat equation, and finally examine different settings for hyperbolic equations. By combining these components, an implementation for the space-dependent FPE can be assembled as it is done in Chapter 7.

5.6.1 Elliptic equations

For the elliptic PDEs on the sphere we verify our numerical results using an analytical reference solution. Stating a relatively arbitrary solution u, its spherical Laplacian $\Delta_{\mathbf{p}} u$ is calculated to obtain the right hand side f, which is then given in the code. To obtain the H^1-error, the gradient $\nabla_{\mathbf{p}} u$ has to be known as well. The concrete calculations for the following two examples can be found in Appendix 10.3.

5.6.1.1 Spherical Poisson equation

The Poisson equation as the most basic elliptic PDE reads

$$-\Delta_{\mathbf{p}} u = f \quad \text{on } \mathbb{S}^2.$$

As analytical solution we give $u = 2z$, see Figure 5.8. The corresponding right hand side is $f = -4z$.

Figure 5.8: *u=2z*

Without an additional requirement the solution of the Poisson equation is not unique since any constant can be added to it. One remedy is to prescribe the integral value $\int_{\mathbb{S}^2} u = c$, where the constant c is adjusted to the chosen function u, here $c = 0$. A wrong constant can crystallize in a wrong EOC for the L^2-norm but an expected

EOC for the H^1-norm. Below we compare the L^2- and the H^1-errors, and the associated EOCs for both a quadrilateral and a triangular mesh.

i)	$\mathbf{Q}_1$ (with linear transformation)				$\mathbf{Q}_2$ (with quadratic transformation)			
level	L^2-err	EOC	H^1-err	EOC	L^2-err	EOC	H^1-err	EOC
3	5.06e-2	-	3.25e-1	-	2.84e-05	-	1.35e-4	-
4	1.29e-2	1.9694	1.64e-1	0.9827	1.79e-06	3.9865	8.62e-6	3.9668
5	3.25e-3	1.9923	8.24e-2	0.9957	1.12e-07	3.9966	5.42e-7	3.9909
6	8.12e-4	1.9981	4.12e-2	0.9989	7.02e-09	3.9985	3.40e-8	3.9941
7	2.08e-4	1.9995	2.06e-2	0.9997	4.55e-10	3.9477	2.64e-9	3.6897

ii)	$\mathbf{P}_1$ (with linear transformation)				$\mathbf{P}_2$ (with quadratic transformation)			
level	L^2-err	EOC	H^1-err	EOC	L^2-err	EOC	H^1-err	EOC
3	2.34e-2	-	1.79e-1	-	2.66e-5	-	4.21e-4	-
4	5.90e-3	1.9881	8.99e-2	0.9920	1.67e-6	3.9922	5.26e-5	3.0009
5	1.48e-3	1.9970	4.50e-2	0.9979	1.05e-7	3.9957	6.57e-6	3.0003
6	3.70e-4	1.9992	2.25e-2	0.9995	6.83e-9	3.9412	8.22e-7	3.0001
7	9.25e-5	1.9998	1.13e-2	0.9999	1.77e-9	1.9488	1.03e-7	2.9988

Table 5.9: *Convergence results for the spherical Poisson equation with $u = 2z$.*

Since our domain is uniformly discretized and solution u is sufficiently smooth, for shape functions from P_r the following basic estimates hold for the Poisson equation:

$$\|u - u_h\|_{L^2(\Omega)} \leq c_1 h^{r+1} \qquad \text{and} \qquad \|u - u_h\|_{H^1(\Omega)} \leq c_2 h^r.$$

Thus, for (multi-)linear shape functions ($r = 1$), the error in the L^2-norm ideally behaves like $\mathcal{O}(h^2)$, i.e., it reduces to a quarter if h is halved, whereas in the H^1-norm a behavior like $\mathcal{O}(h)$ can be expected, i.e., the error is halved if h is halved. In fact, for $\mathbf{P}_1/\mathbf{Q}_1$ we observe the EOC 2 for the L^2-error and the EOC 1 for the H^1-error in Table 5.9.
For quadratic elements we expect the EOC 3 in the L^2-norm and the EOC 2 in the H^1-norm. These expectations are exceeded in all cases, although there are differences between the meshes. This superconvergence effect seems to be related to the analytical solution. Replacing $u = 2z$ with $u = xy$, the EOC reduces to the expected value. Finally, note that for $\mathbf{P}_2/\mathbf{Q}_2$ the EOC stagnates or is even getting worse on level 7, because the solver can no longer improve for errors in the order of e-9 or e-10.

Transformation. Typically, for $\mathbf{P}_1/\mathbf{Q}_1$ a linear transformation between reference and physical element is sufficient, whereas we use a quadratic transformation for $\mathbf{P}_2/\mathbf{Q}_2$. If quadratic elements are combined only with a linear transformation, we expect the order of convergence to be reduced to that of (multi-)linear elements. However, in some cases the higher order remains, especially in the H^1-norm. Conversely, joining (multi-)linear elements with a quadratic transformation, for the Poisson equation no improvement was detectable, whereas in the hyperbolic case, described in Section 5.6.3, a significantly increased order, not that far away from the

order expected for quadratic elements, was observed. Combining $\mathbf{P}_1/\mathbf{Q}_1$ with a quadratic instead of a linear transformation has another advantage. While for the standard transformation quantities like $\mathrm{tr}(\mathbf{A})$ or the mass of the solution u seem to be only approximately one, they become exact with the quadratic transformation.

In what follows, we restrict ourselves to the quadrilateral mesh, because usually switching to a triangular mesh causes no relevant changes.

5.6.1.2 Spherical diffusion-reaction equation

Another possibility to ensure uniqueness of the elliptic PDE is to extend the Poisson equation to a diffusion-reaction equation, that is,

$$-\Delta_{\mathbf{p}} u + u = f \quad \text{on } \mathbb{S}^2.$$

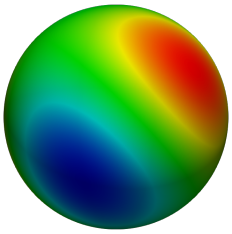


Figure 5.10: *u=xy*

We choose $u = xy$ as reference solution, see Figure 5.10. Again, relevant calculations can be found in Appendix 10.3. Especially, the right hand side reads $f = 7xy$. Once more we obtain the optimal order of convergence.

	$\mathbf{Q}_1$ (with linear transformation)				$\mathbf{Q}_2$ (with quadratic transformation)			
level	L^2-err	EOC	H^1-err	EOC	L^2-err	EOC	H^1-err	EOC
3	5.91e-3	-	2.14e-1	-	3.16e-4	-	1.13e-2	-
4	1.50e-3	1.9783	2.14e-1	1.0015	4.03e-5	2.9735	2.84e-3	1.9882
5	3.77e-4	1.9944	5.33e-2	1.0004	5.06e-6	2.9934	7.12e-4	1.9970
6	9.43e-5	1.9986	2.67e-2	1.0001	6.33e-7	2.9984	1.78e-4	1.9992
7	2.36e-5	1.9996	1.33e-2	1.0000	7.91e-8	2.9996	4.45e-5	1.9998

Table 5.11: *Results for the diffusion-reaction equation on the sphere with $u = xy$.*

Solver. For the linear systems resulting from the previous elliptic PDEs or the following parabolic PDE, the SSOR ('symmetric successive overrelaxation') method was used as preconditioner and the PCG ('preconditioned conjugate gradient') method as main solver. The PCG method requires a symmetric positive definite system matrix [Mei15]. Consequently, due to Lemma 3.5 it is suitable for both the stiffness matrix from the Poisson equation, and for the mass matrix resulting the reactive term or the temporal term.

5.6.2 Parabolic equations

Adding a time derivative to the Poisson equation we obtain the heat equation

$$\frac{\partial u}{\partial t} - \Delta_{\mathbf{p}} u = f \quad \text{on } \mathbb{S}^2 \times (0, T].$$

We choose $T = 1$. In contrast to the Poisson equation, the solution of this time-dependent problem is already uniquely determined by the initial condition $u(0) = u_0$. Choosing a suitable time stepping scheme and an appropriate time step size Δt, we can expect the same optimal order as for the Poisson equation.

We use Crank-Nicolson as time stepping scheme for both (multi-)linear and quadratic elements. As second-order method Crank-Nicolson is certainly suitable for $\mathbf{Q}_1$. In combination with $\mathbf{Q}_2$ it could be replaced by a third-order time stepping scheme. Alternatively, the refinement factor for the time has to be $\sqrt{8} \approx 2.828$:

For cubic convergence in space, halving mesh size yields an error smaller by a factor of $2^3 = 8$. For quadratic convergence in space, mesh size divided by A leads to an error divided by A^2. We need that $A^2 = 8$, that is, $A = \sqrt{8}$. In practice, Crank-Nicolson can be used as a third-order time stepping scheme by not just doubling, but tripling the number of time steps when switching to a finer spatial level.

Inhomogeneous heat equation. First, we consider an inhomogeneous heat equation. As an exact solution we use an extension of the function, which we have already used for the diffusion-reaction equation, namely

$$u(x, y, t) = xy \exp(-t).$$

Accordingly, the initial condition is $u_0(x, y) = xy$, see also Figure 5.10. The right hand side reads $f(x, y) = -u + 6u = 5xy \exp(-t)$. Starting from level 3, we use 10, 20, 40, 80 and 160 steps for $\mathbf{Q}_1$ and 20, 60, 180, 540 and 1620 steps for $\mathbf{Q}_2$.

	$\mathbf{Q}_1$ (with linear transformation)				$\mathbf{Q}_2$ (with quadratic transformation)			
level	L^2-err	EOC	H^1-err	EOC	L^2-err	EOC	H^1-err	EOC
3	6.71e-3	-	7.66e-2	-	1.19e-4	-	4.15e-3	-
4	1.72e-3	1.9599	3.90e-2	0.9744	1.50e-5	2.9916	1.05e-3	1.9901
5	4.34e-4	1.9898	1.96e-2	0.9935	1.87e-6	3.0002	2.62e-4	1.9974
6	1.09e-4	1.9974	9.81e-3	0.9935	2.34e-7	3.0017	6.55e-5	1.9993
7	2.72e-5	1.9994	4.91e-3	0.9996	2.92e-8	3.0015	1.64e-5	1.9998

Table 5.12: *Convergence results for the spherical heat equation with* $u = xy \exp(-t)$.

The convergence results documented in Table 5.12 are optimal. The calculations for $\mathbf{Q}_1$ are finished within several seconds. Making use of $\mathbf{Q}_2$, however, the computational time increases to more than half an hour. This is not only due to the changed space, but in particular due to the large number of steps.

Homogeneous heat equation. Secondly, we consider the homogeneous heat equation, where $f = 0$. A general solution of the spherical Laplace equation is given by a linear combination of the so-called spherical harmonics, see, e.g., [Nol13]. Considering only parts of this linear combination and extending them by a suitable temporal term, an exact solution of the homogeneous heat equation reads [TP01]

$$u(\theta, t) = \frac{1}{4\pi} \sum_{n=0}^{\infty} (2n + 1) P_n(\cos\theta) \exp(-n(n + 1)t). \tag{5.18}$$

Here, the P_n are the Legendre polynomials and $\theta \in [0, \pi]$ is the polar angle, see Section 10.1.

In principle, $u(\theta, 0)$ describes the δ-distribution at the north pole of the sphere. However, an unrealistically large number of terms is necessary to approximate the δ-distribution for $t = 0$. In practice, formula (5.18) is not suitable to describe the transition from the Dirac distribution to a Gaussian distribution [TP01]. For slightly larger t, the infinite series can be well approximated by only very few terms because the temporal term is of the form $\exp(-n(n+1)t)$.

Each individual summand of the series in (5.18) fulfills the homogeneous heat equation. For our simulation, we use the partial sum consisting of the first four summands of the infinite series. Substituting $\cos\theta := \xi$ we obtain

$$\begin{aligned} u(\theta, t) &= \frac{1}{4\pi} \sum_{n=0}^{3} (2n+1) P_n(\cos\theta) \exp(-n(n+1)t) \\ &= \frac{1}{4\pi} \Big(P_0(\xi) + 3P_1(\xi) e^{-2t} + 5P_2(\xi) e^{-6t} + 7P_3(\xi) e^{-12t} \Big) \\ &= \frac{1}{4\pi} \Big(1 + 3\xi e^{-2t} + \frac{5}{2}(3\xi^2 - 1) e^{-6t} + \frac{7}{2}(5\xi^3 - 3\xi) e^{-12t} \Big). \end{aligned} \tag{5.19}$$

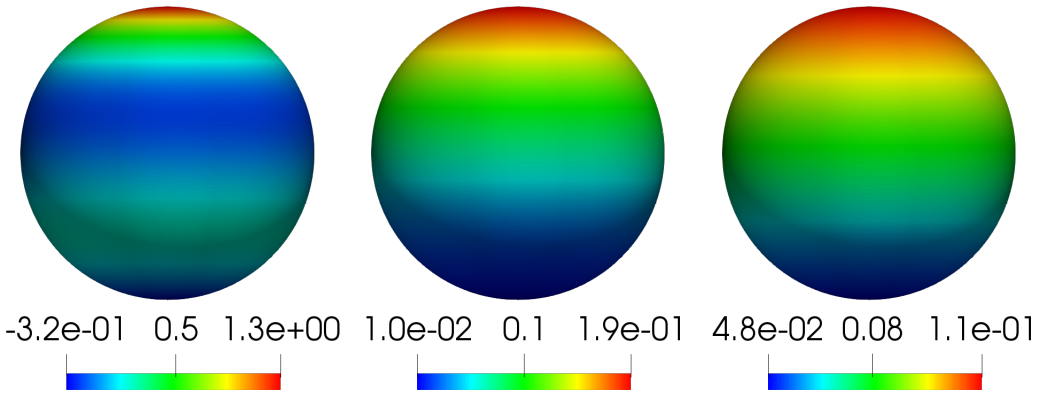


Figure 5.13: *Solution (5.19) of the homogeneous heat equation at the times 0, 0.5 and 1.*

Using 25, 50, 100, 200 and 400 time steps on the interval $[0, 1]$ for $\mathbf{Q}_1$ and 100, 300, 900, 2700 and 8100 steps for $\mathbf{Q}_2$, we obtain the optimal order of convergence, see Table 5.14.

	$\mathbf{Q}_1$ (with linear transformation)				$\mathbf{Q}_2$ (with quadratic transformation)			
level	L^2-err	EOC	H^1-err	EOC	L^2-err	EOC	H^1-err	EOC
3	3.09e-3	-	7.04e-3	-	3.15e-06	-	1.60e-5	-
4	8.01e-4	1.9493	2.94e-3	1.2594	4.15e-07	2.9249	3.53e-6	2.1761
5	2.02e-4	1.9873	1.37e-3	1.0963	5.35e-08	2.9566	8.59e-7	2.0381
6	5.06e-5	1.9968	6.74e-4	1.0282	6.49e-09	3.0427	2.14e-7	2.0084
7	1.27e-5	1.9992	3.35e-4	1.0076	7.59e-10	3.0968	5.33e-8	2.0018

Table 5.14: *Convergence results for the homogeneous spherical heat equation at time $T = 1$.*

5.6.3 Hyperbolic equations

Last but not least, we solve transport equations on the sphere, that is,

$$\frac{\partial u}{\partial t} + \nabla_{\mathbf{p}} \cdot (\mathbf{v}u) = 0 \quad \text{on } \mathbb{S}^2 \times (0, T].$$

A suitable bilinear form for the convective term was determined in Section 5.5. For the different test cases, the velocity must be tangential to the sphere. This reflects that the surface divergence has to be applied to a function, which lives in the tangent plane of the surface, see Definition 5.16.

5.6.3.1 Extended Jeffery equation

First, we consider the extended Jeffery equation (2.11). The corresponding velocity field $\mathbf{v} = \dot{\mathbf{p}} = \dot{\mathbf{p}}(\mathbf{D}(\nabla\mathbf{u}), \mathbf{W}(\nabla\mathbf{u}))$ is an element of the tangent plane $T_p\mathbb{S}^2$, see Lemma 2.2. As initial configuration we assume the isotropic distribution, so that the exact solution is given by Corollary 2.5.

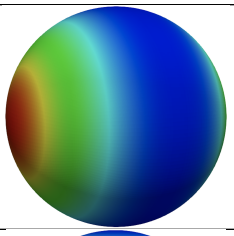
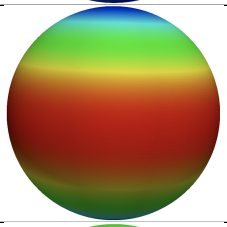
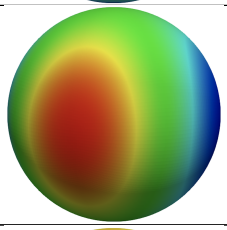
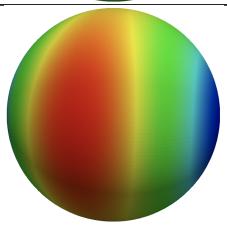
$\nabla\mathbf{u}$ and flow field name	$\mathbf{A}$ and $[u_{\min}, u_{\max}]$	visualization of u
$\begin{pmatrix} 0.02 & & \\ & -0.01 & \\ & & -0.01 \end{pmatrix}$ **uniaxial elongation**	$\begin{pmatrix} 0.349 & & \\ & 0.325 & \\ & & 0.325 \end{pmatrix}$ $u \in [7.5, 9.0] \times 10^{-2}$	
$\begin{pmatrix} 0.01 & & \\ & 0.01 & \\ & & -0.02 \end{pmatrix}$ **biaxial elongation**	$\begin{pmatrix} 0.341 & & \\ & 0.341 & \\ & & 0.318 \end{pmatrix}$ $u \in [7.1, 8.4] \times 10^{-2}$	
$\begin{pmatrix} 0 & 0.05 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$ **simple shear**	$\begin{pmatrix} 0.334 & 0.020 & \\ 0.020 & 0.332 & \\ & & 0.333 \end{pmatrix}$ $u \in [6.9, 9.2] \times 10^{-2}$	
$\begin{pmatrix} -0.005 & 0.05 & \\ & -0.005 & \\ & & 0.01 \end{pmatrix}$ **shear stretch**	$\begin{pmatrix} 0.330 & 0.019 & \\ 0.019 & 0.329 & \\ & & 0.341 \end{pmatrix}$ $u \in [6.7, 9.0] \times 10^{-2}$	

Table 5.15: *Solution of extended Jeffery equations at time $T = 2$ for different $\nabla\mathbf{u}$. The solution u is described by its interval, the orientation tensor $\mathbf{A}$ and a visualization of u.*

In Table 5.15, four different flow fields defined by a Jacobian $\nabla \mathbf{u}$ are compared. The results of the corresponding Jeffery equations at $T = 2$ are described by the quantities $\mathbf{A}$ and $[u_{\min}, u_{\max}]$. Moreover, the solutions are visualized. We look at the spheres in Table 5.15 from the midpoint between the positive x-, y- and z-direction, see Figure 5.16.

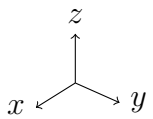


Figure 5.16: *Perspective.*

Since the exact solution is known for each point in time, in principle the final time T can be chosen arbitrarily. However, all four flow fields form peaks over time. The transition from a problem with smooth data to a problem with discontinuous data would at least require the usage of a limiting algorithm.

We choose $T = 2$. The convergence results for the four fields are very similar. We consider the shear stretch case as it seems to be the most interesting one. The results are summarized in Table 5.17.

	$\mathbf{Q}_1$		$\mathbf{Q}_2$	
level	L^2-error	EOC	L^2-error	EOC
3	5.64e-3	-	1.13e-5	-
4	1.42e-3	1.9942	2.14e-6	2.3980
5	3.54e-4	1.9985	4.87e-7	2.1313
6	8.86e-5	1.9995	1.19e-7	2.0362
7	2.22e-5	1.9997	2.94e-8	2.0137

Table 5.17: *Convergence results for the extended Jeffery equation with shear stretch velocity field.*

In combination with $\mathbf{Q}_1$, we applied both Crank-Nicolson and BDF2 as time stepping schemes, while the spatial discretization $\mathbf{Q}_2$ was combined with both Crank-Nicolson (tripling the number of time steps from level to level) and with BDF3. In each case, the errors were identical.

For the advection equation as order of convergence with respect to the L^2-norm only $\mathcal{O}(h)$ can be proven in the general case and only $\mathcal{O}(h^{\frac{3}{2}})$ if a suitable stabilization is used, see, e.g., [LT05, Don18]. Using $\mathbf{Q}_1$, however, we even obtain the order 2, see Table 5.17. This fits, for instance, with observations in [Loh19, Tab. 3.1], where under optimal conditions (smooth data, uniform mesh), also order 2 could be observed when solving the advection equation with linear elements.
An order reduced by one compared to the optimal order for elliptic equations is observed, when we use quadratic elements. For the remaining test cases we therefore restrict ourselves to $\mathbf{Q}_1$.

Solver. The previous combination of SSOR and PCG fails, when we switch to convective equations because the system matrix is no longer symmetric positive definite. Instead, we use the SPAI preconditioner and BiCGStab as the solver, as we did in Section 4.4.2.

5.6.3.2 Self-reproducing configurations

Next, we use velocity fields, whose initial configuration is reproduced after orbiting around the sphere for a time interval of size 1. Consequently, the exact solution is known at that point of time. We consider both a temporally constant and a deformational flow field, and combine both of these fields with two different initial configurations, so that four different test cases are examined.

Initial configurations. In contrast to the previous setup we do not start with an isotropic distribution anymore, but the initial configurations describe a continuous and a discontinuous geometry. so-called Gaussian Hill defined by

$$u_0(x,y,z) = \exp\big(-5\big((x-x_0)^2 + (y-y_0)^2 + (z-z_0)^2\big)\big)$$

as continuous configuration, see, e.g., [LSPT12]. Even though an exponential function is often unfavorable because no zero boundary conditions can be prescribed, this is not a disadvantage on the sphere without boundary.

The triple (x_0, y_0, z_0) defines position and magnitude of the center of the Gaussian Hill. For x_0, y_0 or $z_0 \gg 1$ the exponential function u_0 converges to zero. We choose $(x_0, y_0, z_0) = (1, 0, 0)$, that is,

$$u_0(x,y,z) = \exp\big(-5\big((x-1)^2 + y^2 + z^2\big)\big).$$

Then, $\max u_0 = 1$ and the hill is oriented in positive x-direction, see Figure 5.18.

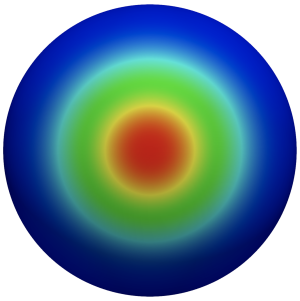


Figure 5.18: *The Gaussian Hill configuration on a sphere.*

Slotted Cylinders. As discontinuous configuration we choose two slotted cylinders. This geometry is already known from the numerical studies in Section 4.4. However, the manifold complicates the definition. It now reads [LSPT12]

$$u_0(\theta,\varphi) = \begin{cases} 1 & \text{if } r_i \le r \text{ and } |\varphi-\varphi_i| \ge \frac{r}{6} \quad \text{for } i=1,2, \\ 1 & \text{if } r_1 \le r \text{ and } |\varphi-\varphi_1| < \frac{r}{6} \quad \text{and} \quad \theta-\theta_1 < -\frac{5}{12}r, \\ 1 & \text{if } r_2 \le r \text{ and } |\varphi-\varphi_2| < \frac{r}{6} \quad \text{and} \quad \theta-\theta_2 > \frac{5}{12}r, \\ 0 & \text{otherwise,} \end{cases}$$

Figure 5.19: *Two slotted cylinders on a sphere.*

where

$$r = 0.5,\ \theta_1 = \theta_2 = 0,\ \varphi_1 = \frac{5\pi}{6},\ \varphi_2 = \frac{7\pi}{6};$$
$$r_1 = \arccos\big(\cos(\theta_1)\cos(\theta) + \sin(\theta_1)\sin(\theta)\cos(\varphi-\varphi_1)\big)$$
$$r_2 = \arccos\big(\cos(\theta_2)\cos(\theta) + \sin(\theta_2)\sin(\theta)\cos(\varphi-\varphi_2)\big).$$

Radius r_1 refers to the left circle/cylinder, while radius r_2 refers to the right one. The first condition for u_0 defines the four circular segments. The second and third condition are responsible for the area behind the slots, see Figure 5.19

Flow fields. Both velocity fields are of the form

$$\mathbf{v} = v_1 \mathbf{e}_\varphi - v_2 \mathbf{e}_\theta.$$

The basis vectors $\mathbf{e}_\varphi$ and $\mathbf{e}_\theta$ are given in Appendix 10.1. This representation ensures that $\mathbf{v}$ is tangential to the sphere since $\mathbf{e}_\varphi, \mathbf{e}_\theta \perp \mathbf{e}_r$. The zonal component v_1 and the meridional component v_2 are specified in the following two examples.

Temporally constant flow field. This test case, called solid body rotation in what follows, can be found in [PBR13] or, historically older but with the correct sign of v_2, in [WDH+92]. It reads

$$\begin{aligned} v_1(\varphi, \theta) &= \quad 2\pi(\sin\theta\cos\alpha + \cos\varphi\cos\theta\sin\alpha), \\ v_2(\varphi, \theta) &= -2\pi\sin\varphi\sin\alpha. \end{aligned} \tag{5.20}$$

The velocity field is solenoidal, since

$$\begin{aligned} \nabla_{\mathbf{p}} \cdot \mathbf{v} &= \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta\, v_2(\varphi,\theta)) + \frac{1}{\sin\theta}\frac{\partial u_1(\varphi,\theta)}{\partial\varphi} \\ &= 2\pi\sin\varphi\,\sin\alpha\,\frac{\cos\theta}{\sin\theta} - 2\pi\sin\varphi\,\sin\alpha\,\frac{\cos\theta}{\sin\theta} \quad = 0. \end{aligned}$$

The rotation angle α provides the orientation of the flow, see Figure 5.26. For $\alpha = 0$ the configuration moves around the sphere in the xy-plane, that is, across the equator. As the angle α is increased, the circle gets smaller until there is no movement anymore for $\alpha = \frac{\pi}{2}$. For $\alpha > \frac{\pi}{2}$ we observe a change of direction until for $\alpha = \pi$ the configuration follows the same path as for $\alpha = 0$ but in the opposite direction. The same pattern of circles becoming smaller until there is a change of direction can be observed for $\alpha \in (\pi, 2\pi)$.

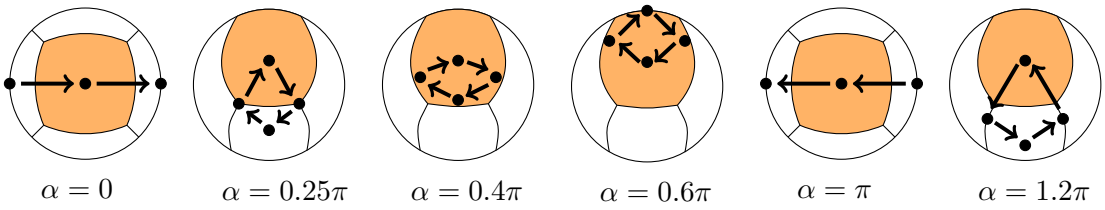


Figure 5.26: *Motion described by the velocity field* $\mathbf{v}(\alpha)$*. Observing the orange area, we look in the positive x-direction; the y-axis points to the left and the z-axis upwards.*

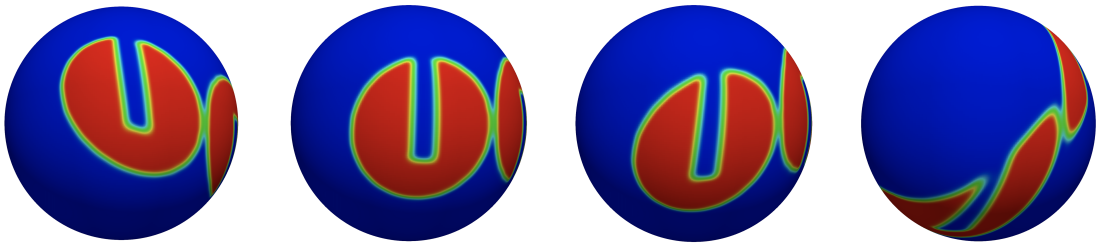
Deformational flow field. The velocity field can also be time-dependent. With minor variations such a deformational flow field can be found in [LSPT12, PBR13]. The coefficients v_1 and v_2 read

$$\begin{aligned} v_1(\varphi, \theta, t) &= 2\beta\sin^2(\varphi')\cos(\theta)\sin(\theta)\cos(\pi t) + 2\pi\sin(\theta), \\ v_2(\varphi, \theta, t) &= \beta\sin(2\varphi')\sin(\theta)\cos(\pi t), \end{aligned} \tag{5.21}$$

where $\varphi' = \varphi - 2\pi t$. The velocity field has also been constructed to be solenoidal, so that

$$\begin{aligned}\nabla_{\mathbf{p}} \cdot \mathbf{v} &= \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}(\sin^2\theta)\,\beta\sin(2\varphi')\cos(\pi t) + \frac{1}{\sin\theta}\frac{\partial}{\partial\varphi}(\sin^2(\varphi'))\,2\beta\cos\theta\sin\theta\cos(\pi t)\\ &= 2\beta\cos(\pi t)\cos(\theta)\left[\sin(2\varphi') - 2\cos(\varphi')\sin(\varphi')\right] \quad = \quad 0.\end{aligned}$$

The coefficient β describes the degree of deformation, see Figure 5.31 below.



$\beta = -2$ $\beta = 0.5$ $\beta = 2$ $\beta = 10$

Figure 5.31: *Deformation of the Slotted Cylinders for different β at time $t = 0.1$.*

The time-dependence increases the computational effort, since parameter t is passed in each time step and the convection matrix has to be reassembled every time.

Remark 5.26. In the underlying sources, [WDH+92], [LSPT12] and [PBR13], an 'American convention' for the spherical coordinates is used. To ensure consistency within this work, however, we modify and adapt those expressions to use the common physical convention. Technical details are described in Section 10.1.

Numerical results

Temporally constant velocity field. We start with a solid body rotation defined by the temporally constant velocity field (5.20) with $\alpha = 0$. The usage of other α did not lead to new findings.

First, we apply the baseline Galerkin approach and use Crank-Nicolson as time stepping scheme. We switch from the levels 3-7 to the levels 5-9, because especially for the discontinuous configuration visual improvements can still be observed on the higher levels. On the time interval $[0, 1]$ and starting from level 5, we use 200, 400, 800, 1600 and 3200 steps. The results for both he Gaussian Hill and the Slotted Cylinder configuration can be found in Table 5.32.

The EOCs in Table 5.32 could be expected. For the smooth configuration the EOC with respect to the L^2-norm is 2 as in the case of the extended Jeffery equation, see Table 5.17. For the discontinuous configuration, the EOC is naturally significantly reduced. Moreover, we documented the range $[u_{\min}, u_{\max}]$ at the final time $T = 1$. For the discontinuous configuration there are strong overshoots and undershoots, see Figure 5.33 below.
For the Gaussian Hill here we refrain from visualization, since the final configuration throughout looks similar to the initial configuration illustrated in Figure 5.18. However, there are undershoots even for this smooth configuration.

Therefore, all the following simulations are performed using MCL combined with SSP-RK2. The number of time steps is chosen in a way that the CFL-condition

	Gaussian Hill			Slotted Cylinders		
level	L^2-err	EOC	$[u_{\min}, u_{\max}]$	L^2-err	EOC	$[u_{\min}, u_{\max}]$
5	1.09e-2	-	[-1.51e-3, 0.9980]	6.13e-1	-	[-7.76e-1, 1.5897]
6	2.70e-3	2.01	[-3.65e-4, 0.9999]	4.86e-1	0.34	[-6.01e-1, 1.8009]
7	6.75e-4	2.00	[-9.49e-5, 1.0000]	3.58e-1	0.44	[-6.49e-1, 1.5461]
8	1.69e-4	2.00	[-2.52e-5, 1.0000]	3.05e-1	0.23	[-5.46e-1, 1.5235]
9	4.22e-5	2.00	[-6.30e-6, 1.0000]	2.38e-1	0.35	[-5.38e-1, 1.5127]

Table 5.32: *Results for the baseline Galerkin scheme in combination with Crank-Nicolson when the temporally constant flow field ($\alpha = 0$) is used.*

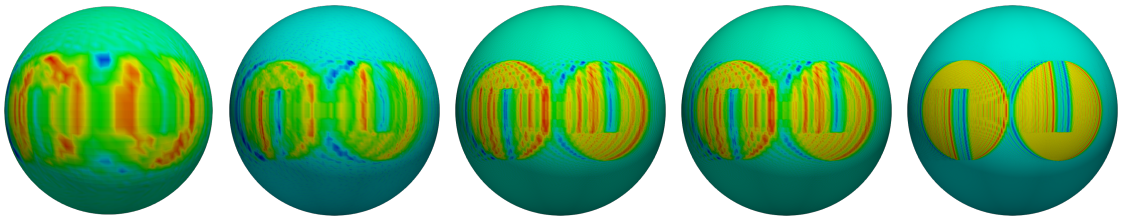


Figure 5.33: *(cf. Table 5.32) Visual results for the Slotted Cylinder configuration at the levels 5 to 9 after a full cycle. The data range is adapted to the extremal values.*

(4.16) is met. Employing the same temporally constant flow field as before, we use 540, 1080, 2160, 4320 and 8640 steps for the levels 5 to 9.
The absolute errors in the L^2-norm, the corresponding EOCs and the minimal and maximal value at the final time of each level are documented in Table 5.34.

	Gaussian Hill			Slotted Cylinders		
level	L^2-err	EOC	$[u_{\min}, u_{\max}]$	L^2-err	EOC	$[u_{\min}, u_{\max}]$
5	2.53e-2	-	[2.24e-9, 0.8830]	5.60e-1	-	[-2.16e-09, 0.8474]
6	4.83e-3	2.39	[2.13e-9, 0.9608]	4.07e-1	0.46	[-8.10e-11, 0.9952]
7	9.59e-4	2.33	[2.09e-9, 0.9870]	2.93e-1	0.48	[-7.32e-14, 1.0000]
8	1.96e-4	2.29	[2.07e-9, 0.9957]	2.22e-1	0.40	[-2.80e-15, 1.0000]
9	4.05e-5	2.27	[2.06e-9, 0.9986]	1.69e-1	0.40	[-4.09e-17, 1.0000]

Table 5.34: *Results for MCL in combination with SSP-RK2 when the temporally constant flow field ($\alpha = 0$) is used.*

For the Gaussian Hill the correct physical bounds are kept, while there are still undershoots for the Slotted Cylinders. However, the undershoots are extremely

small. At the higher levels they are even in the range of machine accuracy. Moreover, the overshoots can no longer be identified in the visualization, see Figure 5.35.

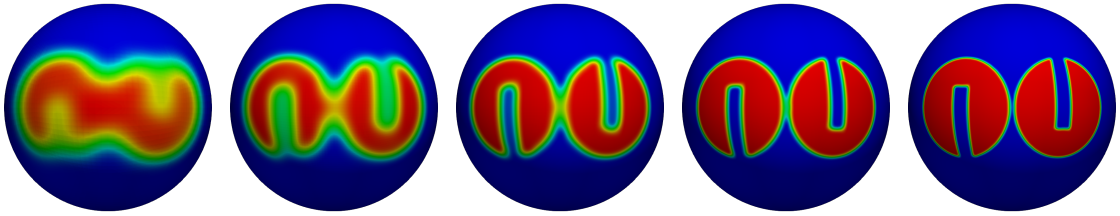


Figure 5.35: *(cf. Table 5.34) Visual results for the Slotted Cylinder configuration at the levels 5 to 9 after a full cycle. The data range is adapted to the extremal values.*

Deformational flow field. Finally, we use the time-dependent velocity field given by (5.21). Using $\beta = 10$ this is a relatively demanding test case with strong deformation. The MCL algorithm is applied in combination with SSP-RK2 as time stepping scheme. Again, the number of time steps is determined by the CFL-condition (4.16). The CFL-condition is influenced by the velocity field. Starting from level 5 we even use 2000, 4000, 8000, 16000 and 32000 steps for the interval $[0, 1]$. The Tables and Figures 5.36 and 5.37 show the results.

Gaussian Hill. For the smooth Gaussian Hill initial configuration we obtain:

level	L^2-err	EOC	linear transformation		quadratic transformation		
			$u_{\min}$				$u_{\max}$
			(global)	$(T = 1)$	(global)	$(T = 1)$	
5	3.05e-1	-	-4.68e-8	1.25e-8	2.06e-9	1.35e-8	0.39
6	1.62e-1	0.91	2.06e-9	2.44e-9	2.06e-9	2.45e-9	0.62
7	5.91e-2	2.46	2.06e-9	2.26e-9	2.06e-9	2.25e-9	0.82
8	1.31e-2	2.28	2.06e-9	2.14e-9	2.06e-9	2.14e-9	0.93
9	2.24e-3	2.53	2.06e-9	2.09e-9	2.06e-9	2.09e-9	0.98

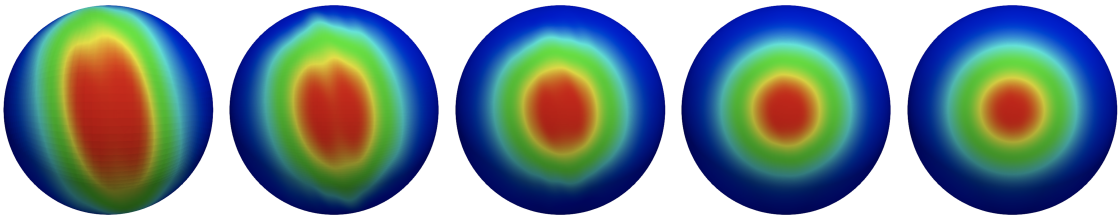


Table and Figure 5.36: *Quantitative and qualitative results for the Gaussian Hill after a full cycle at levels 5-9 for a deformational flow field if MCL is applied.*

Slotted Cylinders. For the discontinuous configuration the results read:

level	L^2-err	EOC	linear transformation $u_{\min}$ (global)	linear transformation $u_{\min}$ $(T=1)$	quadratic transformation $u_{\min}$ (global)	quadratic transformation $u_{\min}$ $(T=1)$	$u_{\max}$
5	0.74	-	-1.29e-06	-1.70e-07	-3.59e-07	-9.19e-09	0.72
6	0.58	0.35	-1.03e-07	-4.08e-09	-3.11e-08	-3.15e-10	0.97
7	0.45	0.38	-1.42e-08	-2.51e-11	-4.56e-09	-5.41e-12	1.00
8	0.33	0.45	-1.78e-09	-2.16e-13	-5.72e-10	-2.39e-14	1.00
9	0.24	0.42	-2.23e-10	-3.24e-16	-7.17e-11	-1.35e-16	1.00

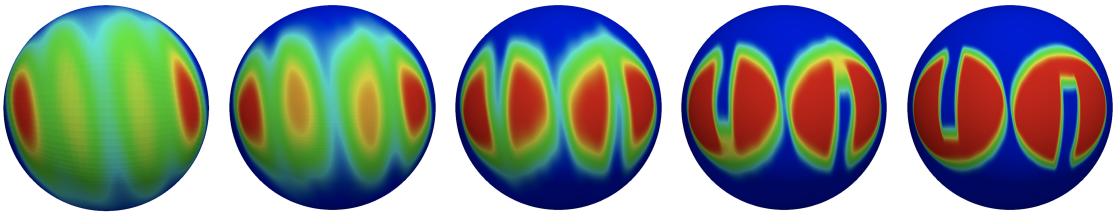


Table and Figure 5.37: *Quantitative and qualitative results for the Slotted Cylinders after a full cycle at levels 5-9 for a deformational flow field if MCL is applied.*

Despite limiting, we observe undershoots for the Slotted Cylinder configuration. This can be explained by the fact that the surface elements are not given exactly, but that they are mapped to 2d elements. For this reason, we compare the linear and the quadratic transformation although only the space $\mathbf{Q}_1$ is used.

With respect to the two different transformations no difference can be found for the L^2-errors, for the maximal values $u_{\max}$ at the final time and for the visualization. However, slight differences are identified for the minimal values $u_{\min}$. For both transformations the global minimum, that is, the minimal value over all time steps and the minimal value at the final time $T = 1$ are compared. Throughout the magnitude of the negative values at the end of the simulation is significantly smaller than the global minimum. In addition, the quadratic transformation is on average one decimal place more accurate than the linear transformation.

While the improvement of the L^2-error does not seem that impressive for the discontinuous configuration, the visualization demonstrates that the shape is restored significantly better from level to level. Furthermore, we recognize that in case of the deformational flow fields it is even more challenging to restore the original shape.

6 Techniques for the coupled system

This chapter covers technique for the full Fokker-Planck equation as well as theory for the Navier-Stokes equations. The focus of the thesis is on the Fokker-Planck equation for $\psi = \psi(\mathbf{x}, \mathbf{p}, t)$, concretely

$$\begin{aligned} \frac{\partial \psi}{\partial t} + \nabla_{\mathbf{x}} \cdot (\mathbf{u}\psi) + \nabla_{\mathbf{p}} \cdot (\dot{\mathbf{p}}\psi) &= \Delta_{\mathbf{p}}(D_r \psi) \qquad \text{in } \Omega \times \mathbb{S}^2 \times (0, T], \\ \text{where} \quad \dot{\mathbf{p}} &= \mathbf{W}\mathbf{p} + \lambda_e \left[\mathbf{D}\mathbf{p} - (\mathbf{D} : (\mathbf{p} \otimes \mathbf{p}))\, \mathbf{p}\right], \\ \psi &= \psi_0 \quad \text{in } \Omega \times \mathbb{S}^2 \times \{0\}, \\ \text{and} \quad \psi &= \psi_{\text{bc}} \quad \text{on } \partial\Omega \times \mathbb{S}^2 \times (0, T]. \end{aligned}$$

Considering this system, the Fokker-Planck equation (2.2) is complemented by the Jeffery equation (2.5) and by generic initial and boundary conditions. We introduced a splitting of the full Fokker-Planck equation into an advection equation in the physical space and a convection equation in the configuration space already at the beginning of this thesis, see the equations (2.4a) and (2.4b). This splitting is not only highly recommended, but even necessary to solve the equation numerically within a tolerable time. Therefore, the basic principles of operator splitting are presented in Section 6.1. While Section 6.1.1 covers the topic in general, Section 6.1.2 is dedicated more specifically to the Fokker-Planck equation, including aspects of implementation.

For more advanced problems the velocity field $\mathbf{u}$ is not given analytically anymore, but it is measured in experiments or calculated numerically by the Navier-Stokes equations. For fiber suspensions a suitable formulation of them reads

$$\begin{aligned} \rho \left(\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u} \right) = \rho \mathbf{g} - \nabla p + \nabla \cdot \boldsymbol{\tau}, \quad \nabla \cdot \mathbf{u} = 0 \qquad \text{in } \Omega \times (0, T], \\ \text{where} \quad \boldsymbol{\tau} = 2\mu_I (\mathbf{D} + N_p \mathbb{A} : \mathbf{D} + N_s (\mathbf{D}\mathbf{A} + \mathbf{A}\mathbf{D})), \end{aligned}$$

see (2.23), (2.29) and (2.33). When we couple the Fokker-Planck equation to the Navier-Stokes equations, limiting the individual subproblems of the former is not enough anymore. To obtain a conservative scheme, additionally the equation as a whole has to be modified. An algorithm for this is given in Section 6.2.

The Jeffery equation determines vector $\dot{\mathbf{p}}$ based on the tensors $\mathbf{D} = \mathbf{D}(\nabla \mathbf{u})$ and $\mathbf{W} = \mathbf{W}(\nabla \mathbf{u})$. Consequently, the Jacobians $\nabla \mathbf{u}$ have to be reconstructed from the velocity field $\mathbf{u}$. Section 6.3 addresses a possible gradient recovery. Finally, Section 6.4 describes requirements and methods to solve the Navier-Stokes equations.

6.1 Operator splitting

6.1.1 Basic idea of operator splitting

Classification. Operator splitting is a technique to solve ODEs and PDEs by breaking them down into smaller and simpler subproblems. It can be interpreted as a time stepping method. Let us consider the generic equation

$$\frac{\mathrm{d}u}{\mathrm{d}t} + \mathcal{L}u = f \tag{6.1}$$

and a two-term splitting

$$\mathcal{L} = \mathcal{L}_1 + \mathcal{L}_2. \tag{6.2}$$

On the one hand, we distinguish between dimension and physical splitting [HV03]. Dimension splitting separates the operator $\mathcal{L}$ with respect to the space dimensions. Instead of solving one huge problem in d dimensions, we solve d one-dimensional problems. This methodology is also known as alternating direction approach [HV03]. Its origins go back to the 1950s, when 'alternating direction implicit' (ADI) was developed for parabolic and elliptic equations using finite differences [PR55].
Physical splitting means that we split between the terms associated with different physical phenomena such as convection, diffusion and reaction.

On the other hand, we distinguish between differential and algebraic splitting [QV08]. Using differential splitting, the decomposition is realized on the continuous level, that is, $\mathcal{L}_1$ and $\mathcal{L}_2$ represent different differential operators of a PDE. The splitting of the boundary conditions has to be kept in mind as well.
Algebraic splitting is applied on the (semi-)discrete level. In this case, the operator $\mathcal{L}$ from (6.2) corresponds to the matrix $\mathcal{L} = M^{-1}A$, where mass matrix M and matrix A result from the space discretization. Since we are on the algebraic stage, the matrices already include the boundary conditions [KH15].

Another name for the operator splitting is fractional-step algorithm. It is often used synonymously in the literature. Following [QV08], however, call the decomposition (6.2) itself operator splitting but the concrete schemes fractional-step methods. Three of them are listed below.

Specific schemes. A basic method is the Marchuk-Yanenko splitting [QV08]

$$\begin{cases} \dfrac{u^{n+\frac{1}{2}} - u^n}{\Delta t} + \mathcal{L}_1 u^{n+\frac{1}{2}} = 0, & \text{(6.3a)} \\ \dfrac{u^{n+1} - u^{n+\frac{1}{2}}}{\Delta t} + \mathcal{L}_2 u^{n+1} = f^n, & \text{(6.3b)} \end{cases}$$

where $n \geq 0$. We have two implicit problems. They are similar to backward Euler steps, but it is not intuitive that the time step size Δt is used twice even though only the interval (u^n, u^{n+1}) is considered. However, adding (6.3a) and (6.3b) yields

$$\frac{u^{n+1} - u^n}{\Delta t} + \mathcal{L}u^{n+1} = f^n + \mathcal{L}_1(u^{n+1} - u^{n+\frac{1}{2}}). \tag{6.4}$$

Equation (6.3b) can be reformulated to

$$u^{n+1} - u^{n+\frac{1}{2}} = \Delta t\,(f^n - \mathcal{L}_2 u^{n+1}). \tag{6.5}$$

Substituting (6.5) into (6.4), we then find that [QV08]

$$\frac{u^{n+1} - u^n}{\Delta t} + \mathcal{L} u^{n+1} = f^n + \Delta t \mathcal{L}_1 (f^n - \mathcal{L}_2 u^{n+1}). \tag{6.6}$$

Expression (6.6) demonstrates that the Marchuk-Yanenko splitting is well-posed and first-order accurate, in the sense that the splitting error is $\mathcal{O}(\Delta t)$. It is straightforward to extend (6.6) to a splitting with more than two operators. However, the intermediate results are no consistent approximations to the exact solution [HV03].

This is the other way round for the Peaceman-Rachford scheme [PR55, Glo03, QV08]

$$\begin{cases} \dfrac{u^{n+\frac{1}{2}} - u^n}{\Delta t/2} + \mathcal{L}_1 u^{n+\frac{1}{2}} = f^{n+\frac{1}{2}} - \mathcal{L}_2 u^n, & \text{(6.7a)} \\ \dfrac{u^{n+1} - u^{n+\frac{1}{2}}}{\Delta t/2} + \mathcal{L}_2 u^{n+1} = f^{n+\frac{1}{2}} - \mathcal{L}_1 u^{n+\frac{1}{2}}, . & \text{(6.7b)} \end{cases}$$

We recognize that there is a symmetry between the operators $\mathcal{L}_1$ and $\mathcal{L}_2$. The scheme is consistent also at the intermediate levels, since both operators are incorporated in both steps. Eliminating $u^{n+\frac{1}{2}}$ we find that [QV08]

$$\frac{u^{n+2} - u^n}{\Delta t} + \frac{1}{2}\mathcal{L}(u^n + u^{n+1}) = f^{n+\frac{1}{2}} - (\Delta t)^2 \mathcal{L}_1 \mathcal{L}_2 \left(\frac{u^{n+1} - u^n}{\Delta t} \right). \tag{6.8}$$

Consequently, the Peaceman-Rachford scheme is second-order accurate.
Another second-order method is the three-step Strang splitting [KH15]

$$\begin{cases} \dfrac{u^{n+\frac{1}{4}} - u^n}{\Delta t/2} + \mathcal{L}_1 u^{n+\frac{1}{4}} = 0, & \text{(6.9a)} \\ \dfrac{u^{n+\frac{3}{4}} - u^{n+\frac{1}{4}}}{\Delta t} + \mathcal{L}_2 u^{n+\frac{3}{4}} = 0, & \text{(6.9b)} \\ \dfrac{u^{n+1} - u^{n+\frac{3}{4}}}{\Delta t/2} + \mathcal{L}_1 u^{n+1} = f^n. & \text{(6.9c)} \end{cases}$$

A fourth well-known algorithm is Glowinski's θ-method [Glo03], which reads

$$\begin{cases} \dfrac{u^{n+\theta} - u^n}{\theta \Delta t} + \mathcal{L}_1 u^{n+\theta} = f^n - \mathcal{L}_2 u^n, & \text{(6.10a)} \\ \dfrac{u^{n+1-\theta} - u^{n+\theta}}{(1-2\theta)\Delta t} + \mathcal{L}_2 u^{n+1-\theta} = f^{n+\theta} - \mathcal{L}_1 u^{n+\theta}, & \text{(6.10b)} \\ \dfrac{u^{n+1} - u^{n+1-\theta}}{\theta \Delta t} + \mathcal{L}_1 u^{n+1} = f^{n+1-\theta} - \mathcal{L}_2 u^{n+1-\theta}, & \text{(6.10c)} \end{cases}$$

where $\theta \in (0, \frac{1}{2})$. In each time step one operator is treated explicitly and one operator is treated implicitly [KH15]. This three-stage splitting scheme is second-order accurate for $\theta = 1 - \frac{\sqrt{2}}{2}$. A detailed analysis can be found in [Glo03].

Advantages and difficulties. Operator splitting gives a division into subproblems. This has several advantages. Dimension splitting can simplify the matrix structure. For instance, a pentadiagonal matrix might be reduced to a tridiagonal matrix, which then allows to employ particularly efficient strategies such as the Thomas algorithm instead of an LU decomposition.
Applying physical splitting the resulting subproblems can be solved individually and customized discretizations can be choosen. For example, the convective term can be treated explicitly, while the diffusive and reactive terms are treated implicitly [KH15]. Another helpful tool is the subtime stepping, where the number of time steps may vary between the subproblems and can be adapted individually, so that effort can be saved for less restrictive requirements to Δt.
Thanks to smaller subproblems nonlinear systems too large to handle due to inherent couplings are avoided from the start [HV03]. Nonlinear terms, as they arise in the Navier-Stokes equations, are isolated.
A further aspect is the modularity of the corresponding code. Existing code for basic equations can be reused. The overview during programming and debugging is facilitated. Last but not least, the possibility for parallel computing is given.

Despite the overwhelming advantages of the operator splitting, we should keep in mind that it can also have an adverse effect on accuracy and robustness [KH15]. Furthermore, by the subtime stepping some level of coupling is given up. For the coupled FPE-NSE system we discuss this in Section 7.2.2.2.

6.1.2 Application to Fokker-Planck equation

Without operator splitting the full time-dependent FPE with its six dimensions would be computationally intractable. This is due to the 'curse of dimensionality', which means that the computational effort grows exponentially with each additional dimension [KS09a, LOP11]. Splitting the FPE into two subproblems improves the situation. With the natural choice to separate the components of space and orientation, the splitted FPE reads

$$\frac{\partial \psi}{\partial t} + \mathcal{L}\psi \;=\; \frac{\partial \psi}{\partial t} + \mathcal{L}_{\mathbf{x}}\psi + \mathcal{L}_{\mathbf{p}}\psi = 0,$$
$$\text{where } \mathcal{L}_{\mathbf{x}}\psi = \nabla_{\mathbf{x}} \cdot (\mathbf{u}\psi) \text{ and } \mathcal{L}_{\mathbf{p}}\psi = \nabla_{\mathbf{p}} \cdot (\dot{\mathbf{p}}\psi) - \Delta_{\mathbf{p}}(D_r\psi).$$

This alternating direction approach results in a homogeneous spatial advection equation, on the one hand, and a space-independent FPE, on the other hand, see equations (2.4a) and (2.4b) in the modeling chapter. We interpret this decomposition as dimensional splitting, since both the $\mathbf{x}$- and the $\mathbf{p}$-component describe points in space [GT13].
For simplicity, we use each of the two operators $\mathcal{L}_{\mathbf{x}}$ and $\mathcal{L}_{\mathbf{p}}$ exactly once per time step. This is similar to the Marchuk-Yanenko method, even if no implicit treatment is required in our case. For instance, in [HO06] the FPE is solved using the second-order accurate Strang splitting. However, here it is acceptable if we are only first-order accurate, i.e., if we have a splitting error of $\mathcal{O}(\Delta t)$. On the one hand, this is reasonable compared to the modeling error, and since the changes over time are relatively small. On the other hand, the spatial error always dominates, since for

our space discretization including the limiting only $\mathcal{O}(\sqrt{h})$ can be expected [Loh19]. Nevertheless, in practical computations a higher-order time-discretization is applied in each splitting step and better convergence results can be observed, see Section 7.1.

In analogy to the standard finite element approach (3.4), we choose the tensor product approximation

$$\psi(\mathbf{x},\mathbf{p},t) \;\approx\; \psi_h(\mathbf{x},\mathbf{p},t) \;=\; \sum_{i=1}^{N}\sum_{k=1}^{M} \psi_{i,k}(t)\,\varphi_i(\mathbf{x})\,\widetilde{\varphi_k}(\mathbf{p}),$$

where $\varphi(\mathbf{x})$ and $\widetilde{\varphi_k}(\mathbf{p})$ are continuous Lagrange basis functions, so that

$$\psi_h(\mathbf{x}_i,\mathbf{p}_k,t) = \psi_{i,k}(t).$$

The coefficient $\psi_{i,k}(t)$ describes the probability that a fiber located at the node $\mathbf{x}_i \in \Omega$ has orientation $\mathbf{p}_k \in \mathbb{S}^2$ at time t. Using the notations $\psi_h(\mathbf{x}_i,\mathbf{p},t) = \psi_{i,*}(t)$ and $\psi_h(\mathbf{x},\mathbf{p}_k,t) = \psi_{*,k}(t)$, the two subproblems to be solved in each time step $(t^n, t^{n+1}]$ are

1.) For each orientation $\mathbf{p}_k$ solve the homogeneous spatial advection equation

$$\frac{\partial \psi_{*,k}}{\partial t} + \nabla_{\mathbf{x}} \cdot (\mathbf{u}\psi_{*,k}) = 0 \quad \text{in } \Omega, \tag{6.11}$$

where $\psi_{*,k}(t^n) = \psi_{*,k}^n, \quad \widetilde{\psi}_{*,k}^{n+1} = \psi_{*,k}(t^{n+1})$.

2.) In each grid point $\mathbf{x}_i$ solve the space-independent FPE

$$\frac{\partial \psi_{i,*}}{\partial t} + \nabla_{\mathbf{p}} \cdot (\dot{\mathbf{p}}_i \psi_{i,*}) = \Delta_{\mathbf{p}}(D_r \psi_{i,*}) \quad \text{on } \mathbb{S}^2, \tag{6.12}$$

where $\psi_{i,*}(t^n) = \widetilde{\psi}_{*,k}^{n+1}, \quad \psi_{i,*}^{n+1} = \psi_{i,*}(t^{n+1})$.

In the literature, authors often use a heterogeneous alternating direction approach, that is, their approaches for the physical space and the configuration space differ [KS09a, Loh16b]. In this thesis, however, we apply the same methodology for both subproblems. This is an indicator for the strength and the wide applicability of the MCL method.

Implementation. For practical purposes, we store the intermediate results, i.e., the coefficients, in the huge $N \times M$-matrix

$$\mathbf{\Psi} = \{\psi_{i,k}\}_{\substack{i=1,\dots,N \\ k=1,\dots,M}}. \tag{6.13}$$

Updating the i^{th} row corresponds to solving the space-independent FPE for grid point $\mathbf{x}_i$, updating the k^{th} column to solving a homogeneous advection equation for orientation $\mathbf{p}_k$.

Each row of the matrix is part of the configuration space, each column part of the physical space, see Figure 6.1. The updates are realized sequentially.

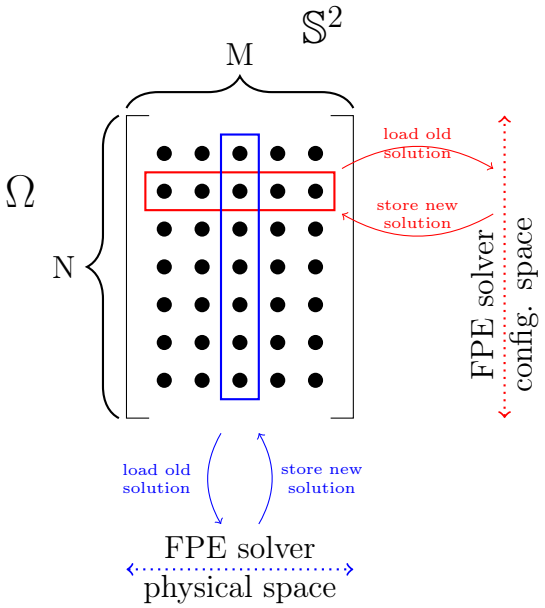


Figure 6.1: *Schematic illustration of the operator splitting for the FPE [Kne06].*

Considering the two subproblems (6.11) and (6.12) again, there are the velocity fields $\mathbf{u}$ without index and $\dot{\mathbf{p}}_i$ with index. This corresponds to practical experiences during assembly. For the term $\nabla_{\mathbf{x}} \cdot (\mathbf{u}\psi)$ we obtain one $N \times N$-matrix, while an $M \times M$-convection matrix has to be assembled for each grid point i.

Last but not least, it has to be weighed up how to choose the number M of spherical grid points in relation to the number N of spatial grid points. In 3d, typically $N > M$. Then, in each time step, we have to solve many space-independent FPEs with relatively small $M \times M$-FE matrices. At the same time, only M advection equations have to be solved, but they produce large linear systems consisting of $N \times N$-FE matrices.

Outlook: Application to the Navier-Stokes equations. To take into account the mutual influence between the fluid and the fibers, finally the FPE and the NSE are coupled. Various types of splitting are also widely used for the NSE.

Since the treatment of the nonlinearity and the incompressibility is particularly challenging, these components should be treated individually. By doing so, also the velocity and the pressure can be handled in a segregated manner. A further option might be to decouple the convective term and the diffusive term. For example, one option is to use Glowinski's θ-scheme (6.10) as an operator splitting approach. Choosing the operators $\mathcal{L}_1$ and $\mathcal{L}_2$ appropriately, see, e.g., [KH15], we have to solve two Stokes problems and a nonlinear convection-diffusion equation.

Section 6.4.2 presents some basic techniques to solve the NSE numerically.

6.2 Limiting for systems

In Chapter 4, we introduced the MCL methodology, which is designed to ensure bound-preserving and conservative numerical results. Switching to the full FPE, however, the basic approach does not ensure conservation of mass anymore.

The basic approach (4.9) for limited fluxes reads $f_{ij}^* = \alpha_{ij} f_{ij}$. Application of MCL to the advection equations (6.11) produces vectors of limited fluxes $f_{ij,k}^* = \alpha_{ij,k} f_{ij,k}$, which include the orientation angle k as well.
For this reason, we split up the limiting into two steps. We do not throw away the benefits of the previous limiting but still use MCL to obtain $f_{ij,k}^*$. In the second step, $f_{ij,k}^*$ is used as auxiliary flux and enhanced to $f_{ij,k}^{**} = \beta_{ij,k} f_{ij,k}^*$. We derive a requirement for $f_{ij,k}^{**}$ and a formula for the associated $\beta_{ij,k}$.

In our framework, the conservation of mass is equivalent to the normalization condition (2.3b). In particular, the additional requirement for the discrete ODF is

$$\int_{\mathbb{S}^{d-1}} \psi_{i,*}^{n+1}(\mathbf{p})\,\mathrm{d}\mathbf{p} \overset{(3.14)}{=} \sum_{k=1}^{M} m_k \psi_{i,k}^{n+1} \overset{!}{=} 1, \tag{6.14}$$

where $m_k = \int_{\mathbb{S}^{d-1}} \widetilde{\varphi}_k(\mathbf{p})\,\mathrm{d}\mathbf{p}$ describes an entry of the lumped mass matrix with respect to the sphere. Using the flux corrected approach, in analogy to (4.17) within a forward Euler time we obtain

$$\psi_{i,k}^{n+1} = \psi_{i,k}^{n} + \frac{\Delta t}{m_i} \sum_{j \in \mathcal{N}_i^*} \left[(d_{ij} - k_{ij})(\psi_{j,k}^n - \psi_{i,k}^n) + f_{ij,k}^{**} \right]. \tag{6.15}$$

Since each explicit SSP-RK method consists of forward Euler steps, the following results can be transferred. Substituting (6.15) into (6.14), the mass on a sphere is

$$\begin{aligned}
\int_{\mathbb{S}^{d-1}} \psi_{i,*}^{n+1}(\mathbf{p})\,\mathrm{d}\mathbf{p} \;&=\; \sum_{k=1}^{M} m_k \psi_{i,k}^{n+1} \\
&= \underbrace{\sum_{k=1}^{M} m_k \psi_{i,k}^{n}}_{=1} + \frac{\Delta t}{m_i} \sum_{j \in \mathcal{N}_i^*} \left((d_{ij} - k_{ij}) \Big(\underbrace{\sum_{k=1}^{M} m_k \psi_{j,k}^n}_{=1} - \underbrace{\sum_{k=1}^{M} m_k \psi_{i,k}^n}_{=1} \Big) + \sum_{k=1}^{M} m_k f_{ij,k}^{**} \right) \\
&= 1 + \frac{\Delta t}{m_i} \sum_{j \in \mathcal{N}_i^*} \sum_{k=1}^{M} m_k f_{ij,k}^{**},
\end{aligned} \tag{6.16}$$

where it was assumed that the normalization property is given at time 'n'. Our goal is that the mass of each sphere remains one at time 'n+1'. With respect to (6.16) this is satisfied if

$$\sum_{k=1}^{M} m_k f_{ij,k}^{**} \overset{!}{=} 0. \tag{6.17}$$

For the low-order method, where $\alpha_{ij,k} = 0$ and hence $f_{ij,k}^{**} = 0$, this is automatically fulfilled. For the general case we adapt the scaling strategy from [LKSM17] to determine the coefficients $\beta_{ij,k}$ in a way, that (6.17) is satisfied. This might be summarized in the following algorithm:

Algorithm 6.1 (Mass limiting)**.**

1.) Prelimiting: $f^*_{ij,k} = \alpha_{ij,k} f_{ij,k}$

2.) Additional limiting: $f^{**}_{ij,k} = \beta_{ij,k} f^*_{ij,k}$, *where*

$$\beta_{ij,k} := \begin{cases} \frac{-\sum \min\{0, m_k f^*_{ij,k}\}}{\sum \max\{0, m_k f^*_{ij,k}\}} & \text{if } \sum m_k f^*_{ij,k} > 0 \text{ and } m_k f^*_{ij,k} > 0, \\ \frac{\sum \max\{0, m_k f^*_{ij,k}\}}{-\sum \min\{0, m_k f^*_{ij,k}\}} & \text{if } \sum m_k f^*_{ij,k} < 0 \text{ and } m_k f^*_{ij,k} < 0, \\ 1 & \text{otherwise.} \end{cases}$$

The $\beta_{ij,k}$ satisfy (6.17). For each entry of the coefficient matrix Ψ the product $m_k f^*_{ij,k}$ is calculated and then it is checked row by row whether the sum of these products is positive or negative. If the sum equals zero, condition (6.17) is already met by $f^*_{ij,k}$ and $\beta_{ij,k}$ is set to one. If the sum is positive, the positive entries have to be reduced by weighting them appropriately. This is done by choosing the weight as the quotient of the sum of the negative and the sum of the positive entries in the row, combined with the correct sign for $\beta_{ij,k}$. In case of a negative row sums the procedure is analogous.

Algebraic example. A small example may illustrate how the additional limiting works. Let

$$m_k = \begin{pmatrix} 1 & 2 & 4 & 0.5 \end{pmatrix} \quad \text{and} \quad f^*_{ij,k} = \begin{pmatrix} 2 & -3 & 0 & -1 \end{pmatrix},$$

so that

$$m_k f^*_{ij,k} = \begin{pmatrix} 2 & -6 & 0 & -0.5 \end{pmatrix} \quad \text{and} \quad \sum_k m_k f^*_{ij,k} = -4.5 < 0.$$

Thus, the weighting factor for the negative entries of $f^*_{ij,k}$ reads

$$\beta_{ij,k} = \frac{\text{sum of positive entries of } m_k f^*_{ij,k}}{-\text{sum of negative entries of } m_k f^*_{ij,k}} = \frac{4}{13},$$

so that $f^{**}_{ij,k} = \beta_{ij,k} f^*_{ij,k} = \begin{pmatrix} 2 & -\frac{12}{13} & 0 & -\frac{4}{13} \end{pmatrix}$. Hence,

$$\sum_{k=1}^{M} m_k f^{**}_{ij,k} = 2 - \frac{24}{13} - \frac{2}{13} = 0,$$

which is the desired result.

Additional coupling. A key change due to the second limiting step is that the single equations cannot be solved independently from each other anymore as already the notation with the three indices i, j and k demonstrates. Originally, solving an advection equation in space corresponded to updating the respective column of the coefficient matrix Ψ. Now we need to know $\alpha_{ij,k}$ for all indices k to calculate $\beta_{ij,k}$ for a given i. All in all, the additional limiting step requires more exchange of information.

6.3 Gradient Recovery

Solving the NSE we obtain a numerical solution u_h. To calculate the tensors $\mathbf{D}$ and $\mathbf{W}$, the gradient ∇u_h has to be reconstructed. Gradient recovery is a common task in computational mathematics. Different options are, for example, presented in [Don18]. We use an approach based on an L^2-projection, see [Kuz10].

Definition 6.2 (L^2-projection, [Loh19])**.** *An L^2-projection projects an arbitrary function $u \in L^2(\Omega)$ into an FE space $V_h \subset L^2(\Omega)$. The L^2-projection operator $\mathcal{P}_h$ is defined by $\mathcal{P}_h : L^2(\Omega) \to V_h$, $u \mapsto \mathcal{P}_h u := u_h$, where*

$$(\mathcal{P}_h u, \varphi_h)_{L^2(\Omega)} = (u, \varphi_h)_{L^2(\Omega),} \quad \textit{that is,} \quad \int_\Omega (u_h - u)\, \varphi_h = 0 \quad \forall \varphi_h \in V_h.$$

Remarks 6.3 (L^2-projection)**.**

i) (Applications.) An L^2-projection can be used to project a given function, e.g., a discontinuous initial solution of a time-dependent problem, to an FE space V_h.

ii) (Properties.) With the admissible test function $\varphi_h \equiv 1$ we obtain $\int_\Omega u_h = \int_\Omega u$, i.e., mass conservation. Moreover, the L^2-projection satisfies the best approximation property $\|u - \mathcal{P}_h u\|_{L^2(\Omega)} = \min_{\varphi_h \in V_h} \|u - \varphi_h\|_{L^2(\Omega)}$, see [Loh19].

Using linear or bilinear basis functions, the gradient

$$\nabla u_h \overset{(3.4)}{=} \sum_j u_j \nabla \varphi_j \ \in (V_h)^d$$

is piecewise constant and discontinuous at the interelement boundaries, so that no direct evaluation at the nodes is possible. A continuous approximation for the gradient can be defined by [Kuz10]

$$\mathbf{g}_h \overset{(3.4)}{=} \sum_j \mathbf{g}_j \varphi_j \ \in (V_h)^d.$$

Our approach applies the L^2-projection

$$\int_\Omega \mathbf{g}_h\, \varphi_i = \int_\Omega \nabla u_h\, \varphi_i \qquad \forall\, \varphi_i \in V_h. \tag{6.18}$$

Choosing the φ_i from the same space as in the overall problem is not mandatory, but natural.
We introduce the discrete gradient operator $\mathbf{C}$, where

$$\mathbf{C} = (\mathbf{c}_{ij})_{i,j=1}^N, \quad \mathbf{c}_{ij} = \int_\Omega \varphi_i \nabla \varphi_j \, \mathrm{d}\mathbf{x}.$$

The boldface notation symbolizes that the single entries are not scalar-, but vector-valued. The size of the vectors equals the space dimension d. Since our application is set in the three-dimensional space, we assume that $d = 3$. A transfer to lower

dimensional spaces is no problem. In 3d, the gradient operator $\mathbf{C}$ can be split into the matrices $\mathbf{C} = (C^x, C^y, C^z)$, where

$$\begin{aligned} C^x &= (c^x_{i,j})_{i,j=1}^N, \quad c^x_{ij} = \int_\Omega \varphi_i \, \partial_x \varphi_j \, \mathrm{d}\mathbf{x}, \\ C^y &= (c^y_{i,j})_{i,j=1}^N, \quad c^y_{ij} = \int_\Omega \varphi_i \, \partial_y \varphi_j \, \mathrm{d}\mathbf{x}, \\ C^z &= (c^z_{i,j})_{i,j=1}^N, \quad c^z_{ij} = \int_\Omega \varphi_i \, \partial_z \varphi_j \, \mathrm{d}\mathbf{x}. \end{aligned}$$

Substituting the expressions for $\mathbf{g}_h$ and ∇u_h into (6.18), rearranging the terms and identifying the formulas for the consistent mass matrix and the discrete gradient operator, we find that

$$\begin{aligned} & \int_\Omega \sum_j \mathbf{g}_j \varphi_j \; \varphi_i \;=\; \int_\Omega \sum_j u_j \nabla \varphi_j \; \varphi_i \qquad \forall\, \varphi_i \in V_h \\ \iff \quad & \sum_j \underbrace{\int_\Omega \varphi_i \varphi_j}_{=m_{ij}} \; \mathbf{g}_j \;=\; \sum_j \underbrace{\int_\Omega \varphi_i \nabla \varphi_j}_{=\mathbf{c}_{ij}} \; u_j \qquad \forall\, \varphi_i \in V_h \\ \iff \quad & M_C \mathbf{g} \;=\; \mathbf{C} u. \end{aligned}$$

Here we approximate M_C by M_L because the inversion of the mass matrix is extremely cheap and for (multi-)linear basis functions the result is accurate enough. The equation to determine the coefficients of $\mathbf{g}_h$ then reads

$$\mathbf{g} = M_L^{-1} \mathbf{C} u. \tag{6.19}$$

For the implementation the question arises how the dimensions in (6.19) fit together, since the entries of both $\mathbf{g}$ and $\mathbf{C}$ are vector-valued. Obviously, the size of $\mathbf{g}$ is related to the size of u. If u is scalar, we obtain a gradient in the classical sense, that is,

$$\mathbf{g} = (\mathbf{g}_i)_{i=1}^N, \qquad \mathbf{g}_i = \begin{pmatrix} \partial_x u_i \\ \partial_y u_i \\ \partial_z u_i \end{pmatrix} =: \begin{pmatrix} g_{i,x} \\ g_{i,y} \\ g_{i,z} \end{pmatrix}.$$

However, u usually is a vector-valued quantity. Considering the FPE, u is a velocity field in each grid point. Strictly speaking, the 'gradients' are Jacobian matrices then. Therefore, we replace the notation $\mathbf{g}$ by $\mathbf{G}$, where

$$\mathbf{G} = (\mathbf{G}_i)_{i=1}^N, \qquad \mathbf{G}_i = \begin{pmatrix} \partial_x u_{i,x} & \partial_y u_{i,x} & \partial_z u_{i,x} \\ \partial_x u_{i,y} & \partial_y u_{i,y} & \partial_z u_{i,y} \\ \partial_x u_{i,z} & \partial_y u_{i,z} & \partial_z u_{i,z} \end{pmatrix} =: \begin{pmatrix} g_{i,xx} & g_{i,xy} & g_{i,xz} \\ g_{i,yx} & g_{i,yy} & g_{i,yz} \\ g_{i,zx} & g_{i,zy} & g_{i,zz} \end{pmatrix}.$$

Using the splitting of $\mathbf{C}$, in the case of a scalar function u the individual components of $\mathbf{g} = (g^x,\ g^y,\ g^z)$ can be written as

$$g^x = M_L^{-1} C^x u, \quad g^y = M_L^{-1} C^y u, \quad g^z = M_L^{-1} C^z u.$$

The elementwise formula then reads

$$g_{i,x} = \frac{1}{m_i} \sum_{j=1}^{N} c_{ij}^x u_j, \quad g_{i,y} = \frac{1}{m_i} \sum_{j=1}^{N} c_{ij}^y u_j, \quad g_{i,z} = \frac{1}{m_i} \sum_{j=1}^{N} c_{ij}^z u_j.$$

The extension to a vector-valued $\mathbf{u} = (u^x, u^y, u^z) \in \mathbb{R}^{N\times 3}$ is quite natural. We only have to substitute u^x, u^y or u^z for u to obtain the final result. Configurations for different sizes are illustrated in Table 6.2. The usage of row or column vector depends on the given structures in the software. Therefore, we do not care whether $\mathbf{G} \in \mathbb{R}^{N\times 3\times 3}$, $\mathbf{G} \in \mathbb{R}^{3\times 3\times N}$ or $\mathbf{G} \in \mathbb{R}^{3\times N\times 3}$.

u scalar	g^x	$=$	M_L^{-1}	C^x	u
considering only one component	$\in \mathbb{R}^N$		$\in \mathbb{R}^{N\times N}$	$\mathbb{R}^{N\times N}$	$\mathbb{R}^N$
u scalar	$\mathbf{g}$	$=$	M_L^{-1}	$\mathbf{C}$	u
considering all components	$\in \mathbb{R}^{3\times N}$		$\in \mathbb{R}^{N\times N}$	$\mathbb{R}^{N\times N\times 3}$	$\mathbb{R}^N$
u vector-valued	$\mathbf{G}$	$=$	M_L^{-1}	$\mathbf{C}$	$\mathbf{u}$
considering all components	$\in \mathbb{R}^{3\times 3\times N}$		$\in \mathbb{R}^{N\times N}$	$\mathbb{R}^{N\times N\times 3}$	$\mathbb{R}^{N\times 3}$

Table 6.2: *Overview of different setting with respect to the involved dimensions.*

Remark 6.4 (Testing the gradient recovery)**.** To check the implementation, we specify an exact velocity field, e.g., $\mathbf{u}(x, y, z) = (-y, x, z^2)$. We require the scheme to be linearity preserving, i.e., if the solution consists of linear functions such as the Jacobian

$$\nabla \mathbf{u} = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 2z \end{pmatrix},$$

it has to be exact.

6.3.1 Trace correction

We assume an incompressible fluid, i.e., $\nabla \cdot \mathbf{u} = 0$. For an analytical flow field $\mathbf{u}$ this is equivalent to $\operatorname{tr}(\nabla \mathbf{u}) = 0$. The gradient recovery approach presented above does not guarantee that this property transfers to the discretized function. Instead, for some points in space, the scalar $\operatorname{tr}(\nabla \mathbf{u})$ even takes values with two or three digits. Therefore, we solve the constrained optimization problem

$$\begin{cases} \min & \|G - \nabla \mathbf{u}\|_F^2 \\ \text{s.t.} & \operatorname{tr}(\nabla \mathbf{u}) = 0, \end{cases}$$

Let G be a Jacobian already approximated with the gradient recovery, while

$$\nabla \mathbf{u} := (z_{ij})_{i,j=1}^3$$

is unknown. We use the Frobenius norm, which is easier to apply than the L^2-norm. For a matrix $A \in \mathbb{R}^{m\times n}$ the squared Frobenius norm is defined by

$$\|A\|_F^2 := \sum_{i=1}^{m}\sum_{j=1}^{n} |a_{ij}|^2. \tag{6.20}$$

To find the minimum of $f(\nabla\mathbf{u}) := \|G - \nabla\mathbf{u}\|_F^2 := \sum_{i,j=1}^{3}(g_{ij} - z_{ij})^2$ subject to the equality constraint $g(\nabla\mathbf{u}) := \operatorname{tr}(\nabla\mathbf{u}) = z_{11} + z_{22} + z_{33} = 0$, we apply the Lagrange formalism, see [For13]. We first have to form the Lagrange function

$$\mathcal{L}(\nabla\mathbf{u}, \lambda) := f(\nabla\mathbf{u}) + \lambda g(\nabla\mathbf{u}),$$

where λ is the so-called Lagrange multiplier. The next step is to determine the stationary point of $\mathcal{L}(\nabla\mathbf{u}, \lambda)$. Differentiating with respect to z_{ij}, $i \neq j$, with respect to $z_{ii}, i \in \{1, 2, 3\}$, and with respect to λ, we find that

$$\frac{\partial\mathcal{L}}{\partial z_{ij}} = -2(g_{ij} - z_{ij}) \overset{!}{=} 0 \quad \Longrightarrow \quad z_{ij} = g_{ij}, \tag{6.21}$$

$$\frac{\partial\mathcal{L}}{z_{ii}} = -2(g_{ii} - z_{ii}) + \lambda \overset{!}{=} 0 \quad \Longrightarrow \quad z_{ii} = g_{ii} - \lambda/2, \tag{6.22}$$

$$\frac{\partial\mathcal{L}}{\partial\lambda} = z_{11} + z_{22} + z_{33} \overset{!}{=} 0. \tag{6.23}$$

While (6.21) gives no new results, substituting (6.22) into (6.23), we obtain

$$g_{11} + g_{22} + g_{33} - \frac{3}{2}\lambda = 0 \quad \Longleftrightarrow \quad \lambda = \frac{2}{3}\operatorname{tr}(G).$$

Inserting λ into (6.22), we end up with

$$z_{ii} = g_{ii} - \frac{1}{3}\operatorname{tr}(G), \tag{6.24}$$

that is, the deviatoric part of tensor G is determined. In summary, using this pleasantly simple formula (6.24), it holds true that

$$\operatorname{tr}(\nabla\mathbf{u}) \overset{(6.24)}{=} \sum_{i=1}^{3} g_{ii} - 3\cdot\frac{1}{3}\operatorname{tr}(G) = \operatorname{tr}(G) - \operatorname{tr}(G) = 0.$$

6.4 Numerical methods for Navier-Stokes equations

Modeling. The incompressible Navier-Stokes equations, consisting of the conservation laws for mass and momentum, were derived in Section 2.4. The stress tensor $\boldsymbol{\tau}$ determines whether a Newtonian or a non-Newtonian fluid is modeled. Leaving tensor $\boldsymbol{\tau}$ unspecified for the time being, the Navier-Stokes equations read [KH15, Loh19]

$$\rho\frac{\partial\mathbf{u}}{\partial t} + \rho\mathbf{u}\cdot\nabla\mathbf{u} - \nabla\cdot\boldsymbol{\tau} + \nabla p = \rho\mathbf{g} \quad \text{in } \Omega\times(0,T] \qquad \text{(momentum equation)},$$

$$
\begin{aligned}
\nabla \cdot \mathbf{u} &= 0 && \text{in } \Omega \times (0,T] && \text{(incompressibility)},\\
\mathbf{u}(\cdot,0) &= \mathbf{u}_0 && \text{in } \Omega && \text{(initial condition)},\\
\mathbf{u} &= \mathbf{u}_D && \text{on } \partial\Omega_D \times (0,T] && \text{(Dirichlet boundary)},\\
\boldsymbol{\tau}\mathbf{n} - p\mathbf{n} &= \rho\mathbf{h} && \text{on } \partial\Omega_N \times (0,T] && \text{(Neumann boundary)},
\end{aligned}
$$

where $\mathbf{u}_0$ is the initial velocity field, $\mathbf{u}_D$ the Dirichlet boundary data and $\mathbf{h}$ the force acting on the fluid in the outward normal direction $\mathbf{n}$ [Loh19].

Initial and boundary conditions. For the instationary Navier-Stokes equations an initial velocity $\mathbf{u}_0$ must be specified. It has to satisfy the incompressibility constraint. This is the case if it is determined as the solution of the stationary Navier-Stokes equations, or the corresponding Stokes equations. Note that $\mathbf{u}_0$ also has to be compatible with the boundary conditions.

On Dirichlet boundaries, the velocity is given as a function. Typically, such boundary conditions, also known as essential boundary conditions, are set at the inflow boundary and at the fixed walls. In the case of pure Dirichlet boundaries we obtain

$$
0 = \int_\Omega \nabla \cdot \mathbf{u} \, \mathrm{d}\mathbf{x} = \int_{\partial\Omega} \mathbf{u} \cdot \mathbf{n} \, \mathrm{d}\mathbf{s} = \int_{\partial\Omega} \mathbf{u}_D \cdot \mathbf{n} \, \mathrm{d}\mathbf{s}
$$

as a compatibility condition [Joh16]. When using pure Dirichlet boundary conditions the pressure solution is only defined up to an additive constant. An option to get a unique solution is to force the mean value over the pressure to be zero [Bra13]. Furthermore, the pressure becomes uniquely defined, when Dirichlet and Neumann conditions are combined. Neumann conditions are typically chosen at the outflow boundary. The exact form of a Neumann boundary condition depends on the given PDE. Neumann conditions are also called natural boundary conditions [Joh16], since they are defined in a way that the integrands of the corresponding boundary integrals reduce to a given function. If they even reduce to zero, we have a so-called do-nothing boundary condition [Joh16].

Numerical methods. We apply numerical methods to approximate a solution of the Navier-Stokes equations, since in the vast majority of cases no analytical solution is known. A space and a time discretization as well as a solver have to be chosen.

The usual methods are applicable for the spatial discretization. However, finite differences are usually only found in ancient publications [Cho68]. In the 1990s finite volumes and finite elements had proven to provide the most accurate and efficient solutions [Tur99]. The trend was more and more towards finite element methods [GS98, Tur99], which we also use below.

6.4.1 Weak formulation, discretization, Stokes elements

Weak formulation. To derive a weak formulation for the continuous conservation laws, as usual the equations are multiplied by a test function and the resulting expression is integrated. However, the test function $\mathbf{w}$ has to be vector-valued, since

the velocity $\mathbf{u}$ is [Loh19]. For the viscous term $\nabla \cdot \boldsymbol{\tau}$ and for the pressure ∇p it applies that

$$\begin{aligned} -\int_\Omega (\nabla \cdot \boldsymbol{\tau}) \cdot \mathbf{w} \, \mathrm{d}\mathbf{x} &= \int_\Omega \boldsymbol{\tau} : \nabla \mathbf{w} \, \mathrm{d}\mathbf{x} - \int_{\partial\Omega} \boldsymbol{\tau}\mathbf{n} \cdot \mathbf{w} \, \mathrm{d}\mathbf{s}, \\ \text{and} \qquad \int_\Omega \nabla p \cdot \mathbf{w} \, \mathrm{d}\mathbf{x} &= -\int_\Omega p(\nabla \cdot \mathbf{w}) \, \mathrm{d}\mathbf{x} + \int_{\partial\Omega} p\mathbf{n} \cdot \mathbf{w} \, \mathrm{d}\mathbf{s}. \end{aligned}$$

Hence, the weak formulation for the whole system reads:

Find $\mathbf{u} \in \mathbf{W}$ and $p \in Q$, so that

$$\begin{aligned} &\rho \int_\Omega \frac{\partial \mathbf{u}}{\partial t} \cdot \mathbf{w} \, \mathrm{d}\mathbf{x} + \rho \int_\Omega (\mathbf{u} \cdot \nabla \mathbf{u}) \cdot \mathbf{w} \, \mathrm{d}\mathbf{x} + \int_\Omega \boldsymbol{\tau} : \nabla \mathbf{w} \, \mathrm{d}\mathbf{x} \\ &\quad - \int_\Omega p(\nabla \cdot \mathbf{w}) \, \mathrm{d}\mathbf{x} - \int_{\partial\Omega} (\boldsymbol{\tau}\mathbf{n} - p\mathbf{n}) \, \mathbf{w} \, \mathrm{d}\mathbf{s} = \rho \int_\Omega \mathbf{g} \cdot \mathbf{w} \, \mathrm{d}\mathbf{x} \quad \forall \mathbf{w} \in \mathbf{W}_0, \end{aligned} \tag{6.25a}$$

$$\text{and} \quad \int_\Omega (\nabla \cdot \mathbf{u}) q = 0 \quad \forall q \in Q. \tag{6.25b}$$

With respect to the highest occurring derivative for the velocity, we choose

$$\begin{aligned} \mathbf{W} &= \{\mathbf{u} \in H^1(\Omega)^d \mid \mathbf{u} = \mathbf{u}_D \text{ on } \partial\Omega_D\}, \\ \text{and} \qquad \mathbf{W}_0 &= \{\mathbf{w} \in H^1(\Omega)^d \mid \mathbf{w} = \mathbf{0} \text{ on } \partial\Omega_D\} \end{aligned}$$

as solution and test space for the velocity. For the pressure $Q = L^2(\Omega)$ is appropriate both as solution and test space, since no derivatives of the pressure appear.
In what follows, the boundary integrals in (6.25a) simplify due to $\mathbf{w} \in \mathbf{W}_0$ and the Neumann condition. Additionally, a minus is inserted into the homogeneous equation (6.25b), so that the expressions for the pressure gradient ∇p and the incompressibility constraint $\nabla \cdot \mathbf{u}$ mirror each other. We obtain:

Find $\mathbf{u} \in \mathbf{W}$ and $p \in Q$, so that

$$\begin{aligned} &\rho \int_\Omega \frac{\partial \mathbf{u}}{\partial t} \cdot \mathbf{w} \, \mathrm{d}\mathbf{x} + \rho \int_\Omega (\mathbf{u} \cdot \nabla \mathbf{u}) \cdot \mathbf{w} \, \mathrm{d}\mathbf{x} + \int_\Omega \boldsymbol{\tau} : \nabla \mathbf{w} \, \mathrm{d}\mathbf{x} \\ &\quad - \int_\Omega p(\nabla \cdot \mathbf{w}) \, \mathrm{d}\mathbf{x} - \rho \int_{\partial\Omega_N} \mathbf{h} \, \mathrm{d}\mathbf{s} = \rho \int_\Omega \mathbf{g} \cdot \mathbf{w} \, \mathrm{d}\mathbf{x} \quad \forall \mathbf{w} \in \mathbf{W}_0, \end{aligned} \tag{6.26a}$$

$$\text{and} \quad -\int_\Omega (\nabla \cdot \mathbf{u}) q = 0 \quad \forall q \in Q. \tag{6.26b}$$

Discretization. Next, we spatially discretize the weak formulation. We use a conforming approach, that is, for the finite-dimensional spaces we have that $\mathbf{W}_h \subseteq \mathbf{W}$ and $Q_h \subseteq Q$. The task reads:

Find $\mathbf{u}_h \in \mathbf{W}_h$ and $p_h \in Q_h$, so that

$$\begin{aligned} &\rho \int_\Omega \frac{\partial \mathbf{u}_h}{\partial t} \cdot \mathbf{w}_h \, \mathrm{d}\mathbf{x} + \rho \int_\Omega (\mathbf{u}_h \cdot \nabla \mathbf{u}_h) \cdot \mathbf{w}_h \, \mathrm{d}\mathbf{x} + \int_\Omega \boldsymbol{\tau}(\mathbf{u}_h) : \nabla \mathbf{w}_h \, \mathrm{d}\mathbf{x} \\ &\quad - \int_\Omega p_h(\nabla \cdot \mathbf{w}_h) \, \mathrm{d}\mathbf{x} - \rho \int_{\partial\Omega_N} \mathbf{h}_h \, \mathrm{d}\mathbf{s} = \rho \int_\Omega \mathbf{g}_h \cdot \mathbf{w}_h \, \mathrm{d}\mathbf{x} \quad \forall \mathbf{w}_h \in \mathbf{W}_0^h, \end{aligned} \tag{6.27a}$$

$$\text{and} \quad -\int_\Omega (\nabla \cdot \mathbf{u}_h)\, q_h \;=\; 0 \quad \forall q_h \in Q_h. \tag{6.27b}$$

The discrete velocity and the discrete pressure are interpolated by [ESW14]

$$\mathbf{u}_h = \sum_{j=1}^{n_u} \mathbf{u}_j(t)\boldsymbol{\varphi}_j(\mathbf{x}) \quad \text{and} \quad p_h = \sum_{l=1}^{n_p} p_l(t)\psi_l(\mathbf{x}),$$

where $\boldsymbol{\varphi}_j$ and ψ_l are the basis functions. While the vector-valued basis functions for the velocity $\mathbf{u}_h$ is helpful for analytical purposes, scalar functions are preferred for practical implementation. The numbers n_u and n_p are the degrees of freedom for which the coefficients $\mathbf{u}_j(t)$ and $p_l(t)$ are to be determined and have not already been specified by Dirichlet boundary values. Substituting the given approaches for $\mathbf{u}_h$ and p_h into the weak formulation (6.27), and setting $\rho = 1$, we obtain the semi-discrete formulation

$$\begin{aligned} M\frac{\partial \mathbf{u}}{\partial t} + K(\mathbf{u})\mathbf{u} + A\mathbf{u} + Bp &= \mathbf{g}, \\ B^T\mathbf{u} &= 0. \end{aligned} \tag{6.28}$$

Here $\mathbf{u}$ and p refer to discrete variables, even if we omit the index h. The vector $\mathbf{g}$ involves both contributions from the right hand side and possible contributions from boundary nodes. The finite element matrices we obtain are the mass matrix

$$M = (m_{ij})_{\substack{i=1,\dots,n_u \\ j=1,\dots,n_u}}, \quad \text{where} \quad m_{ij} := \int_\Omega \boldsymbol{\varphi}_i(\mathbf{x}) \cdot \boldsymbol{\varphi}_j(\mathbf{x})\, \mathrm{d}\mathbf{x},$$

the nonlinear convection matrix

$$K(\mathbf{u}) = (k_{ij})_{\substack{i=1,\dots,n_u \\ j=1,\dots,n_u}}, \quad \text{where} \quad k_{ij} := \int_\Omega (\mathbf{u}_h \cdot \nabla\boldsymbol{\varphi}_j(\mathbf{x})) \cdot \boldsymbol{\varphi}_i(\mathbf{x})\, \mathrm{d}\mathbf{x},$$

the diffusion matrix

$$A = (a_{ij})_{\substack{i=1,\dots,n_u \\ j=1,\dots,n_u}}, \quad \text{where} \quad a_{ij} := \int_\Omega \nabla\boldsymbol{\varphi}_i : 2\mu_I(\mathbf{D}(\boldsymbol{\varphi}_j) + N_p\mathbb{A} : \mathbf{D}(\boldsymbol{\varphi}_j))\, \mathrm{d}\mathbf{x},$$

and the matrix

$$B = (b_{ik})_{\substack{i=1,\dots,n_u \\ k=1,\dots,n_p}}, \quad \text{where } b_{ik} := -\int_\Omega \psi_k(\mathbf{x})\nabla \cdot \boldsymbol{\varphi}_i(\mathbf{x})\, \mathrm{d}\mathbf{x}.$$

The matrices B and B^T are the discrete analogs of the gradient and the divergence operators.
If the standard θ-scheme is applied to (6.28) as time-discretization, we obtain

$$\underbrace{\left[\frac{1}{\Delta t}M + \theta\big(K(\mathbf{u}^{n+1}) + A\big)\right]}_{:= S}\mathbf{u}^{n+1} + Bp^{n+1} = \underbrace{\mathbf{g}^n + \left[\frac{1}{\Delta t}M - (1-\theta)\big(K(\mathbf{u}^n) + A\big)\right]\mathbf{u}^n}_{:= \mathbf{f}},$$

$$B^T\mathbf{u}^{n+1} = \mathbf{0}.$$

Handling the pressure implicitly as $p^{n+\mathbf{1}}$ ensures that the incompressibility constraint $B^T\mathbf{u}^{n+\mathbf{1}} = 0$ can be satisfied. The three matrices M, K and A acting on the velocity vector $\mathbf{u}^{n+1}$ are combined to form the so-called velocity matrix S. Then, finally, the fully discrete system can be written as the saddle point problem

$$\begin{bmatrix} S & B \\ B^T & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ p \end{bmatrix} = \begin{bmatrix} \mathbf{f} \\ 0 \end{bmatrix}, \tag{6.29}$$

where $S \in \mathbb{R}^{n_u \times n_u}$, $B \in \mathbb{R}^{n_u \times n_p}$, and $[\mathbf{u}, p] =: [\mathbf{u}^{n+1}, p^{n+1}]$. Different options to solve this nonlinear saddle point problem are discussed in Section 6.4.2. Before that, we think about the solvability of the system (6.29) and about stable finite element pairs for the velocity and the pressure.

Dimensional consideration. A necessary condition for the saddle point problem to be solvable, is that the number of degrees of freedom for the velocity is greater than or equal to that for the pressure, i.e., $n_u \geq n_p$. Otherwise, the columns of B are linearly dependent, and the saddle point matrix has no full rank. Moreover, even strict inequality should apply, i.e., $n_u > n_p$ [KH15]. Otherwise, velocity $\mathbf{u}$ would already be completely determined by the incompressibility constraint $B^T\mathbf{u} = 0$. Consequently, it is a good rule of thumb to choose the polynomials for the approximation of the velocity one degree higher than those for the pressure. However, pure dimensional considerations are not at all sufficient to guarantee stability [Bra13]. We need a further condition.

Inf-sup condition. Many analytical investigations are restricted to the linear Stokes problem given in Remark 2.11. Fortunately, some of the results transfer to the Navier-Stokes equations, see, e.g., [Ess22]. One such tool is the inf-sup condition, also known as LBB condition. On the discrete level, it requires that there is a positive mesh-independent constant γ, such that [ESW14, Loh19]

$$\inf_{\substack{q_h \in Q_h \\ q_h \neq 0}} \sup_{\substack{\mathbf{v}_h \in \mathbf{W}_h \\ \mathbf{v}_h \neq \mathbf{0}}} \frac{|(q_h, \nabla \cdot \mathbf{v}_h)|}{\|q_h\|_{L^2(\Omega)} \|\mathbf{v}_h\|_{H^1(\Omega)}} \geq \gamma > 0.$$

The condition is sufficient for the solvability of the saddle point problem to the Stokes equations, and it ensures the stability of the pressure [ESW14].

Stokes elements. We present finite element pairs for the velocity and the pressure. Analytical investigations are only available for the Stokes and not for the nonlinear Navier-Stokes equations. However, a finite element pair, which is unstable for the Stokes equations is also unstable for the related Navier-Stokes equations. Conversely, numerical experiments have shown that the stability transfers from the Stokes to the Navier-Stokes equations.

As already the inf-sup condition, which combines $\mathbf{W}_h$ and Q_h in one expression, indicates, the discrete spaces cannot be chosen independently of each other [ESW14]. Since $\mathbf{u} \in \mathbf{W} = H_0^1(\Omega)^d$ and $p \in Q = L^2(\Omega)$ live in different spaces, it is reasonable to choose different discrete spaces $\mathbf{W}_h$ and Q_h, that is, to use a mixed approximation.

The usage of discrete spaces of equal order might also work, but circumventing the basic inf-sup condition requires stabilization terms in the discrete continuity equation [BN83, Ste84, ESW14, KH15, Joh16].
The incompressibility constraint $\nabla \cdot \mathbf{u} = 0$ provides the additional equation, which is necessary to determine the pressure. Conversely, an appropriate approximation for the pressure is essential to ensure discrete freedom of divergence for the velocity.

In what follows, we take a closer look at the capability of different combinations of quadrilateral elements. We refer to the literature for similar considerations about triangular elements, e.g., [Bra13, ESW14, Joh16]. The simplest and seemingly natural choices for $(\mathbf{W}_h, Q_h)$ are inappropriate. For instance, the pairs $\mathbf{Q}_1\mathbf{P}_0$ and $\mathbf{Q}_1\mathbf{Q}_1$, visualized in Figure 6.3a) and b), are unstable. Since, in particular, $\mathbf{Q}_1\mathbf{P}_0$ is 'slightly unstable but highly usable' [GS98], stabilization is an option. Alternatively, the dimension of the approximation for the velocity has to be enhanced in relation to the pressure. Some options for this are $\mathbf{Q}_2\mathbf{Q}_1$, $\mathbf{Q}_2\mathbf{P}_1^{\text{disc}}$ and $\mathbf{Q}_2\mathbf{P}_0$, see Figure 6.3c)-e).

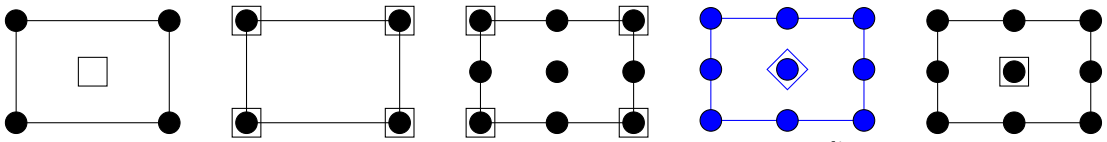


Figure 6.3: *Different element pairs, shown here for 2d. Without further stabilization the pairs a) $\mathbf{Q}_1\mathbf{P}_0$ and b) $\mathbf{Q}_1\mathbf{Q}_1$ are unstable, while c) $\mathbf{Q}_2\mathbf{Q}_1$, d) $\mathbf{Q}_2\mathbf{P}_1^{\text{disc}}$ and e) $\mathbf{Q}_2\mathbf{P}_0$ are stable from the start. The symbol ● denotes the (two) velocity components, □ the pressure, and ◇ the combination of pressure and pressure gradient.*

The so-called Taylor-Hood pair $\mathbf{Q}_2\mathbf{Q}_1$ is quite common. It is the simplest second-order quadrilateral element pair, which has a C^0-pressure [GS98]. The pair $\mathbf{Q}_2\mathbf{P}_0$ is a simplification of $\mathbf{Q}_2\mathbf{P}_1^{\text{disc}}$. However, this is false economy, because the low accuracy of the pressure approximation also influences the velocity approximation [GS98, ESW14].

For this work, the pair $\mathbf{Q}_2\mathbf{P}_1^{\text{disc}}$ was used [Ess22]. The choice of the discontinuous pressure approximation $\mathbf{P}_1^{\text{disc}}$ is permissible, since we only require that $Q_h \subseteq Q := L^2(\Omega)$. The pair $\mathbf{Q}_2\mathbf{P}_1^{\text{disc}}$ is more efficient than $\mathbf{Q}_2\mathbf{Q}_1$, since we have one degree of freedom less for the pressure. In [ESW14, Sec. 3.3.1] it is established that $\mathbf{Q}_2\mathbf{P}_1^{\text{disc}}$ is stable. In general, a discontinuous pressure approximation has some attractive features. Practical experience has shown that the accuracy of the velocity solution increases in the case of a discontinuous pressure approximation [ESW14]. Furthermore, we even obtain local mass conservation, since for every element K the test function q_h can be set as characteristic function, that is, $q_h(\mathbf{x}) = 1$ if $\mathbf{x} \in K$ and $q_h(\mathbf{x}) = 0$ otherwise. Consequently,

$$0 = \int_\Omega q_h \nabla \cdot \mathbf{u}_h \, \mathrm{d}\mathbf{x} = \int_K \nabla \cdot \mathbf{u}_h \, \mathrm{d}\mathbf{x} = \int_{\partial K} \mathbf{u}_h \cdot \mathbf{n} \, \mathrm{d}s.$$

Last but not least, a discontinuous space for the pressure is particularly suitable for multiphase flows, where the pressure has a discontinuous jump at the boundary anyway.

Finally, there are also non-conforming approaches, where $\mathbf{W}_h \not\subseteq \mathbf{W}$ and $Q_h \not\subseteq Q$. An example for such an approach is the Rannacher-Turek element $\widetilde{\mathbf{Q}}_1\mathbf{Q}_0$ consisting of rotated bilinear shape functions for the velocity and a piecewise constant pressure approximation [Tur99]. The pair is cheap, inf-sup stable [KH15] and well suited for the design of multigrid solvers [Tur99]. However, the analysis for non-conforming approaches is associated with increased difficulties.

When switching from 2d to 3d, the transition from rectangles to hexahedra is straightforward. Changing from triangles to tetraeders is non-trivial [ESW14, Sec. 3.3.4].

6.4.2 Solution process

Literature. There is a zoo of methods to solve the Navier-Stokes equations, see, e.g., [Qua93, ESW14, Ran17c]. We applied multigrid to the (linearized) NSE and used the Vanka smoother [Van86]. Before giving a rough overview of the components of this solver, we briefly introduce the overall framework of PSC solvers.

Pressure Schur complement (PSC). The PSC is a basic approach to solve saddle point problems. To derive the PSC approach for our saddle point problem (6.29), we first multiply the momentum equation $S\mathbf{u} + Bp = \mathbf{f}$ from left by $B^T S^{-1}$ and make use of the mass equation $B^T\mathbf{u} = 0$, so that

$$B^T S^{-1} Bp = B^T S^{-1}\mathbf{f}. \tag{6.30a}$$

The matrix $B^T S^{-1} B$ is referred to as pressure Schur complement matrix. Moreover, we rewrite the momentum equation as

$$\mathbf{u} = S^{-1}(\mathbf{f} - Bp). \tag{6.30b}$$

Summing up, we have a mechanism to update pressure and velocity one after another, first p via (6.30a) and then $\mathbf{u}$ via (6.30b).

Classical projection schemes. A well-known option to solve the NSE are the classical continuous projection schemes. Following [Tur99], these schemes can be interpreted as global PSC schemes. This means that preconditioners $C^{-1} \approx (B^T S^{-1} B)^{-1}$ refer to the global PSC matrix.
A first-order representative is the Chorin-Temam scheme [Cho68]. It consists of two steps. In the first step the intermediate velocity field $\mathbf{v}^{n+1}$ is calculated, while the second step gives the new result $(\mathbf{u}^{n+1}, p^{n+1})$. The second step contains the eponymous projection, which is based on the Helmholtz decomposition theorem. This states that every 3d velocity field can be decomposed in a divergence-free and in a curl-free part [Qua93, Sec. 7.2], [Joh16, Th. 3.168].
Shortcomings of the Chorin-Temam scheme, besides its order, are an artificial dependence of a stationary solution on Δt [KH15] and that the previous value of the

pressure is never taken into account. An alternative is, for example, the second-order van Kan projection scheme [Kan86].

These classical projection schemes are in particular preferred for high Reynolds numbers, when small time steps are used anyway. For a small Reynolds number, as it is given in our application, we prefer a monolithic local solver. Compared to the projection schemes, each step is more expensive, but fewer steps are required. The time step size only has to be adjusted for accuracy but not for robustness reasons [Ess22, p. 26], [Tur99, e.g. p. 20].

Local Multilevel Pressure Schur complement solver. Therefore, instead of a classical projection scheme, in this thesis we use a local multilevel pressure Schur complement (MPSC) solver [Tur99, Ess22]. The scheme is called local because it applies preconditioners C_i locally on certain patches Ω_i [Tur99, p. 44; p. 67]. In what follows, we roughly consider the different building block, that is, the Newton scheme as outer solver and the multigrid method as inner solver.

Newton. As nonlinear outer solver Newton's method is used. It is well-known that for a nonsingular Jacobian and a suitable initial value $\mathbf{u}_0$, Newton's method converges quadratically. A drawback of Newton's method is that for larger Reynolds numbers the radius of convergence shrinks, so that better initial guesses $\mathbf{u}_0$ are needed [Ess22].
The first initial value can be calculated using either the Stokes or the steady Navier-Stokes equations. Considering the unsteady NSE, we can later use the result from the previous time step as new initial value. All in all, in our application with its small Reynolds number no problems related to the initial values were observed [Ess22].
As stopping criterion the residual of the nonlinear equation is determined. Specifically, the tolerance was set to 1e-11 [Ess22]. Even if this tends to be overkill, it does not harm since we are still finished after 1 to 3 steps due to the quadratic convergence.

The Newton method is also used for linearization purposes. Specifically, we obtain Oseen problems, which correspond to the Stokes equations extended by a convective term, see, e.g., [Joh16]. We solve these problems using a multigrid algorithm.

Multigrid methods. Multigrid methods [TOS00, Hac13] are a class of advanced solvers. A key advantage is that their order of convergence is independent of the mesh size. However, there is not one multigrid algorithm for all applications.

Classical iteration schemes converge the better, the smaller the spectral radius of their iteration matrix is. As Fourier analysis shows, an iteration error has both smooth and oscillatory parts. The oscillating error components, on the one hand, correspond to the small eigenvalues. They can be damped very quickly [Qua93,

Sec. 2.7]. The smooth components, on the other hand, are associated with the large eigenvalues and hold back convergence.

The idea of multigrid is that an error that is smooth on a fine mesh becomes generally more oscillatory on a coarser mesh. Hence, both fine mesh and coarser mesh problems are solved. In the simplest case, this can be done very efficiently with the V-cycle, see Figure 6.4.
Transferring information from the coarse to the fine mesh is called interpolation or prolongation, while the transfer from the fine to a coarse grid is called restriction. For technical details we refer, e.g., to [Tur99, BHM00, TOS00].

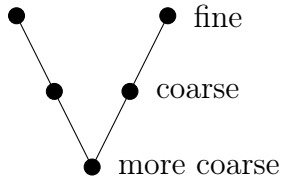


Figure 6.4: *3-grid V-cycle [ESW14].*

The multigrid scheme used to solve the NSE in this thesis is preconditioned with a so-called Vanka smoother. Originally introduced in [Van86], the Vanka smoother is a block Jacobi (as in our case) or a block Gauß-Seidel iteration scheme, see [DR06]. Our Vanka smoother works very locally in the sense that the patches Ω_i are determined by the triangulation of the mesh resulting in tiny non-overlapping patches. Following [Ess22], the FEAT3 software used has an efficient implementation for this case.

Modified diffusion term. Finally, we think about the practical consequences that in the NSE the conventional diffusive term $\Delta\mathbf{u}$ is replaced by the term $\nabla \cdot \boldsymbol{\tau}$ for non-Newtonian fluids. In principle, this can lead to a less well-conditioned system, i.e., more multigrid steps. However, the system remains stable for the parameters N_p and N_s under consideration [Ess22]. Therefore, the main effort compared to the NSE for Newtonian fluids is the more complex assembly of the corresponding FE matrix.

7 Numerical studies for the coupled system

In this chapter, we finally address the fully coupled FPE-NSE system. One challenge is to find meaningful benchmarks. A first benchmark for the Fokker-Planck equation, including all of its terms, is presented in Section 7.1. This benchmark still contains some simplifications, but provides an analytical solution. Operator splitting is the only technique from Chapter 6, which is needed. This changes in Section 7.2, where we move to the second benchmark problem, a three-dimensional axisymmetric contraction. There we obtain the velocity field from the Navier-Stokes equations. To verify our results, at least qualitatively, we compare them with results from the literature.

7.1 Analytical Jeffery Benchmark

7.1.1 Setting

We consider numerical results for the full Fokker-Planck equation. For $\psi = \psi(\mathbf{x}, \mathbf{p}, t)$ the equation reads

$$\begin{aligned} \frac{\partial \psi}{\partial t} + \nabla_{\mathbf{x}} \cdot (\mathbf{u}\psi) + \nabla_{\mathbf{p}} \cdot (\dot{\mathbf{p}}\psi) &= \Delta_{\mathbf{p}}(D_r \psi) \qquad \text{in } \Omega \times \mathbb{S}^2 \times (0, T], \\ \text{where} \quad \dot{\mathbf{p}} &= \mathbf{W}\mathbf{p} + \lambda_e \left[\mathbf{D}\mathbf{p} - (\mathbf{D} : (\mathbf{p} \otimes \mathbf{p}))\, \mathbf{p}\right], \\ \psi(\cdot, \cdot, 0) &= \psi_0, \quad \text{and} \quad \psi(\mathbf{x}, \cdot, \cdot) = \psi_{\text{bc}} \quad \text{for } \mathbf{x} \in \partial\Omega. \end{aligned}$$

We choose the geometry Ω and the velocity field $\mathbf{u}$ in such a way that we have an analytical reference solution. Specifically, we make use of the temporal equivalence between the solution ψ_L of the space-independent FPE

$$\frac{\partial \psi_L}{\partial t} + \nabla_{\mathbf{p}} \cdot (\dot{\mathbf{p}}\psi_L) = \Delta_{\mathbf{p}}(D_r \psi_L)$$

and the solution ψ_E of the space-dependent FPE

$$\frac{\partial \psi_E}{\partial t} + \nabla_{\mathbf{p}} \cdot (\dot{\mathbf{p}}\psi_E) + \nabla_{\mathbf{x}} \cdot (\mathbf{u}\psi_E) = \Delta_{\mathbf{p}}(D_r \psi_E).$$

While the equation for ψ_L describes the Lagrangian perspective, the full problem including the spatial convection term takes the Eulerian perspective, see Section 2.2.2. We also find the idea of a solid body rotation here. If we start from the same initial condition, that is, $\psi_E(t_0) = \psi_L(t_0)$ and use a rotational velocity field in space with

the angular velocity $\omega = 1$, after one complete revolution there is an equivalence again, that is,

$$\psi_E(t_0 + 2\pi) = \psi_L(t_0 + 2\pi).$$

Hence, the solution of the space-independent problem can be used as reference solution for the full Fokker-Planck equation. The fact that the PDE for ψ_L is the extended Jeffery equation gives rise to the name 'Analytical Jeffery Benchmark'.

General setting. As computational area in space we choose the unit circle, i.e.,

$$\Omega = \big\{(x, y) \in \mathbb{R}^2 \,|\, x^2 + y^2 \leq 1\big\}.$$

Accordingly, let the spatial velocity field describe a counter-clockwise rotation around the center of the circle, i.e.,

$$\mathbf{u} = (-y, x)^T.$$

Considering $\Omega \subset \mathbb{R}^2$ instead of $\Omega \subset \mathbb{R}^3$ is a simplification. However, a two-dimensional domain is suitable for the time being, not least because it keeps the computational time within limits. Another advantage of the circle as geometry is that the boundary conditions cannot only be neglected on $\mathbb{S}^2$, but also for Ω, because no mass crosses the boundary, since $\mathbf{u}(\mathbf{s}) \cdot \mathbf{n}(\mathbf{s}) = 0 \quad \forall\, \mathbf{s} \in \partial\Omega$.

As Jacobian matrix necessary to calculate the deformation and spin tensor, which are needed in turn to calculate the rotation velocity $\dot{\mathbf{p}}$, we employ

$$\nabla\mathbf{v} = \operatorname{diag}\big(0.02,\, -0.01,\, -0.01\big).$$

This corresponds to an uniaxial elongation. That $\nabla\mathbf{v} \in \mathbb{R}^{3\times3}$ is chosen independently of $\mathbf{u} \in \mathbb{R}^2$ is another simplification. It ignores the connection between $\mathbf{u}$ and $\nabla\mathbf{v}$, but for our purposes this modified equation works fine. In particular, this way the sphere $\mathbb{S}^2$ does not need to be replaced by the circle $\mathbb{S}^1$ as computational domain for the space-independent FPE.

Initial conditions. The next challenge is to find a reasonable initial condition both with respect to space and orientation. One constraint is that the normalization condition is satisfied for each sphere. We use the analytical solution of the extended Jeffery equation given in Corollary 2.5. This way, not only can the solution ψ_L be used as reference solution for ψ_E, but in the case $D_r = 0$ we even know the analytical solution.
For practical purposes we switch to the discrete level, i.e., to the coefficient matrix $\Psi \in \mathbb{R}^{N\times M}$ introduced by (6.13). For each spatial point $i \in \{1, \ldots, N\}$ an initial condition for the corresponding spherical grid point $k \in \{1, \ldots, M\}$ is given by

$$\Psi_{i,k} := \psi(\mathbf{p}_k, t(i)) = \frac{\psi_0}{\|\mathbf{C}(t(i))\mathbf{p}_k\|^3}, \quad \text{where } \mathbf{C}(t(i)) = \exp(-t(i)(\mathbf{W} + \lambda_e\mathbf{D})). \tag{7.1}$$

The exact solution ψ does not only depend on the orientation $\mathbf{p}$, but also on the time t. This $t(i)$ is used as parameter to specify different smooth configurations on the spheres. Furthermore, it also defines the configuration in space as already the notation $t(i)$ indicates. With a coupling between $t(i)$ and the spatial points $\mathbf{x}_i$ a meaningful initial configuration in space can be ensured as well. We define two different versions of our benchmark.

a) Smooth initial configuration. Within this test case we simply set $t(i)$ to the x-coordinate of $\mathbf{x}_i$, that is,

$$t(i) := x_i.$$

This approach results in a smooth configuration both in space and orientation.

b) Discontinuous initial configuration. The second initial condition is more complex. It is based on the slotted cylinder/sharp cone/smooth hump setting known from Section 4.4. The parameter $t(i)$ more precisely reads $t(r(\mathbf{x}_i))$ here and $r := r(\mathbf{x}_i) := \|\mathbf{x}_i - \mathbf{x}_g^m\|$ defines the distance from the respective point $\mathbf{x}_i$ to the midpoints $\mathbf{x}_g^m$ of the three geometries. The midpoints are set to

$$\mathbf{x}_g^m := \begin{cases} (0.0;\ 0.5) & \text{if } g = SlCy, \\ (0.0;\ -0.5) & \text{if } g = cone, \\ (-0.5;\ 0.0) & \text{if } g = hump. \end{cases}$$

The radius $R := 0.3$ and the claim $r(\mathbf{x}_i) \leq R$ determine the circles $\mathcal{B}_g$ for the particular geometries. For the cylinder to be slotted at least one of the two requirements $|x| \geq 0.05$ and $y \geq 0.6$ has to be satisfied. Finally, we define

$$t(r(\mathbf{x}_i)) := \begin{cases} t_1 & \text{if } \mathbf{x}_i \in \mathcal{B}_{SlCy}, \\ (1 - \omega(r))\, t_2, \text{ where } \omega(r) = r/R & \text{if } \mathbf{x}_i \in \mathcal{B}_{cone}, \\ (1 - \omega(r))\, t_3, \text{ where } \omega(r) = 1/2\,(1 - \cos(\pi\, r/R)) & \text{if } \mathbf{x}_i \in \mathcal{B}_{hump}. \end{cases}$$

The expressions $\omega(r)$ define the concrete shape of cone and hump. They are constructed in a way that $\omega(0) = 1$ and $\omega(R) = 1$, so that $t(0) = t_{2,3}$ and $t(R) = 0$. In particular, $t = 0$ on the boundary of the circles $\mathcal{B}_{cone}$ and $\mathcal{B}_{hump}$ indicates an isotropic distribution there. Consequently, setting the values outside the geometries to $1/4\pi$ as well, guarantees a smooth transition between the areas of cone and hump and their surrounding.

In Figure 7.1, we see two discontinuous configurations as examples. They obviously cover quite different value ranges. Moreover, the two configurations cannot be depicted completely analog, since in the first case, with respect to the values, the geometries actually point 'downwards'.

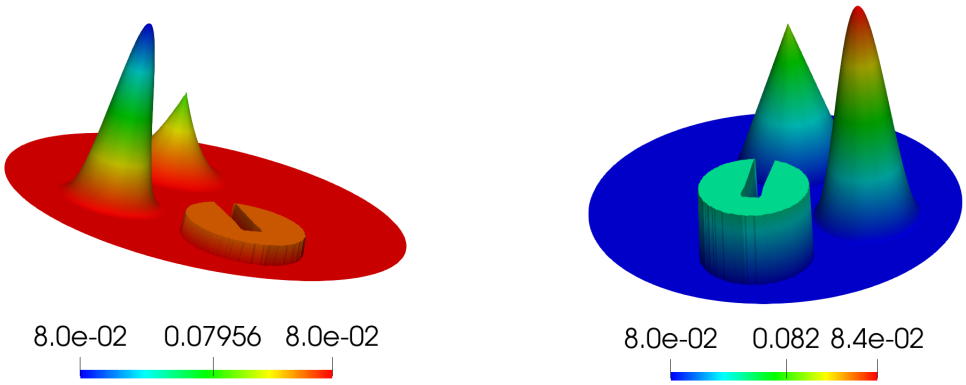


Figure 7.1: *Visualization of the discontinuous initial configuration in space for two different spherical mesh points; $k = 6$ on the left side and $k = 24$ on the right side.*

Remark 7.1 (Choice of $\nabla\mathbf{v}$)**.** We choose $\nabla\mathbf{v}$ independently of $\mathbf{u}$. The Jacobian matrix corresponding to the extended velocity field $\tilde{\mathbf{u}} = (-y, x, 0)^T$ would read

$$\nabla\tilde{\mathbf{u}} = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

However, compared to $\nabla\mathbf{v} = \operatorname{diag}\big(0.02, -0.01, -0.01\big)$, the choice of $\nabla\tilde{\mathbf{u}}$ would significantly shift the magnitudes of the convective terms relative to each other. In particular, multiplicating the values in the matrix $\nabla\mathbf{v}$ is equivalent to multiplicating the final time T: strong discontinuities would arise, see Section 5.6.3.1, and lead to numerical problems.

Moreover, the basic initial condition (7.1) would reduce to a constant. Since the above $\nabla\tilde{\mathbf{u}}$ is skew-symmetric, it applies that $\mathbf{D} = \mathbf{0}$, $\mathbf{W} = \nabla\tilde{\mathbf{u}}$ and consequently $\mathbf{C}(t) = \exp(-t\mathbf{W})$. Let $\mathbf{V}$ be the matrix, whose columns consist of the eigenvalues of $\mathbf{W}$, and let the diagonal matrix $\mathbf{E}$ contain the corresponding eigenvalues. Applying the diagonalization $\mathbf{W} = \mathbf{V}\mathbf{E}\mathbf{V}^{-1}$, the formula for the matrix exponential reads $\exp(-t\mathbf{W}) = \mathbf{V}\exp(-\mathbf{E}t)\,\mathbf{V}^{-1}$. In our case this means that

$$\begin{aligned}
\mathbf{C}\mathbf{p} &= \exp(-t\mathbf{W})\mathbf{p} = \mathbf{V}\exp(-t\mathbf{E})\mathbf{V}^{-1}\mathbf{p} \\
&= \begin{pmatrix} 0 & i & -i \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{pmatrix} \begin{pmatrix} e^0 & 0 & 0 \\ 0 & e^{-it} & 0 \\ 0 & 0 & e^{it} \end{pmatrix} \begin{pmatrix} 0 & 0 & 1 \\ -i/2 & 1/2 & 0 \\ i/2 & 1/2 & 0 \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} \\
&= \begin{pmatrix} \cos(t) & \sin(t) & 0 \\ -\sin(t) & \cos(t) & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} = \begin{pmatrix} \cos(t)p_1 + \sin(t)p_2 \\ -\sin(t)p_1 + \cos(t)p_2 \\ p_3 \end{pmatrix}.
\end{aligned}$$

Hence,

$$\begin{aligned}
\|\mathbf{C}\mathbf{p}\|^2 = \;& \cos^2(t)p_1^2 + 2\cos(t)\sin(t)p_1p_2 + \sin^2(t)p_2^2 \\
& + \sin^2(t)p_1^2 - 2\sin(t)\cos(t)p_1p_2 + \cos^2(t)p_2^2 + p_3^2 \\
= \;& p_1^2 + p_2^2 + p_3^2 = \|\mathbf{p}\| = 1.
\end{aligned}$$

and therefore $\mathbf{\Psi}_{i,k} = \psi_0 \quad \forall\, i, k$.

7.1.2 Numerical results

a) Smooth initial configuration.

i) Space-independent FPE for ψ_L.

	L^2-error			EOC		
level	min	max	**average**	min	max	**average**
2	2.42e-2	2.59e-3	2.50e-2	-	-	-
3	6.35e-3	6.92e-3	6.61e-3	1.929	1.906	1.918
4	1.62e-3	1.77e-3	1.69e-3	1.972	1.964	1.968
5	4.07e-4	4.47e-4	4.25e-4	1.990	1.989	1.990
6	1.02e-4	1.12e-4	1.07e-4	1.996	1.996	1.996

ii) Full FPE for ψ_E.

	L^2-error			EOC		
level	min	max	**average**	min	max	**average**
2	2.273-2	2.66e-2	2.50e-2	-	-	-
3	6.06e-3	7.43e-3	6.61e-3	1.966	1.842	1.918
4	1.50e-3	2.13e-3	1.69e-3	2.017	1.804	1.968
5	3.75e-4	6.98e-4	4.26e-4	1.997	1.608	1.988
6	9.41e-5	2.86e-4	1.07e-4	1.995	1.287	1.992

b) Discontinuous initial configuration.

i) Space-independent FPE for ψ_L.

	L^2-error			EOC		
level	min	max	**average**	min	max	**average**
2	2.50e-2	2.58e-2	2.50e-2	-	-	-
3	6.61e-3	6.88e-3	6.63e-3	1.919	1.907	1.917
4	1.69e-3	1.76e-3	1.69e-3	1.968	1.964	1.968
5	4.25e-4	4.44e-4	4.26e-4	1.990	1.989	1.990
6	1.07e-4	1.11e-4	1.07e-4	1.996	1.996	1.999

ii) Full FPE for ψ_E.

	L^2-error			EOC		
levels	min	max	**average**	min	max	**average**
2	2.44e-2	2.85e-2	2.51e-2	-	-	-
3	5.98e-3	9.42e-3	6.67e-3	2.031	1.595	1.910
4	1.52e-3	3.14e-3	1.75e-3	1.979	1.584	1.934
5	3.81e-4	2.38e-3	4.90e-4	1.992	0.399	1.832
6	9.55e-5	2.11e-3	1.60e-4	1.998	0.179	1.612

Table 7.2: *Numerical results for both the space-independent and the space-dependent Fokker-Planck equation. The L^2-error with respect to the spheres and the corresponding EOC are calculated at the time $T = 2\pi$.*

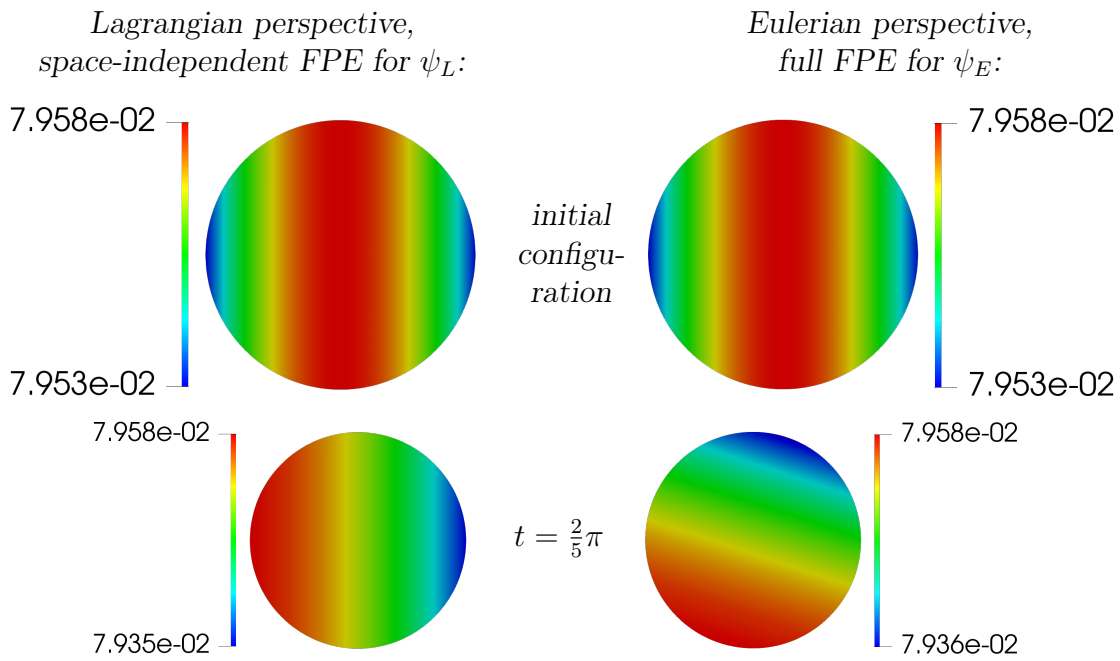
Let $D_r = 0$ and $T = 2\pi$. For the above tests, MCL was applied with $\mathbf{Q}_1$ as polynomial space and we decided to always use the same mesh refinement level for the spatial and the spherical mesh.

For $D_r = 0$ and the chosen initial conditions, we know the exact solution for ψ_L on each sphere at every time. In particular, however, after one revolution, that is, at time $T = 2\pi$, we also know the exact solution for ψ_E.

Section 5.6.3.1 already discussed numerical results for the extended Jeffery equation on a single sphere. Considering the Fokker-Planck equation we solve the Jeffery equation on N spheres, i.e., once for each spatial mesh point.

We calculate the L^2-error for each sphere. The smallest, the largest, and the averaged error are documented. Most of the time all the three EOCs and, in particular, the EOCs corresponding to the averaged L^2-errors are very close to two. Hence, regardless of the spatial initial condition the optimal order of convergence is reached. This makes sense because we primarily calculate the EOC with respect to the spheres, whose initial configurations are always smooth. However, the additional spatial convection term from the full FPE does not only change the configuration in space, but also influences the directly coupled configuration on the sphere. This explains why the averaged EOC for the full FPE deteriorates in the case of the discontinuous spatial initial configuration, see Table 7.2b)ii).
This effect becomes even more apparent when we consider the EOCs for the full FPE with respect to the largest L^2-error, both for the continuous and the discontinuous configuration in space. Overall, compared to the space-independent FPE, the minimal absolute L^2-errors for the full FPE become a little smaller again, while the maximal absolute L^2-errors increase significantly, see Table 7.2.

In Figure 7.3, the Lagrangian and the Eulerian perspective are compared by visualizing a spatial configuration of them during a time interval of length 2π.



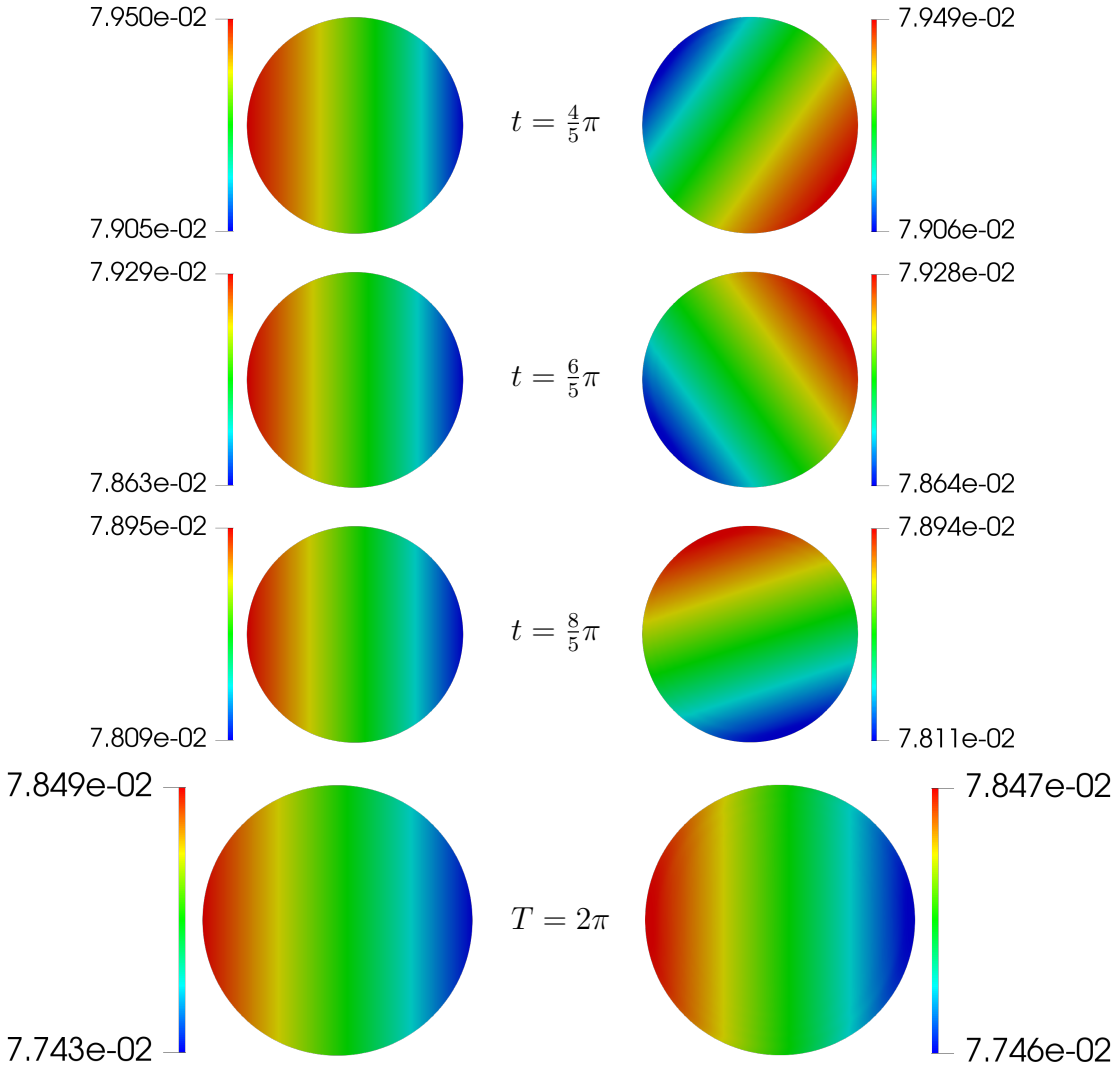


Figure 7.3: *Development of the continuous initial configuration for ψ_L and ψ_E. The visualization is for the spherical grid point $k = 6$. Level 5 was used for both meshes.*

Considering Figure 7.3, first and foremost, for ψ_E the effect of the spatial convection term, which describes a rotation, is remarkable. For both the Lagrangian and the Eulerian framework, for the specific spherical grid point k, the values become smaller over time. The final states at $T = 2\pi$ only show minor deviations in their extreme values and cannot be distinguished visually.

7.2 Axisymmetric Contraction Benchmark

7.2.1 Setting

FPE-NSE system. Now we are able to implement what was already introduced at the beginning of this thesis: the fiber-flow coupling. The Fokker-Planck equation, which governs the behavior of the fibers on the microscopic scale, is coupled to the Navier-Stokes equations, which model the macroscopic flow. The FPE-NSE system and the corresponding data exchange are visualized in Figure 7.4.

$$\textbf{FPE:} \quad \frac{\partial \psi}{\partial t} + \nabla_{\mathbf{x}} \cdot (\mathbf{u}\psi) + \nabla_{\mathbf{p}} \cdot (\dot{\mathbf{p}}\psi) = \Delta_{\mathbf{p}}(D_r \psi);$$
$$\dot{\mathbf{p}} = \mathbf{W}\mathbf{p} + \lambda_e \left[\mathbf{D}\mathbf{p} - (\mathbf{D} : (\mathbf{p} \otimes \mathbf{p}))\,\mathbf{p}\right];$$
$$\mathbf{D} = \mathbf{D}(\nabla \mathbf{u}), \ \mathbf{W} = \mathbf{W}(\nabla \mathbf{u})$$

$\mathbf{u}\ (\&\ \nabla \mathbf{u})$ ⇑ ⇓ $\mathbf{A}(\psi)\ \&\ \mathbb{A}(\psi)$

$$\textbf{NSE:} \quad \rho \left(\frac{\partial \mathbf{u}}{\partial t} + (\mathbf{u} \cdot \nabla)\mathbf{u} \right) = \rho \mathbf{g} - \nabla p + \nabla \cdot \boldsymbol{\tau};$$
$$\nabla \cdot \mathbf{u} = 0;$$
$$\boldsymbol{\tau} = 2\mu_I(\mathbf{D} + N_p \mathbb{A} : \mathbf{D} + N_s(\mathbf{D}\mathbf{A} + \mathbf{A}\mathbf{D}))$$

Figure 7.4: *Coupling between the full FPE and the generalized NSE.*

The upper orange block involves the Fokker-Planck equation combined with Jeffery's equation. The lower reddish block consists of the generalized incompressible Navier-Stokes equations, that is, the conservation laws for mass and momentum. Because of the extended stress tensor a non-Newtonian fluid is described.

All the threads of the thesis come together. Operator splitting is the prerequisite for everything else. We split between the FPE and the NSE and, in particular, we split within the FPE and within the NSE. In contrast to the previous benchmark, limiting for the FPE as a whole becomes necessary to ensure the normalization property of the PDF ψ, see Section 6.2.

Moreover, instead of using an exact velocity field, a discrete velocity field $\mathbf{u}$ is computed with the NSE. To solve the FPE, not only the velocity $\mathbf{u}$, but also the corresponding Jacobians $\nabla \mathbf{u}$ are required, since $\dot{\mathbf{p}} = \dot{\mathbf{p}}(\mathbf{D}(\nabla \mathbf{u}), \mathbf{W}(\nabla \mathbf{u}))$. The $\nabla \mathbf{u}$ are reconstructed from $\mathbf{u}$ to take into account the relationship between $\mathbf{u}$ and $\nabla \mathbf{u}$. A possible procedure is described in Section 6.3.

A striking enhancement from the Jeffery to the current benchmark is the computational area. We have 3d geometry instead of a 2d geometry. This increases the computational effort enormously, but the benchmark comes closer to real applications. The new domain gave our 'Axisymmetric Contraction Benchmark' its name.

Remark (Data transfer from the NSE to the FPE)**.** Following Figure 7.4, we only had to pass the velocity field $\mathbf{u}$ to transfer the data from the NSE to the FPE. In practical implementation, however, we pass the data bundled in two structs,

which contain all the data necessary to solve the FPE in both the physical and the configuration space. Thus, all the velocity dependent data such as the convection matrices are calculated only once for each step and not multiple times if, for instance, subtime stepping is used.

One- and Two-way coupling. We distinguish between one- and two-way coupling, also named decoupled and coupled approach. One-way coupling means that the flow kinematics influence the fiber orientation, whereas the flow kinematics remain unaffected by the fiber orientation. This changes in the framework of two-way coupling, where both phases influence each other.

Restricting the simulations to a one-way coupling is a good starting point [FVDC92, ZMLD+12]. Furthermore, some setups justify to compute the fluid motion without reference to the fiber orientation. In [Tuc91], it is stated that a one-way coupling is often in good agreement with experimental results, although such an approach is not consistent with theory.

Ultimately, however, the goal is a two-way coupling between the Fokker-Planck equation and the Navier-Stokes equations, since not only the flow influences the fibers, but vice versa also the fibers have an effect on the fluid. To transfer information about the fiber orientation from the FPE to the NSE we use the orientation tensors $\mathbf{A}$ and $\mathbb{A}$, which are calculated from the PDF ψ by formula (7.2). The orientation tensors are needed to determine the stress tensor for the generalized NSE. In fact, the fibers influence the stress, which in turn influences the kinematics.

Research papers, which show that a two-term coupling is important to faithfully model the complicated rheological properties of a suspension are, for example, [Ver98, VT02, Kne06, KOM09] or [FMAA20]. In [LDHB88] it is stated that the assumption of streamlines unaffected by the presence of fibers is even incorrect at fiber concentrations below 0.1%. In [FHH+08] the difference between the coupled and uncoupled approach was not found to be significant for the orientation but for the velocity field.

Axisymmetric Conctraction. The contraction is a common benchmark geometry in both 2d and 3d [OP02, Ch.8]. We consider a 3d contraction. As depicted in Figure 7.5 the geometry is composed of a long wide pipe, where the fluid flows in and a smaller narrow pipe, where the fluid flows out. The fluid passes a circular entry and then moves from a tube of radius R into another tube of smaller radius r. Specifically, we use diameter $D = 0.045$ and length $L = 0.09$ at the upstream side and $d = 0.01$ combined with $l = 0.02$ on the downstream side. Consequently, $L/l = 4.5$ and, in particular, $D/d = 4.5$, which defines the contraction ratio 4.5:1. Different contraction ratios are used in the literature, and sometimes the results for various ratios are contrasted [OP02, Ch.8], [OFCH04]. For our simulations we make use of a 4.5:1 contraction since in this case different reference solutions are available in [LDHB88, Ver98, VT02] and [Loh19]. All these works have in common that they use the Folgar-Tucker instead of the Fokker-Planck equation. They differ in the simplifications they make and the closures they apply.

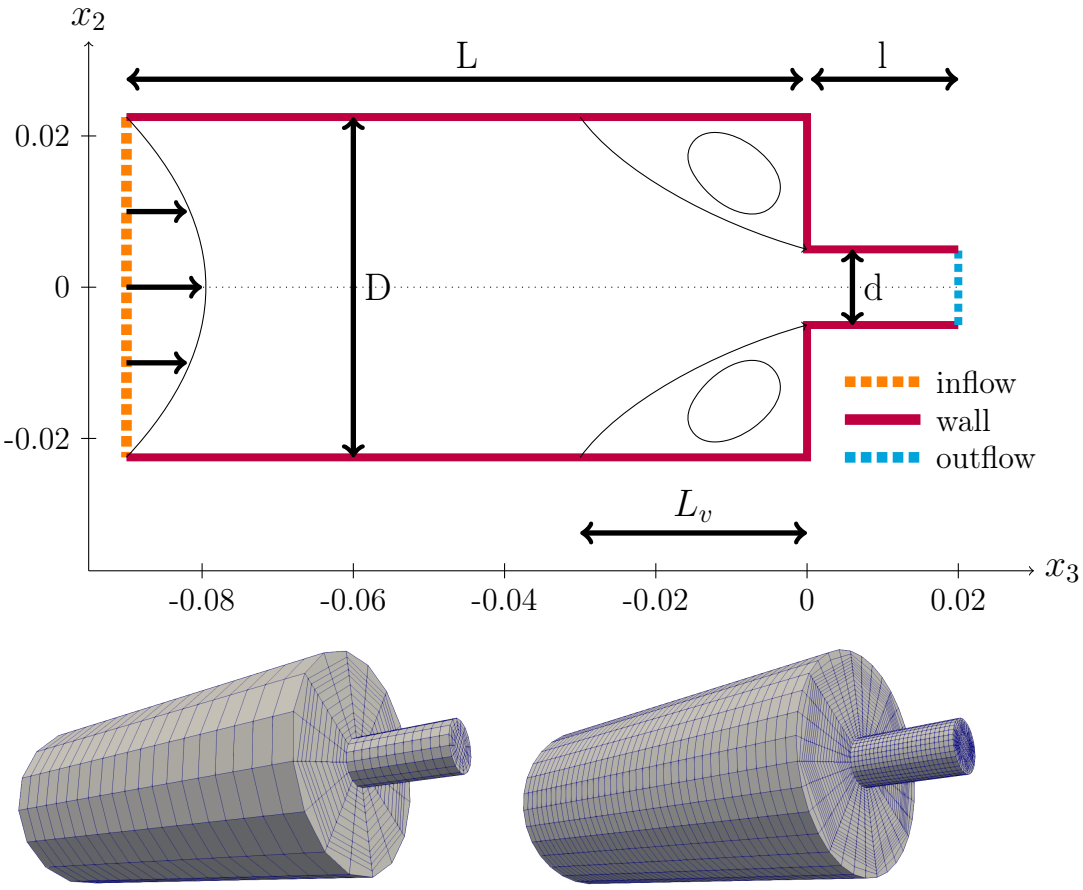


Figure 7.5: *The axisymmetric 4.5:1 contraction. At the top: Schematic representation of the cross-section through the geometry. At the bottom: Discretization with a coarse mesh (level 1; 6969 nodes) and with a finer mesh (level 2; 53393 nodes).*

Further geometries like a plate [BT92a], an axisymmetric expansion or a center-gated disk [Ver98, VT02] are conceivable. Within the plate different layers are observed. The expansion can be considered as an inversion of the contraction geometry. A corresponding flow can be found in many injection molding processes. Using the center-gated disk a radially diverging flow results. However, a huge advantage of our contraction benchmark is that despite the simple geometry Newtonian and non-Newtonian fluids exhibit a significantly different behavior.

Boundary conditions for the NSE. To define a benchmark apart from the computational area, initial and boundary conditions are necessary. We specify the velocity for wall, inflow and outflow boundaries. For the former two we apply Dirichlet boundaries, while for the outflow Neumann boundaries are used. We set $\mathbf{u}_{\text{wall}} = \mathbf{0}$, which is referred to as no-slip condition and reflects that the velocity of the solid wall is adopted. At the inflow we assume a parabolic velocity profile, namely

$$\mathbf{u}_{\text{in}} = u_{\max}\left(1 - \frac{x_1^2 + x_2^2}{R^2}\right)\mathbf{e}_3, \text{ where } u_{\max} = \frac{1}{45}.$$

Hence, radial symmetry is given. The velocity reaches its maximum $u_{\max}$ at the center of the cylinder ($x_1 = x_2 = 0$), whereas the velocity disappears at the edge

$(x_1^2 + x_2^2 = R^2)$. This way even pointwise compatibility between $\mathbf{u}_{\text{in}}$ and $\mathbf{u}_{\text{wall}}$ is obtained. On the outflow boundary we apply homogeneous Neumann conditions, that is, $\boldsymbol{\tau}\mathbf{n} - p\mathbf{n} = \mathbf{0}$. This reflects that no forces act there. The initial velocity is obtained solving the Stokes problem corresponding to our Navier-Stokes problem.

Initial and boundary conditions for the FPE. The initial condition for the Fokker-Planck equation describes the orientation of the fibers at the beginning of the simulation. The two basic configurations are the isotropic/random state and the fully aligned state.
In the isotropic case, the fibers are uniformly distributed in space, which can be approximated by the orientation tensor $\mathbf{A} = \operatorname{diag}(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. The corresponding orientation distribution reads $\psi(\mathbf{p}) = \frac{1}{4\pi}$, see (2.13). Each row of the coefficient matrix Ψ has to reflect this random-in-space distribution. We accomplish this by choosing

$$\psi_{*,k} = \frac{1}{4\pi}.$$

Positivity and symmetry of the FE function ψ_h are obvious. In addition, the normalization property is satisfied, since

$$\int_{\mathbb{S}^2} \psi_h \, \mathrm{d}\mathbf{p} \overset{L3.6c)}{=} \sum_{k=1}^{M} m_k \psi_{*,k} \overset{L3.6b)}{=} \frac{1}{4\pi} |\mathbb{S}^2| = \frac{4\pi}{4\pi} = 1.$$

Full alignment means that the fibers are aligned along the main flow direction. If this, for example, is the x_1-direction, the orientation tensor reads $\mathbf{A} = \operatorname{diag}(1, 0, 0)$. For arbitrary directions, which do not correspond to the orientation of a coordinate axis, the tensors become more complex. However, all tensors have the eigenvalues 0, 0 and 1; and the eigenvector to eigenvalue 1 is oriented in the main flow direction. On the continuous level, the fully aligned state is described by a delta distribution $\delta(\mathbf{p_0})$. To fill the coefficient matrix Ψ, for each point in space two opposite points on the sphere, which fit the direction of the velocity the best, are determined. Let us denote the corresponding indices by $\underline{k}$ and $\overline{k}$. We set the entries of Ψ to

$$\psi_{*,k} = \begin{cases} \frac{1}{2m_k}, & \text{if } k \in \{\underline{k}, \overline{k}\}, \\ 0, & \text{otherwise.} \end{cases}$$

The positivity of the m_k guarantees the non-negativity of ψ, symmetry is ensured by construction, and the mass for every point in space is given by $\frac{1}{2}\left(\frac{m_{\underline{k}}}{m_{\underline{k}}} + \frac{m_{\overline{k}}}{m_{\overline{k}}}\right) = 1$, so that the normalization property is satisfied here as well. Since the orientation tends to be only partially aligned in practice, the corresponding initial condition might also be approximated by experimental results [KOM09, Fig. 3].

In our axisymmetric contraction benchmark we simulate until the steady state is reached. However, especially at the beginning of the simulation the initial condition influences the overall result. Both fully aligned and isotropic condition are reasonable, whereby the latter is usually assumed in the literature. We also impose an isotropic initial condition in order to compare with the results in [Loh19].

Furthermore, it is of utmost importance to specify a boundary condition as well. For the previous Jeffery benchmark we could only avoid this because of the circular geometry. Consistent with the initial condition, we choose an isotropic boundary condition. If the boundary condition is omitted completely, an aligned state is unintentionally created in the inflow region. In [KOM09] it is reported that depending on how the fluid enters the geometry the influence of the one- vs. two-way coupling varies. While an aligned orientation orientation state only marginally changes when mutual coupling is activated, starting from an isotropic state there is a significant difference between the coupled and the uncoupled approach. In particular, the fibers were observed to align more quickly in the case of two-way coupling.

Implementation aspects. To solve the Fokker-Planck equation we use an operator splitting approach and update the coefficient matrix as outlined in Figure 6.1. Let us first consider the advection equations in the physical space Ω. For the Jeffery benchmark, a global analytical velocity field $\mathbf{u}$ was given, whereas now we have an individual velocity field for each grid point of the spatial mesh. Therefore, technically the assembly of the spatial convection matrix deviates. In both cases, however, a single convection matrix $K_{\mathbf{x}} \in \mathbb{R}^{N\times N}$ results, and only one system matrix has to be assembled per time step.

To solve the convection-diffusion equations on the sphere we change the 'direction of solving', that is, the rows of coefficient matrix are updated instead of its columns. For each row, that is, for each of the N points in physical space, an individual convection matrix $K_{\mathbf{p}} \in \mathbb{R}^{M\times M}$ has to be assembled. This makes the updates in the configuration space expensive. For the practical implementation, it has to be weighed up, whether to assemble the respective convection matrix again every time it is needed or whether to assemble all the matrices one time at the beginning and save them in a suitable data structure. While the former approach is computational intensive, the latter requires a huge amount of memory. We chose the second method, since our compute servers provided enough memory.

To obtain the orientation tensors $\mathbf{A}$ and $\mathbb{A}$ it is neither feasible nor desirable to compute the integrals given in (2.21) analytically. Instead, we apply Lemma 3.6c) to approximate the entries of the tensors numerically. The formulas read

$$a_{ij} = \sum_{k=1}^{M} m_k p_k(i) p_k(j) \psi_k, \tag{7.2a}$$

$$a_{ijlm} = \sum_{k=1}^{M} m_k p_k(i) p_k(j) p_k(l) p_k(m) \psi_k. \tag{7.2b}$$

The m_k are the entries of the lumped mass matrix with respect to the sphere, $p_k(i)$ is the i^{th} entry of the orientation vector in k^{th} direction of the sphere, the ψ_k are the entries of the discrete FE solution vector of $\psi(\mathbf{x}, \mathbf{p}, t)$ and M is the number of mesh points on the discretized sphere.

Parameters. Finally, the different parameters have to be set in order to define the concrete problem. For the Fokker-Planck equation, involving Jeffery's equation, we choose

- $\lambda_e = 99/101$ (shape parameter)
- $D_r = C_I\sqrt{\mathbf{D} : \mathbf{D}} = 2\text{e-}3\sqrt{\mathbf{D} : \mathbf{D}}$ (diffusion coefficient).

For the Navier-Stokes equations, involving tensor $\boldsymbol{\tau}$, we set

- $\rho = 1$ (density),
- $\mu_I = 0.1$ (viscosity parameter),
- $N_s = 0$ (shear-rate number),
- $N_p = 6$ (particle number).

The choice $N_s = 0$ reduces the extended tensor to $\boldsymbol{\tau} = 2\mu_I(\mathbf{D} + N_p\mathbb{A} : \mathbf{D})$, since equation (2.33a) simplifies to (2.33b). Consequently, only the fourth-order orientation tensors $\mathbb{A}$ have to be transferred from the FPE to the NSE. With the given parameters, the Reynolds number can be defined by

$$\mathrm{Re} := \frac{u_{\max} \cdot R}{\mu_I} = 0.005.$$

Since the pipe with the larger radius R dominates the contraction geometry, the radius R is used instead of r. The given Reynolds number indicates an extremely laminar flow, so that the Stokes equations might be applied, see Remark 2.11. However, solving the NSE in our coupled system does not require the lion's share of the computational time. Therefore, we stick to the more general NSE.

The reference simulation is performed using

- level 1 in space, i.e., 6969 nodes,
- level 4 on the sphere, i.e., 1538 nodes,
- final time $T = 25$,
- 2500 FPE-NSE steps.

We vary the number of FPE-NSE steps in Section 7.2.2.2. Subsequently, in Section 7.2.2.3 the consequences of different spherical levels are examined, while we compare the results for level 1 and level 2 in space in Section 7.2.2.4. The effect of a varied particle number N_p, which involves the difference between one- and two-way coupling, is addressed in Section 7.2.2.5.

Chosen discretizations. For the Navier-Stokes equations we employ $\mathbf{Q}_2\mathbf{P}_1^{\text{disc}}$, that is, we use triquadratic functions for the velocity and discontinuous linear functions for the pressure, see Section 6.4.

Since $\mathbf{Q}_1$ is used to solve the Fokker-Planck equation numerically, the orientation tensors $\mathbf{A}$ and $\mathbb{A}$ are described by linear elements as well. The tensors from the space of linear functions are projected to the space of quadratic functions to adapt them to the velocity space $\mathbf{Q}_2$ [Ess22].

There is not much difference between solving the steady or the unsteady Navier-Stokes equations because of our very small Reynolds number. We consider the unsteady equation and use BDF(2) as a time stepping scheme matching to $\mathbf{Q}_2$.

Algorithm. The basic procedure to solve the FPE-NSE system reads [Kne06]

1.) Initialize the system.
2.) Update the velocity field $\mathbf{u}$ using the NSE solver.
3.) Update the coefficient matrix $\boldsymbol{\Psi}$ in Ω by iterating over the mesh points in $\mathbb{S}^2$ and by solving a homogeneous advection equation.
4.) Update the coefficient matrix $\boldsymbol{\Psi}$ in $\mathbb{S}^2$ by iterating over the spatial mesh points and by solving the space-independent FPEs.
5.) Update the stress tensor $\boldsymbol{\tau}$. Return to step 2.), and repeat until the final time is reached or a termination condition is met.

7.2.2 Numerical results

We perform the numerical studies for the 3d axisymmetric 4.5:1 contraction. In doing so, we are interested in the flow and the orientation state. The results are visualized in the x_2x_3-plane; no differences were found due to a changed cross-section.

7.2.2.1 Results for the reference simulation

Flow state. The behavior of the flow involves the velocity. This, of course, increases in the smaller pipe and especially in the center of this pipe. More interesting are the streamlines and, in particular, the vortices that arise in the angles at the end of the large pipe.

The results related to the flow state are visualized in Figure 7.6. The initial flow field is obtained by solving the NSE for $N_p = N_s = 0$. Already in this starting solution small recirculating corner vortices have formed and their size can be estimated to **0.36R**.

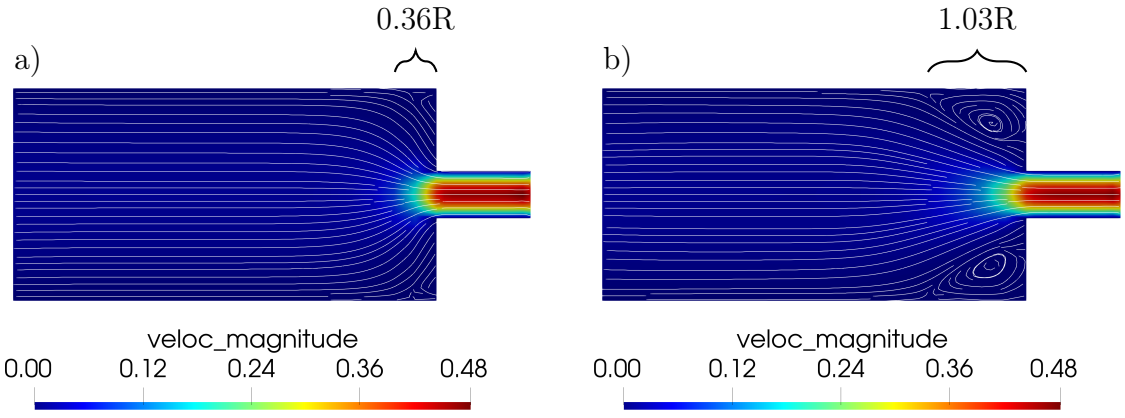


Figure 7.6: *Streamlines, vortices and the magnitude of the velocity at a) $t = 0$ (one-way coupled) and at b) $T = 25$ (two-way coupled).*

The size of the vortices increases when we set $N_p = 6$ in the further course of the simulation. There is some freedom in specifying where the vortices end. Both a lower limit of 0.022, i.e., 0.98R and an upper limit of 0.0255, i.e., 1.08R are

justified. Choosing the mean value **1.03R** our vortex detachment point is in excellent agreement with results from the literature.

In [VT02], where the FTE was solved with an ORF closure, a vortex size of 1.06R was obtained. This is very similar to the results measured for experimental data in [LDHB88, Fig.10a]. In [Loh19] the value 1.04R was obtained as vortex size, when the FTE was solved in combination with a so-called natural B closure, while 0.96R was observed in combination with a less advanced ORS-closure.
Summing up, the streamlines presented in Figure 7.6 demonstrate that the flow behavior is captured correctly by our method.

Orientation state. To explore the orientation of the fibers, we consider the largest eigenvalues of the second-order orientation tensors. An isotropic distribution is described when all eigenvalues, and thus the maximum eigenvalue, are one-third, while an aligned distribution is approximated when the maximum eigenvalue is close to one.

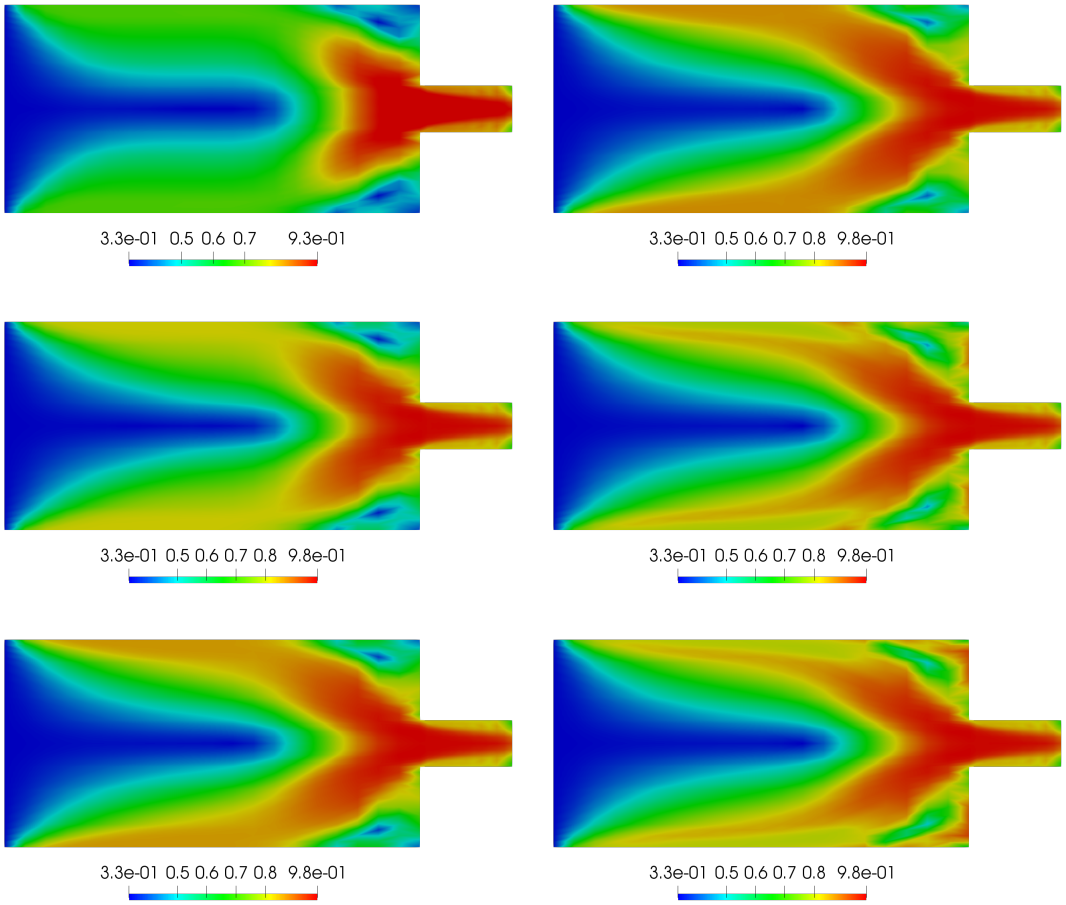


Figure 7.7: *Largest eigenvalues in the x_2x_3-plane at $t = 1.25; 2.5; 3.75$ (left column) and $t = 5; 12.5$ and 25 (right column).*

The development of the maximum eigenvalues is depicted in Figure 7.7. At the beginning the fibers that are nearly perfectly aligned are found at the end of the

larger pipe, in the transition between the two pipes and in the smaller pipe. The largest eigenvalue is 0.93.
The area where the fibers are nearly perfectly aligned slightly spreads to the boundaries when the the simulation progresses and the largest eigenvalue grows to 0.98. The value of about 0.98 demonstrates that a perfect alignment is achieved nowhere. This can be explained by the diffusive component.

Convergence to steady state. Let $a_1^n \leq a_2^n \leq a_3^n$ be the eigenvalues of $\mathbf{A}$ at the current time step. A measure of how stationary the largest eigenvalue a_3^n already is can be defined, for example, by the following two errors:

$$\mathrm{err}_2 := \frac{\|a_3^n - a_3^{n-1}\|_2}{N\Delta t} := \frac{\sqrt{|(a_3)_i^n - (a_3)_i^{n-1}|^2}}{N\Delta t}, \tag{7.3a}$$

$$\mathrm{err}_\infty := \frac{\|a_3^n - a_3^{n-1}\|_\infty}{\Delta t} := \max_{i\in\{1,\dots,N\}} \frac{|(a_3)_i^n - (a_3)_i^{n-1}|}{\Delta t}. \tag{7.3b}$$

For both the Euclidean and the maximum norm the error is divided by Δt. This scaling makes the value larger, but it is reasonable because it allows comparability of the errors for different Δt. The Euclidean norm is additionally averaged over the number of spatial grid points. Checking for '$\mathrm{err}_2 < \mathrm{TOL}_2$' or '$\mathrm{err}_\infty < \mathrm{TOL}_\infty$' could be used as stopping criterion.
We chose $[0, 25]$ as a fixed time interval, since the 'eyeball norm' indicates that we converge to a stationary limit and that at the latest after this period of time not only the velocity field but also the orientation state has reached its stationary limit. From the development of the largest eigenvalues in Figure 7.7, it can be concluded that the major changes of orientation take place at the beginning of the simulation. The same observation is made by studying the quantitative results below in Table 7.8 or Table 7.9. The distance between the maximum eigenvalues decreases faster at the beginning, this change slows down over time until the reached state is said to be stationary.

Trace correction. In the previous chapter we addressed the gradient recovery of the Jacobian $\nabla\mathbf{u}$ from $\mathbf{u}$, including a simple trace correction with respect to the Frobenius norm to ensure that $\mathrm{tr}(\nabla\mathbf{u}) = 0$, see Section 6.3.1. When we performed the gradient reconstruction for our Axisymmetric Contraction Benchmark without any correction of trace as worst cases we saw $\min(\mathrm{tr}(\nabla\mathbf{u})) = -34.50$ and $\max(\mathrm{tr}(\nabla\mathbf{u})) = 28.54$. Consequently, trace correction seems to be reasonable or even necessary.
It is all the more surprising that adding a trace correction has no significant effects. Hardly any differences can be detected in the visualization. In the documentation of err_2 and err_∞ no difference is observed at all. At the same time, simulation times for the trace-corrected case and the uncorrected case are extremely similar. For the remaining test cases, we therefore always include our trace correction.

'Quality management'. Finally, we review the overall quality of the simulation with respect to the property preservation. First, we check whether on a single sphere

the mass of one, representing the normalization property, is maintained. In fact, the largest deviation at $T = 25$ downward with respect to *all* spatial points is -2.1e-12, whereas the largest deviation upward is 1.3e-12. The average deviation is 6.3e-14, which is close to the machine accuracy. All in all, this is perfectly acceptable, especially since we already discussed in Chapter 5 that the transformation from a Cartesian to a curved element can introduce a slight error.
A further aspect to check is that for the eigenvalues it always holds that $0 \leq a_1 \leq a_2 \leq a_3 \leq 1$. Figure 7.7 indicates the largest eigenvalue is always smaller than one. Analogously we checked that the smallest eigenvalue is always greater than zero, so that no ad hoc correction for our orientation tensors $\mathbf{A}$ is necessary.

7.2.2.2 Number of basic steps

When coupling the FPE and the NSE, the question arises how often to switch between the PDEs, that is, how to choose the basic time step size Δt. This is a heuristic choice. So far we have used 2500 steps to solve the NSE on the time interval $[0, 25]$, i.e., $\Delta t = 0.01$.
To evaluate the effect of different numbers of basic time steps, we compare the results for 1000, 2500 and 5000 FPE-NSE steps. The other parameters remain unchanged compared to the reference setting. The number of substeps for the space-dependent and -independent part of the FPE is determined individually via the CFL-condition.

Pros and Cons. An argument for fewer basic steps is a reduced computational time. In fact, our local MPSC solver for the NSE can handle significantly larger time steps than $\Delta t = 0.01$. However, the savings are not passed on one-to-one to the total computational time because more substeps are required for the FPE.
An argument that justifies more basic steps, despite of the increased numerical effort, is the stronger coupling between FPE and NSE. However, especially in our benchmark with an extremely small Reynolds number, we have to check if there is still a change in the velocity field that is worth to be calculated at all.

basic time steps	1000		2500		5000	
time	err_2	err_∞	err_2	err_∞	err_2	err_∞
0.05	2.13e-2	15.38	1.70e-2	14.85	1.61e-2	13.61
0.10	9.06e-3	7.14	8.03e-3	5.30	7.71e-3	4.30
0.15	5.35e-3	2.21	4.94e-3	1.40	4.86e-3	1.38
0.20	4.14e-3	1.09	3.97e-3	1.06	3.92e-3	1.04
0.25	3.51e-3	0.965	3.42e-3	0.941	3.40e-3	0.932
2.50	3.91e-4	0.076	3.91e-4	0.075	3.91e-4	0.076
25.00	1.23e-5	0.012	1.16e-5	0.010	1.13e-5	0.0009

Table 7.8: *For a different number of basic steps: Distance of the largest eigenvalues comparing the given step and its predecessor in the norms defined by (7.3).*

In Table 7.8 we observe overall quite similar results for a different number of time steps -as it is desirable. However, as it might be expected, err_2 and err_∞ become slightly smaller when more basic steps are used.

Computational time. The number of basic steps primarily affects the time needed to solve the NSE. This time is proportional to the number of basic time steps. The time spent on the FPE increases with more basic steps as well, but only to a limited extend. This is because the larger number of basic steps in relevant parts is compensated by a smaller number of substeps. In general, the computational time is influenced by the tolerance required by the solver for the NSE.

7.2.2.3 Different levels on the sphere.

The level of the spherical mesh also has a significant impact on the duration of the simulation. In Section 5.6, we did different numerical studies on a single sphere. Meanwhile, however, we consider a sphere for each spatial grid point, that is, a few thousand spheres. Consequently, refining the mesh increases the computational effort drastically. In this context, the question arises whether a large number of orientation points is necessary at all, since not the PDF ψ but the orientation tensors $\mathbf{A}$ couple the FPE to the NSE. In fact, it is assumed that the accuracy of ψ is not that critical, see also [Kne06, KS09a].

As before, the largest eigenvalues of the orientation tensors are used as metric. In analogy to Table 7.8, Table 7.9 quantifies to what extent the orientation has become stationary at different times.

sphere level	2		3		4		5	
time	err_2	err_∞	err_2	err_∞	err_2	err_∞	err_2	err_∞
0.05	1.26e-2	12.96	1.59e-2	14.81	1.70e-2	14.85	1.73e-2	14.65
0.10	5.92e-3	4.50	7.18e-3	5.07	8.03e-3	5.30	8.94e-3	5.45
0.15	4.07e-3	1.27	4.69e-3	1.40	4.94e-3	1.40	5.09e-3	1.39
0.20	3.35e-3	0.959	3.78e-3	1.02	3.97e-3	1.06	4.03e-3	1.07
0.25	2.91e-3	0.787	3.29e-3	0.902	3.42e-4	0.941	3.46e-3	0.947
2.50	3.31e-4	0.091	3.80e-4	0.083	3.91e-4	0.075	3.99e-4	0.079
25.00	7.23e-6	0.008	8.82e-6	0.008	1.16e-5	0.010	1.27e-5	0.011

Table 7.9: *For different spherical meshes: Distance of the largest eigenvalues comparing the given step and its predecessor in the norms defined by (7.3).*

First, we expect that the errors are getting smaller when the mesh is refined. In fact, however, even a trend in the opposite direction is observed for both err_2 and err_∞. The distance measured in the maximum norm and in the Euclidean norm on average is smaller for the coarser mesh.
In Figure 7.10 the orientation states at time $T = 25$ are visualized. Taking a look at the plots the results obtained with sphere level 2 can be questioned, whereas the other meshes lead to plots, which can be hardly distinguished by the eye. Thus, the grid for level 2 is probably so coarse that reliable convergence to the correct result cannot be expected everywhere.

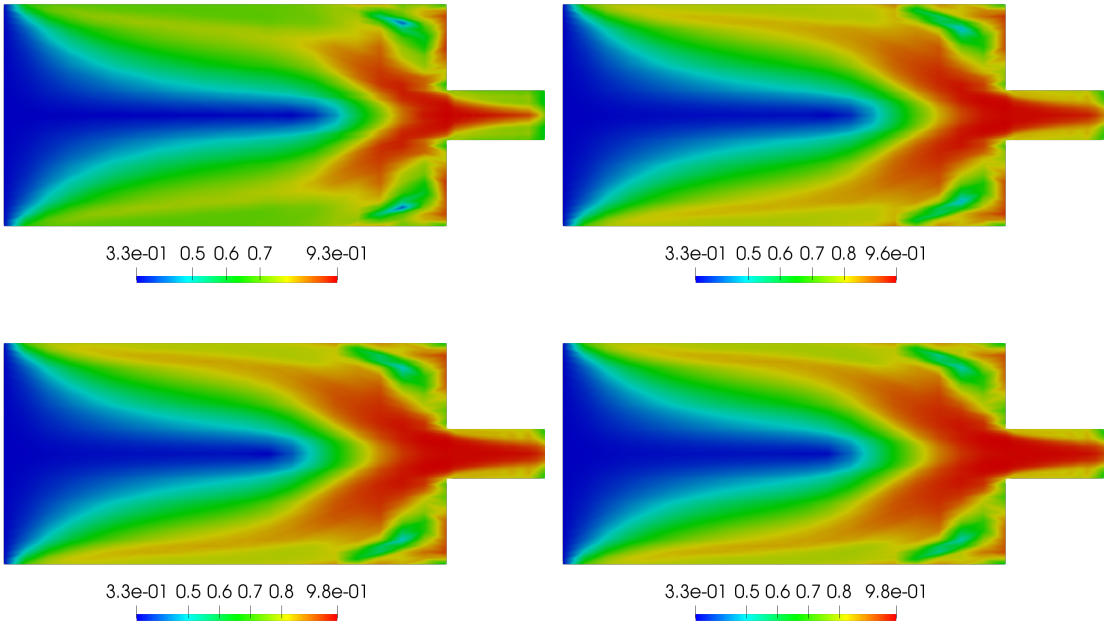


Figure 7.10: *Largest eigenvalues for $T = 25$ when the levels 2, 3 (left), 4 and 5 (right) are used for the spherical mesh.*

In Paragraph 7.2.2.2, where the number of basic steps was varied, the simulation time for the NSE was influenced, while that for the FPE remained fixed. Varying the number of spherical mesh points, it is the other way round. The simulation time for the NSE remains unaffected, whereas there is a direct correspondence between the number of spherical vertices and the time needed to solve the FPE.
The quadrilateral mesh is designed in a way that the number of vertices quadruples from level to level, see Section 5.5.2. This is reflected in the computational time for the FPEs, which quadruples or nearly quintuples in some cases. Summing up, it is reasonable to choose the spherical mesh of level 3 with its 386 vertices.

7.2.2.4 Different levels in space.

For this paragraph we used the spherical mesh on level 3 to compare the results for the spatial meshes on level 1 and 2. No differences were identified with respect to the vortex detachment point.

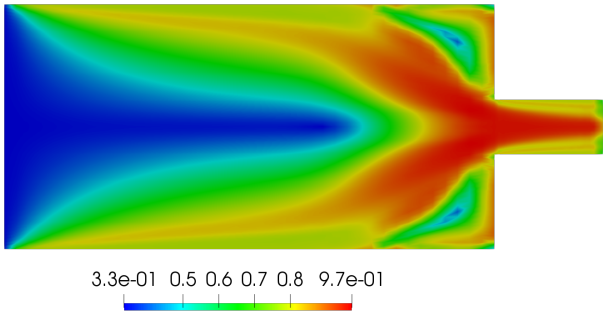


Figure 7.11: *Final orientation state at $T = 25$ for $N_p = 6$ if level 2 is used for the spatial mesh.*

The only noticeable difference compared to our previous simulation results, see, e.g., Figure 7.7, is that the small isotropic area in the vortex region is preserved better over time.

7.2.2.5 Different particle numbers N_p

Previously, $N_p = 6$ was used as a typical value. To get a feeling for the practical implications of different N_p, we also look at $N_p = 0$ and $N_p = 20$. The shear parameter $N_s = 0$ will not be changed.

Choosing $N_p = 0$ reduces the generalized NSE back to a Newtonian NSE. This breaks the flow/orientation coupling. The velocity field has to be calculated only once at the beginning and can then be used throughout the whole simulation, that is, it is not affected by the orientation. Accordingly, the streamlines and vortices are always the same as in Figure 7.6a).

Conversely, $N_p = 20$ represents an increased fiber volume. We expect the suspension to exhibit a stronger non-Newtonian behavior. In fact, in Figure 7.12 we observe a further enlarged vortex, which can be estimated to be approximately **1.34R**.

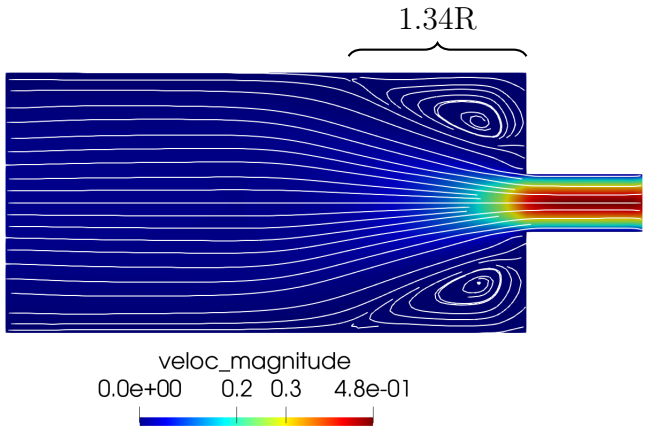


Figure 7.12: *Final flow state at $T = 25$ for $N_p = 20$.*

In Figure 7.13, we contrast the fiber orientation for $N_p = 0$ and $N_p = 20$. We observe that the geometry of the area, where we have an isotropic distribution, is different. The plot of the largest eigenvalue for $N_p = 20$ is similar to the last plot in Figure 7.6 and Figure 7.10, where $N_p = 6$ was used.

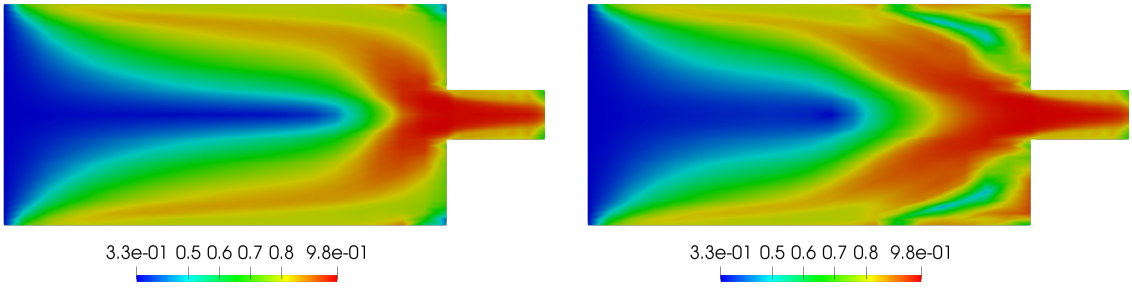


Figure 7.13: *Largest eigenvalues at $T = 25$ for $N_p = 0$ (left) and $N_p = 20$ (right).*

8 Conclusions

8.1 Summary

Each Fokker-Planck equation models a probability distribution function that does live in both a physical and a configuration space. The result is a high-dimensional problem, which makes the Fokker-Planck equation a numerically challenging problem. Applying an operator splitting approach the high-dimensional problem can be divided into smaller subproblems, i.e., in a linear advection equations in the physical space and in a convection-diffusion equations in the configuration space.

Various numerical approaches from the literature to solve Fokker-Planck equations are summarized in **Chapter 1**. In this thesis, we used a deterministic approach and applied property-preserving finite element discretizations for the subproblems.

In **Chapter 2**, our specific Fokker-Planck equation was introduced. It models the orientation of fibers in a suspension, so that the configurational space is chosen as the unit sphere $\mathbb{S}^2$. If we restrict the Fokker-Planck equation to its space-independent terms and ignore the diffusive term, an analytical solution is known. We use this several times as a reference solution.

Other PDEs relevant to this thesis were presented and their relationship was worked out. Jeffery's equation provides an expression for the velocity of the fibers, whereas the velocity field of the fluid is simulated by the Navier-Stokes equations. However, in the first part of the thesis we do not consider the fully coupled FPE-NSE system but the FPE as a single building block.

We derived the Folgar-Tucker equation from the Fokker-Planck equation as an alternative and a low-order approximation of it. While the Fokker-Planck equation directly computes the orientation distribution function and provides an extremely precise description of the orientation, the Folgar-Tucker equation uses tensors to characterize the orientation, which significantly reduce the numerical effort. The orientation tensors require empirical reconstructions, but if these are well chosen, good results are possible. To define the stress tensor, that is responsible for the coupling to the Navier-Stokes equations, the orientation tensors are needed anyway.

Chapter 3 covered the finite element method. We discussed the choice of different function spaces and basis functions. With regard to the necessary property-preserving procedure we decided to use the continuous Galerkin approach, where the numerical solution is expressed in terms of (bi-)linear Lagrange basis functions. Considering the resulting finite element matrices, the consistent mass matrix is often

replaced by a lumped mass matrix for reasons of efficiency and stability or because it is necessary for our limiting strategy.

Chapter 4 explained the AFC methodology, which works on the level of finite element matrices. We first contrasted low- and high-order schemes for a scalar advection equation. Via an additional artificial diffusion matrix and a lumped mass matrix, the low-order scheme ensures that no physical boundaries are violated. Since this happens at the expense of the order of convergence, we must find a mechanism that replaces the high-order method by a low-order method only where necessary. The result is a limited high-order method.

In this thesis we implemented an MCL approach. Considering a scalar advection equation, each solution value at a new time can be written as a linear combination of previous solution values. Restricting this linear combination via a CFL-condition it even becomes a convex combination, and thus ensures that the new value stays in a given invariant domain. Using a straightforward extension of the MCL methodology to convection-diffusion equations, the adapted CFL-condition is the more restrictive the larger the magnitude of the diffusion coefficient becomes.

To design a numerical test problem for the advection equation, a suitable velocity field is needed. We chose a so-called solid body rotation. As configurations to be rotated both continuous and discontinuous geometries were used. By considering not only the results of the MCL algorithm but also from the low-order and unlimited high-order scheme, the need for limiting is well illustrated. For instance, we observe how much the solution smears when the low-order scheme is applied. In addition, the influence of the chosen time stepping is demonstrated. While the backward Euler method performs well, we obtain strong oscillations when the forward Euler method is combined with the baseline Galerkin approach. However, in combination with MCL the second-order SSP-RK scheme, which is based on a convex combination of forward Euler steps, is an excellently suited time stepping scheme.

In **Chapter 5**, we paid special attention to the part of the FPE living on the sphere. We embedded the sphere in the larger context of manifolds. This included the insight that each small piece of the unit sphere $\mathbb{S}^2 \subset \mathbb{R}^3$ can locally be identified with an element in 2d, so that one dimension is saved in numerical computations. The numerical tests demonstrate that there arises a small error due to the transformation between 2d and 3d space. However, this error can be minimized by increasing the order of the transformation and by refining the mesh.

In order to define differential operators and integration on manifolds, we need the Gram matrix, which acts as metric tensor. Moreover, the concept of tangent planes is relevant, not least because the surface divergence can only be applied to fields that are tangential to the manifold.

Overall, we can conclude that despite some additional mathematical requirements, essential software components can be reused for PDEs on surfaces.

With **Chapter 6** we did the last steps towards the coupled FPE-NSE system. First, we focused on the concrete implementation of the operator splitting approach for the FPE. The PDEs in physical and configuration space are solved alternately. For

the practical realization the coefficients of the FE solution are stored in the rows and columns of a large matrix. We benefit from the modularity of the code and from the subtime stepping, which avoids that the smallest time step must be chosen globally. Another potential advantage, especially in the context of our high-dimensional FPE, is the option to parallelize the computations for all PDEs in physical space and, analogously, for all PDEs in configuration space. This possibility is reduced by the normalization limiter, which introduces a coupling between both subproblems. However, this additional limiter is essential to ensure that the normalization property of the ODF is preserved, i.e., that the mass on each sphere stays one. Concretely, our MCL-limiter has to be supplemented by further algebraic corrections.

To calculate the rotational velocity of the FPE, the Jacobian of the suspension velocity is needed. In simpler test cases, we provided it analytically. Once the velocity of the suspension is obtained from the NSE, the associated Jacobians must be reconstructed numerically. We do this with a standard L^2-projection. Without further modification, the trace of these Jacobians is not zero, although it should be in the case of solenoidal velocity fields. Therefore, we performed a trace correction with respect to the Frobenius norm, even though the effect on the overall result of the simulation turned out to be almost negligible.

Finally, in Section 6.4 we derived how the NSE can be written as a saddle point problem. We explained why $\mathbf{Q}_2\mathbf{P}_1^{\text{disc}}$ is a suitable FE pair for velocity and pressure, and why a local multilevel pressure Schur complement solver was used for the NSE in our application.

Chapter 7 covers two benchmarks, one for the full FPE and one for the coupled FPE-NSE system. The former takes up and develops the basic idea of solid body rotation. For a suitably chosen spatial velocity field, there is a temporal equivalence between the solution of the full FPE and the solution of the space-independent FPE. In the case of vanishing diffusion, an analytical solution for the space-independent FPE is known an can be used as a reference solution for the full FPE.

The coupled FPE-NSE system was studied in the framework of the axisymmetric contraction benchmark. This geometry has some appeal to the rheological community, since Newtonian and non-Newtonian fluids behave significantly different. For non-Newtonian fluids the coupling, which takes into account the mutual influence of fibers and fluid, is achieved by using an effective stress tensor. This tensor depends on the second- and fourth-order moments of the orientation distribution function.

In the numerical results we could observe the development of the largest eigenvalues of the orientation tensors over time. These provide information about the primary orientation direction of the fibers. Moreover, we could verify that with increasing particle number, corresponding to an increasing fiber volume, the size of the vortices at the end of the larger pipe of the contraction increases as well. The effect of different sizes of the meshes and the time steps on the accuracy of the numerical results was investigated.

8.2 Outlook

This thesis provides the various building blocks to solve the Fokker-Planck equation and also the coupled FPE-NSE system numerically. The focus was on a theoretically validated approach and, in particular, on preserving physical properties. However, for multiphase flows and Fokker-Planck equations a reduction of the inherently huge computational time is highly relevant as well.

This goes beyond the application of an operator splitting approach. Depending on the application it might be justified to give up some generality. If an in-plane orientation can be expected, it is possible to assume a 2d instead of a 3d distribution. If the configuration is modeled on the surface of a sphere and the orientation distribution is symmetric because we do not distinguish between the beginning and the end of the fibers, the sphere can be reduced to a hemisphere. Then one has to think about the implementation of suitable periodic boundary conditions for the hemisphere, but ideally up to half of the computational time would be saved. Last but not least, the numerous possibilities of code optimization and parallelization on both the software and hardware level open up a wide range of options.

If the problem should be extended, geometries beyond the axisymmetric contraction benchmark are worth to be examined. Moreover, among the open questions is the numerical treatment of free surfaces in injection molding simulations.

9 Bibliography

[ADS+24] Nazih A Al Ayoubi, Hugues Digonnet, Luisa R Da Silva, Christophe Binetruy, Thierry Renault, and Sebastien Comas-Cardona. Simulation of the fiber orientation through a finite element approach to solve the Fokker-Planck Equation. *Journal of Non-Newtonian Fluid Mechanics*, 331:105284, 2024.

[Ali16] Ramzan Ali. *Numerical techniques for the simulation of PDEs on surfaces for biomathematical problems*. PhD thesis, TU Dortmund, 2016.

[AT87] Suresh G Advani and Charles L Tucker. The use of tensors to describe and predict fiber orientation in short fiber composites. *Journal of rheology*, 31(8):751–784, 1987.

[AT90] Suresh G Advani and Charles L Tucker. Closure approximations for three-dimensional structure tensors. *Journal of Rheology*, 34(3):367–386, 1990.

[Bär10] Christian Bär. *Elementare Differentialgeometrie*. De Gruyter, 2010.

[Bat70] GK Batchelor. Slender-body theory for particles of arbitrary cross-section in Stokes flow. *Journal of Fluid Mechanics*, 44(3):419–440, 1970.

[BB73] Jay P Boris and David L Book. Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works. *Journal of computational physics*, 11(1):38–69, 1973.

[BH82] Alexander N Brooks and Thomas JR Hughes. Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. *Computer methods in applied mechanics and engineering*, 32(1-3):199–259, 1982.

[BHM00] William L Briggs, Van Emden Henson, and Steve F McCormick. *A multigrid tutorial*. SIAM, 2000.

[BJK16] Gabriel R Barrenechea, Volker John, and Petr Knobloch. Analysis of algebraic flux correction schemes. *SIAM Journal on Numerical Analysis*, 54(4):2427–2451, 2016.

[BJKR18] Gabriel R Barrenechea, Volker John, Petr Knobloch, and Richard Rankin. A unified analysis of algebraic flux correction schemes for convection–diffusion equations. *SeMA Journal*, 75(4):655–685, 2018.

[BN83] JM Boland and Roy A Nicolaides. Stability of finite elements under divergence constraints. *SIAM Journal on Numerical Analysis*, 20(4):722–731, 1983.

[Bra13] Dietrich Braess. *Finite Elemente: Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*. Springer, 2013.

[BSK19] Kevin Breuer, Markus Stommel, and Wolfgang Korte. Analysis and evaluation of fiber orientation reconstruction methods. *Journal of Composites Science*, 3(3):67, 2019.

[BT92a] Randy S Bay and Charles L Tucker. Fiber orientation in simple injection moldings. Part I: Theory and numerical methods. *Polymer composites*, 13(4):317–331, 1992.

[BT92b] Randy S Bay and Charles L Tucker. Fiber orientation in simple injection moldings. Part II: Experimental results. *Polymer composites*, 13(4):332–341, 1992.

[Cho68] Alexandre J Chorin. Numerical solution of the Navier-Stokes equations. *Mathematics of computation*, 22(104):745–762, 1968.

[CK02] Du Hwan Chung and Tai Hun Kwon. Invariant-based optimal fitting closure approximation for the numerical prediction of flow-induced fiber orientation. *Journal of rheology*, 46(1):169–194, 2002.

[CL04a] Cédric Chauviere and Alexei Lozinski. Simulation of complex viscoelastic flows using the Fokker-Planck equation: 3D FENE model. *Journal of Non-Newtonian Fluid Mechanics*, 122(1-3):201–214, 2004.

[CL04b] Cédric Chauvière and Alexei Lozinski. Simulation of dilute polymer solutions using a Fokker-Planck equation. *Computers & fluids*, 33(5-6):687–696, 2004.

[CT95] Joaquim S Cintra and Charles L Tucker. Orthotropic closure approximations for flow-induced fiber orientation. *Journal of Rheology*, 39(6):1095–1122, 1995.

[DA84] Steven M Dinh and Robert C Armstrong. A rheological equation of state for semiconcentrated fiber suspensions. *Journal of Rheology*, 28(3):207–227, 1984.

[DE07] Gerhard Dziuk and Charles M Elliott. Surface finite elements for parabolic equations. *Journal of Computational Mathematics*, pages 385–407, 2007.

[DE08] Gerhard Dziuk and Charles M Elliott. Eulerian finite element method for parabolic PDEs on implicit surfaces. *Interfaces and Free Boundaries*, 10(1):119–138, 2008.

[DE13] Gerhard Dziuk and Charles M Elliott. Finite element methods for surface PDEs. *Acta Numerica*, 22:289–396, 2013.

[DH03] Jean Donea and Antonio Huerta. *Finite element methods for flow problems*. John Wiley & Sons, 2003.

[DH22] Sina Dahm and Christiane Helzel. Hyperbolic systems of moment equations describing sedimentation in suspensions of rod-like particles. *Multiscale Modeling & Simulation*, 20(3):1002–1039, 2022.

[Don18] Daniel Donner. Gradientenrekonstruktionstechniken zur Stabilisierung konvektiver Terme bei stetigen Galerkin-Diskretisierungen hoher Ordnung. Master's thesis, TU Dortmund, 2018.

[DP99] Donald A Drew and Stephen L Passman. *Theory of multicomponent fluids*. New York, Springer, 1999.

[DR06] Wolfgang Dahmen and Arnold Reusken. *Numerik für Ingenieure und Naturwissenschaftler*. Springer-Verlag, 2006.

[Dre83] Donald A Drew. Mathematical modeling of two-phase flow. *Annual review of fluid mechanics*, 15(1):261–291, 1983.

[DV99] F Dupret and V Verleye. Modelling the flow of fiber suspensions in narrow gaps. In *Rheology Series*, volume 8, pages 1347–1398. Elsevier, 1999.

[Dzi88] Gerhard Dziuk. *Finite elements for the Beltrami operator on arbitrary surfaces*, volume 1357 of Lecture Notes in Mathematics, pages 142–155. Springer, 1988.

[ES10] Charles M Elliott and Björn Stinner. Modeling and computation of two phase geometric biomembranes using surface finite elements. *Journal of Computational Physics*, 229(18):6585–6612, 2010.

[Ess22] Maximilian Esser. A local multilevel pressure schur complement solver for the incompressible Navier-Stokes equations with a tensor-valued diffusion. Master's thesis, TU Dortmund, 2022.

[ESW14] Howard C Elman, David J Silvester, and Andrew J Wathen. *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*. Oxford university press, 2014.

[Eva10] Lawrence C Evans. *Partial Differential Equations*, volume 19. American Mathematical Soc., 2010.

[Fan85] Xi-Jun Fan. Viscosity, first normal-stress coefficient, and molecular stretching in dilute polymer solutions. *Journal of Non-Newtonian Fluid Mechanics*, 17(2):125–144, 1985.

[Fan89] Xi-Jun Fan. Molecular models and flow calculations: II. Simulation of steady planar flow. *Acta Mechanica Sinica*, 5(3):216–226, 1989.

[FCL98] J Feng, CV Chaubal, and LG Leal. Closure approximations for the Doi theory: Which to use in simulating complex flows of liquid-crystalline polymers? *Journal of Rheology*, 42(5):1095–1119, 1998.

[FHH+08] J Férec, M Heniche, MC Heuzey, Gilles Ausias, and PJ Carreau. Numerical solution of the Fokker–Planck equation for fiber suspensions: application to the Folgar-Tucker-Lipscomb model. *Journal of Non-Newtonian Fluid Mechanics*, 155(1-2):20–29, 2008.

[FMAA20] Julien Ferec, Dihya Mezi, Suresh G Advani, and Gilles Ausias. Axisymmetric flow simulations of fiber suspensions as described by 3d probability distribution function. *Journal of Non-Newtonian Fluid Mechanics*, 284:104367, 2020.

[Fok14] Adriaan D Fokker. Die mittlere Energie rotierender elektrischer Dipole im Strahlungsfeld. *Annalen der Physik*, 348(5):810–820, 1914.

[Fol82] MJ Folkes. *Short fibre reinforced thermoplastics*. Research Studies Press, 1982.

[For12] Otto Forster. *Analysis 3*. Springer, 2012.

[For13] Otto Forster. *Analysis 2*. Springer, 2013.

[FT84] Fransisco Folgar and Charles L Tucker. Orientation behavior of fibers in concentrated suspensions. *Journal of reinforced plastics and composites*, 3(2):98–119, 1984.

[FVDC92] H Henry De Frahan, V Verleye, François Dupret, and MJ Crochet. Numerical prediction of fiber orientation in injection molding. *Polymer Engineering & Science*, 32(4):254–266, 1992.

[GGOS20] Umesh Gandhi, Sebastian Goris, Tim A Osswald, and Yu-Yang Song. *Discontinuous fiber-reinforced composites: fundamentals and applications*. Carl Hanser Verlag GmbH Co KG, 2020.

[GKS11] Sigal Gottlieb, David I Ketcheson, and Chi-Wang Shu. *Strong stability preserving Runge-Kutta and multistep time discretizations*. World Scientific, 2011.

[Glo03] Roland Glowinski. Finite element methods for incompressible viscous flow. *Handbook of Numerical Analysis*, 9:3–1176, 2003.

[GNPT18] Jean-Luc Guermond, Murtazo Nazarov, Bojan Popov, and Ignacio Tomas. Second-order invariant domain preserving approximation of the Euler equations using convex limiting. *SIAM Journal on Scientific Computing*, 40(5):A3211–A3239, 2018.

[GP16] Jean-Luc Guermond and Bojan Popov. Invariant domains and first-order continuous finite element approximation for hyperbolic systems. *SIAM Journal on Numerical Analysis*, 54(4):2466–2489, 2016.

[GPT19] Jean-Luc Guermond, Bojan Popov, and Ignacio Tomas. Invariant domain preserving discretization-independent schemes and convex limiting for hyperbolic systems. *Computer Methods in Applied Mechanics and Engineering*, 347:143–175, 2019.

[Gri09] Daniel Grieser. Der Laplace-Operator auf einer Riemannschen Mannigfaltigkeit. Technical report, Universität Oldenburg, 2009.

[GS98] Philip M Gresho and Robert L Sani. *Incompressible flow and the finite element method, Volume 1: Advection-diffusion and isothermal laminar flow*. Wiley, 1998.

[GT13] Sashikumaar Ganesan and Lutz Tobiska. Operator-splitting finite element algorithms for computations of high-dimensional parabolic problems. *Applied Mathematics and Computation*, 219(11):6182–6196, 2013.

[Hac13] Wolfgang Hackbusch. *Multigrid methods and applications*, volume 4. Springer Science & Business Media, 2013.

[Haj22] Hennes Hajduk. *Algebraically constrained finite element methods for hyperbolic problems with applications in geophysics and gas dynamics*. PhD thesis, TU Dortmund, 2022.

[Han62] George L Hand. A theory of anisotropic fluids. *Journal of Fluid Mechanics*, 13(1):33–46, 1962.

[Han94] Peter Hansbo. Aspects of conservation in finite element flow computations. *Computer methods in applied mechanics and engineering*, 117(3-4):423–437, 1994.

[Har84] Ami Harten. On a class of high resolution total-variation-stable finite-difference schemes. *SIAM Journal on Numerical Analysis*, 21(1):1–23, 1984.

[Hig05] Nicholas J Higham. The scaling and squaring method for the matrix exponential revisited. *SIAM Journal on Matrix Analysis and Applications*, 26(4):1179–1193, 2005.

[Hin16] Jochen Hinz. Isogeometric analysis of a reaction-diffusion-model for human brain development. Master's thesis, TU Delft, 2016.

[HL76] EJ Hinch and LG Leal. Constitutive equations in suspension mechanics. part 2. approximate forms for a suspension of rigid particles affected by brownian rotations. *Journal of Fluid Mechanics*, 76(1):187–208, 1976.

[HLHN11] J Hämäläinen, Stefan B Lindström, T Hämäläinen, and H Niskanen. Papermaking fibre-suspension flow simulations at multiple scales. *Journal of engineering mathematics*, 71:55–79, 2011.

[HO06] Christiane Helzel and Felix Otto. Multiscale simulations for suspensions of rod-like molecules. *Journal of Computational Physics*, 216(1):52–75, 2006.

[HV03] Willem H Hundsdorfer and Jan G Verwer. *Numerical solution of time-dependent advection-diffusion-reaction equations*, volume 33. Springer, 2003.

[JA21] Abhinav Jha and Naveed Ahmed. Analysis of flux corrected transport schemes for evolutionary convection-diffusion-reaction equations. *preprint arXiv:2103.04776*, 2021.

[Jef22] George B Jeffery. The motion of ellipsoidal particles immersed in a viscous fluid. *Proceedings of the Royal Society of London. Series A, Containing papers of a mathematical and physical character*, 102(715):161–179, 1922.

[Joh16] Volker John. *Finite element methods for incompressible flow problems*, volume 51. Springer, 2016.

[KA13] Peter Knabner and Lutz Angermann. *Numerik partieller Differentialgleichungen: eine anwendungsorientierte Einführung*. Springer, 2013.

[Kam13] Dimitrios Kamilis. Numerical methods for the PDEs on curves and surfaces. Master's thesis, Umeå University, 2013.

[Kan86] J van Kan. A second-order accurate pressure-correction scheme for viscous incompressible flow. *SIAM journal on scientific and statistical computing*, 7(3):870–891, 1986.

[Keu04] Roland Keunings. Micro-macro methods for the multiscale simulation of viscoelastic flow using molecular models of kinetic theory. *Rheology reviews*, pages 67–98, 2004.

[KH15] Dmitri Kuzmin and Jari Hämäläinen. Finite element methods for computational fluid dynamics: a practical guide. *SIAM Rev*, 57(4):642, 2015.

[KH23] Dmitri Kuzmin and Hennes Hajduk. *Property-preserving numerical schemes for conservation laws.* World Scientific, 2023.

[KLT05] Dmitri Kuzmin, Rainald Löhner, and Stefan Turek. *Flux-corrected transport: Principles, algorithms, and applications.* Springer, 2005.

[Kne06] David Knezevic. Finite element methods for deterministic simulation of polymeric fluids. Technical report, University of Oxford, 2006.

[Kne08] David Knezevic. *Analysis and implementation of numerical methods for simulating dilute polymeric fluids*. PhD thesis, University of Oxford, 2008.

[KOM09] Paul J Krochak, James A Olson, and D Mark Martinez. Fiber suspension flow in a tapered channel: The effect of flow/fiber coupling. *International journal of multiphase flow*, 35(7):676–688, 2009.

[KS09a] David J Knezevic and Endre Süli. A heterogeneous alternating-direction method for a micro-macro dilute polymeric fluid model. *ESAIM: Mathematical Modelling and Numerical Analysis*, 43(6):1117–1156, 2009.

[KS09b] David J Knezevic and Endre Süli. Spectral Galerkin approximation of Fokker-Planck equations with unbounded drift. *ESAIM: Mathematical Modelling and Numerical Analysis*, 43(3):445–485, 2009.

[Kuz10] Dmitri Kuzmin. *A guide to numerical methods for transport equations*. FAU Erlangen-Nürnberg, 2010.

[Kuz12] Dmitri Kuzmin. Algebraic flux correction I. In *Flux-corrected transport*, pages 145–192. Springer, 2012.

[Kuz14] Dmitri Kuzmin. An optimization-based approach to enforcing mass conservation in level set methods. *Journal of Computational and Applied Mathematics*, 258:78–86, 2014.

[Kuz18] Dmitri Kuzmin. Planar and orthotropic closures for orientation tensors in fiber suspension flow models. *SIAM Journal on Applied Mathematics*, 78(6):3040–3059, 2018.

[Kuz20] Dmitri Kuzmin. Monolithic convex limiting for continuous finite element discretizations of hyperbolic conservation laws. *Computer Methods in Applied Mechanics and Engineering*, 361:112804, 2020.

[LC03] Alexei Lozinski and Cédric Chauviere. A fast solver for Fokker-Planck equation applied to viscoelastic flows calculations: 2D FENE model. *Journal of Computational Physics*, 189(2):607–625, 2003.

[LDHB88] GG Lipscomb, Morton M Denn, DU Hur, and David V Boger. The flow of fiber suspensions in complex geometries. *Journal of Non-Newtonian Fluid Mechanics*, 26(3):297–325, 1988.

[LeV92] Randall J LeVeque. *Numerical methods for conservation laws*. Springer Basel AG, 1992.

[LeV96] Randall J LeVeque. High-resolution conservative algorithms for advection in incompressible flow. *SIAM Journal on Numerical Analysis*, 33(2):627–665, 1996.

[Lin08] Stefan Lindström. *Modelling and simulation of paper structure development*. PhD thesis, Mid Sweden University, 2008.

[LKSM17] Christoph Lohmann, Dmitri Kuzmin, John N Shadid, and Sibusiso Mabuza. Flux-corrected transport algorithms for continuous Galerkin methods based on high order Bernstein finite elements. *Journal of Computational Physics*, 344:151–186, 2017.

[Loh16a] Christoph Lohmann. Efficient algorithms for constraining orientation tensors in Galerkin methods for the Fokker–Planck equation. *Computers & Mathematics with Applications*, 71(5):1059–1073, 2016.

[Loh16b] Christoph Lohmann. *Galerkin-Spektralverfahren für die Fokker-Planck-Gleichung*. Springer, 2016.

[Loh19] Christoph Lohmann. *Physics-compatible finite element methods for scalar and tensorial advection problems*. Springer, 2019.

[LOP11] Alexei Lozinski, Robert G Owens, and Timothy N Phillips. The Langevin and Fokker-Planck equations in polymer rheology. In *Handbook of numerical analysis*, volume 16, pages 211–303. Elsevier, 2011.

[Loz03] Alexei Lozinski. *Spectral methods for kinetic theory models of viscoelastic fluids*. PhD thesis, 2003.

[LSPT12] Peter Hjort Lauritzen, William C Skamarock, MJ Prather, and MA Taylor. A standard test case suite for two-dimensional linear transport on the sphere. *Geoscientific Model Development*, 5(3):887–901, 2012.

[LT05] Stig Larsson and Vidar Thomée. *Partielle Differentialgleichungen und numerische Methoden*. Springer, 2005.

[LU07] Stefan B Lindström and Tetsu Uesaka. Simulation of the motion of flexible fibers in viscous fluid flow. *Physics of fluids*, 19(11):113307, 2007.

[LV10] Christoph Landsberg and Axel Voigt. A multigrid finite element method for reaction-diffusion systems on surfaces. *Computing and visualization in science*, 13(4):177–185, 2010.

[MB05] Arif Masud and Lawrence A Bergman. Application of multi-scale finite element methods to the solution of the Fokker–Planck equation. *Computer Methods in Applied Mechanics and Engineering*, 194(12-16):1513–1526, 2005.

[Mei15] Andreas Meister. *Numerik linearer Gleichungssysteme*. Springer, 2015.

[MSHJS11] Stephen Montgomery-Smith, Wei He, David A Jack, and Douglas E Smith. Exact tensor closures for the three-dimensional Jeffery's equation. *Journal of Fluid Mechanics*, 680:321–335, 2011.

[MSJS11] Stephen Montgomery-Smith, David Jack, and Douglas E Smith. The fast exact closure for Jeffery's equation with diffusion. *Journal of Non-Newtonian Fluid Mechanics*, 166(7-8):343–353, 2011.

[NKA20] Alexei Novikov, Dmitri Kuzmin, and Omid Ahmadi. Random walk methods for Monte Carlo simulations of Brownian diffusion on a sphere. *Applied Mathematics and Computation*, 364:124670, 2020.

[Nol13] Wolfgang Nolting. *Grundkurs Theoretische Physik 3*. Springer, 2013.

[NTL05] Ramachandran D Nair, Stephen J Thomas, and Richard D Loft. A discontinuous Galerkin transport scheme on the cubed sphere. *Monthly Weather Review*, 133(4):814–828, 2005.

[OF01] Stanley Osher and Ronald P Fedkiw. Level set methods: an overview and some recent results. *Journal of Computational physics*, 169(2):463–502, 2001.

[OFCH04] James A Olson, Ian Frigaard, Candice Chan, and Jari P Hämäläinen. Modeling a turbulent fibre suspension flowing in a planar contraction: The one-dimensional headbox. *International Journal of Multiphase Flow*, 30(1):51–66, 2004.

[OP02] Robert G Owens and Timothy N Phillips. *Computational rheology*. World Scientific, 2002.

[Ött12] Hans C Öttinger. *Stochastic processes in polymeric fluids: tools and examples for developing simulation algorithms*. Springer Science & Business Media, 2012.

[PBR13] Kara Peterson, Pavel Bochev, and Denis Ridzal. Optimization-based conservative transport on the cubed-sphere grid. In *International Conference on Large-Scale Scientific Computing*, pages 205–212. Springer, 2013.

[Pla17] Max Planck. Über einen Satz der statistischen Dynamik und seine Erweiterung in der Quantentheorie. *Sitzungsbericht Preuss. Akad. Wiss*, 24, 1917.

[PR55] Donald W Peaceman and Henry H Rachford, Jr. The numerical solution of parabolic and elliptic differential equations. *Journal of the Society for industrial and Applied Mathematics*, 3(1):28–41, 1955.

[PT09] Jay H Phelps and Charles L Tucker. An anisotropic rotary diffusion model for fiber orientation in short-and long-fiber thermoplastics. *Journal of Non-Newtonian Fluid Mechanics*, 156(3):165–176, 2009.

[QdLK22] Manuel Quezada de Luna and David I Ketcheson. Maximum principle preserving space and time flux limiting for diagonally implicit runge–kutta discretizations of scalar convection-diffusion equations. *Journal of Scientific Computing*, 92(3):102, 2022.

[Qua93] L Quartapelle. *Numerical solution of the incompressible Navier-Stokes equations*, volume 113. Springer Science & Business Media, 1993.

[QV08] Alfio Quarteroni and Alberto Valli. *Numerical approximation of partial differential equations*, volume 23. Springer Science & Business Media, 2008.

[Ran17a] Rolf Rannacher. *Numerik 1: Numerik gewöhnlicher Differentialgleichungen*. Heidelberg University Publishing, 2017.

[Ran17b] Rolf Rannacher. *Numerik 2: Numerik partieller Differentialgleichungen*. Heidelberg University Publishing, 2017.

[Ran17c] Rolf Rannacher. *Numerik 3: Probleme der Kontinuumsmechanik und ihre numerische Behandlung*. Heidelberg University Publishing, 2017.

[Ris96] Hannes Risken. *The Fokker-Planck Equation*. Springer, 1996.

[Sch13] Ben Schweizer. *Partielle Differentialgleichungen*. Springer, 2013.

[SJA02] Jason KC Suen, Yong Lak Joo, and Robert C Armstrong. Molecular orientation effects in viscoelasticity. *Annual review of fluid mechanics*, 34(1):417–444, 2002.

[SO88] Chi-Wang Shu and Stanley Osher. Efficient implementation of essentially non-oscillatory shock-capturing schemes. *Journal of computational physics*, 77(2):439–471, 1988.

[SS03] James A Sethian and Peter Smereka. Level set methods for fluid interfaces. *Annual review of fluid mechanics*, 35(1):341–372, 2003.

[SSK18] Markus Stommel, Marcus Stojek, and Wolfgang Korte. *FEM zur Berechnung von Kunststoff-und Elastomerbauteilen*. Carl Hanser Verlag GmbH Co KG, 2018.

[Ste84] Rolf Stenberg. Analysis of mixed finite elements methods for the Stokes problem: a unified approach. *Mathematics of computation*, 42(165):9–23, 1984.

[TOS00] Ulrich Trottenberg, Cornelius W Oosterlee, and Anton Schuller. *Multigrid*. Elsevier, 2000.

[TP01] Vladimir Tulovsky and Lech Papiez. Formula for the fundamental solution of the heat equation on the sphere. *Applied mathematics letters*, 14(7):881–884, 2001.

[Tuc91] Charles L Tucker. Flow regimes for fiber suspensions in narrow gaps. *Journal of Non-Newtonian Fluid Mechanics*, 39(3):239–268, 1991.

[Tuc22] Charles L Tucker. *Fundamentals of fiber orientation: Description, measurement and prediction*. Carl Hanser Verlag GmbH Co KG, 2022.

[Tur99] Stefan Turek. *Efficient solvers for incompressible flow problems: an algorithmic and computational approach*. Springer, 1999.

[Van86] S Pratap Vanka. Block-implicit multigrid solution of Navier-Stokes equations in primitive variables. *Journal of Computational Physics*, 65(1):138–158, 1986.

[VCD94] V Verleye, A Couniot, and F Dupret. Numerical prediction of fibre orientation in complex injection-moulded parts. *WIT Transactions on Engineering Sciences*, 4, 1994.

[Ver98] Brent E VerWeyst. *Numerical predictions of flow-induced fiber orientation in three-dimensional geometries.* PhD thesis, University of Illinois, 1998.

[VT02] Brent E VerWeyst and Charles L Tucker. Fiber suspensions in complex geometries: Flow/orientation coupling. *The Canadian Journal of Chemical Engineering*, 80(6):1093–1106, 2002.

[WDH+92] David L Williamson, John B Drake, James J Hack, Rüdiger Jakob, and Paul N Swarztrauber. A standard test set for numerical approximations to the shallow water equations in spherical geometry. *Journal of computational physics*, 102(1):211–224, 1992.

[WJ72] Harold R Warner Jr. Kinetic theory and rheology of dilute suspensions of finitely extendible dumbbells. *Industrial & Engineering Chemistry Fundamentals*, 11(3):379–387, 1972.

[WKT24] Katharina Wegener, Dmitri Kuzmin, and Stefan Turek. Efficient numerical solution of the Fokker–Planck equation using physics–conforming finite element methods. *Journal of Numerical Mathematics*, 32(3):217–232, 2024.

[Zal79] Steven T Zalesak. Fully multidimensional flux-corrected transport algorithms for fluids. *Journal of computational physics*, 31(3):335–362, 1979.

[ZMLD+12] Evgeniy Zharovsky, Amin Moosaie, Anne Le Duc, Michael Manhart, and Bernd Simeon. On the numerical solution of a convection-diffusion equation for particle orientation dynamics on geodesic grids. *Applied Numerical Mathematics*, 62(10):1554–1566, 2012.

10 Appendix

10.1 Spherical Coordinates

Basics. The need for spherical coordinates comes from the fact that we are working with the sphere $\mathbb{S}^2$. Since the $\mathbb{S}^2$ is a unit sphere, its radius is $r = 1$. We set

a) the polar angle $\theta \in [0, \pi]$ and
b) the azimuthal angle $\varphi \in [0, 2\pi)$.

Figure 10.1 visualizes the arrangement of those angles. For the relationship between the Cartesian and the spherical coordinates it holds true that

$$\begin{aligned} \sin(\theta) = a \quad (\star), \qquad & z = \cos(\theta), \\ \sin(\varphi) = \frac{y}{a} \qquad & \overset{(\star)}{\Longrightarrow} y = \sin(\theta)\sin(\varphi), \\ \cos(\varphi) = \frac{x}{a} \qquad & \overset{(\star)}{\Longrightarrow} x = \sin(\theta)\cos(\varphi), \end{aligned}$$

and summing up

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \sin\theta\cos\varphi \\ \sin\theta\sin\varphi \\ \cos\theta \end{pmatrix}.$$

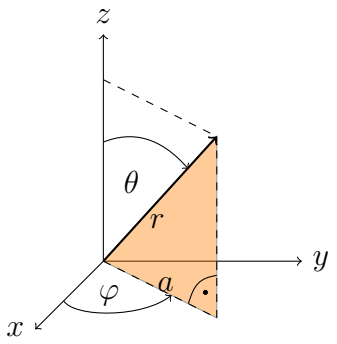


Figure 10.1: *Convention for the spherical coordinates*

In the context of manifolds, the above description for points on the sphere can also be written as parametrization $\tau : \mathbb{R}^2 \to \mathbb{R}^3$, where

$$\begin{aligned} \tau(\theta,\varphi) &= (\ \sin\theta\cos\varphi,\ \sin\theta\sin\varphi,\ \cos\theta\ )^T =: \mathbf{e}_r, \\ \frac{\partial\tau}{\partial\theta}(\theta,\varphi) &= (\ \cos\theta\cos\varphi,\ \cos\theta\sin\varphi,\ -\sin\theta\ )^T =: \mathbf{e}_\theta, \\ \frac{\partial\tau}{\partial\varphi}(\theta,\varphi) &= (-\sin\theta\sin\varphi,\ \sin\theta\cos\varphi,\ 0\ )^T =: \sin\theta\ \mathbf{e}_\varphi. \end{aligned}$$

Typically these three expressions in terms of τ are known as the orthonormal basis vectors $\mathbf{e}_r$, $\mathbf{e}_\theta$ and $\mathbf{e}_\varphi$, where it has to be ensured that $\mathbf{e}_\varphi$ is normalized as well.

Different conventions. The version of the spherical coordinates presented so far comes from physics and is by far the most commonly used convention. However, especially in American mathematics, the polar angle $\theta \in [0, \pi]$ often does not start at the north pole anymore but is replaced by $\tilde{\theta} \in [-\frac{\pi}{2}, \frac{\pi}{2}]$, which is zero in the xy-plane.

The new situation is shown in Figure 10.2. We obtain $\sin\tilde{\theta} = z$ and $\cos\tilde{\theta} = a$. Everything else remains unchanged, so that

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} \cos\tilde{\theta}\cos\varphi \\ \cos\tilde{\theta}\cos\varphi \\ \sin\tilde{\theta} \end{pmatrix}.$$

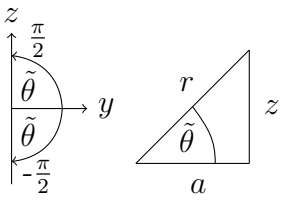


Figure 10.2: *Amercian convention with $\tilde{\theta}$.*

For a uniform setting in our numerical studies we sometimes had to convert from the American convention used in the corresponding literature to the physical convention. We compare the basis vectors as well as two flow fields used in Section 5.6.3.

Physical convention	American convention
$\mathbf{e}_r = \begin{pmatrix} \sin\theta\cos\varphi \\ \sin\theta\sin\varphi \\ \cos\theta \end{pmatrix}$	$\mathbf{e}_r = \begin{pmatrix} \cos\tilde{\theta}\cos\varphi \\ \cos\tilde{\theta}\sin\varphi \\ \sin\tilde{\theta} \end{pmatrix}$
$\mathbf{e}_\theta = \begin{pmatrix} \cos\theta\cos\varphi \\ \cos\theta\sin\varphi \\ -\sin\theta \end{pmatrix}$	$\mathbf{e}_{\tilde{\theta}} = \begin{pmatrix} -\sin\tilde{\theta}\cos\varphi \\ -\sin\tilde{\theta}\sin\varphi \\ \cos\tilde{\theta} \end{pmatrix}$
$\mathbf{e}_\varphi = \begin{pmatrix} -\sin\varphi \\ \cos\varphi \\ 0 \end{pmatrix}$	
Solid Body Rotation	
$u = 2\pi(\sin\theta\cos\alpha + \cos\theta\cos\varphi\sin\alpha)$	$u = 2\pi(\cos\tilde{\theta}\cos\alpha + \sin\tilde{\theta}\cos\varphi\sin\alpha)$
$v = -2\pi\sin\varphi\sin\alpha$	
Deformational Flow Field	
$\varphi' = \varphi - 2\pi t$	
$u = \beta\sin^2(\varphi')\sin(2\theta)\cos(\pi t) + 2\pi\sin(\theta)$	$u = \beta\sin^2(\varphi')\sin(2\tilde{\theta})\cos(\pi t) + 2\pi\cos(\tilde{\theta})$
$v = \beta\sin(2\varphi')\sin(\theta)\cos(\pi t)$	$v = \beta\sin(2\varphi')\cos(\tilde{\theta})\cos(\pi t)$

Table 10.3: *Comparison of physical an American convention for selected quantities.*

The term $\sin\theta$ becomes $\cos\tilde{\theta}$ and vice versa $\cos\theta$ becomes $\sin\tilde{\theta}$, since $\tilde{\theta} = \frac{\pi}{2} - \theta$. Furthermore, $\sin(2\tilde{\theta}) = \sin(\pi - 2\theta) = \sin(2\theta) = 2\sin(\theta)\cos(\theta)$.

The relationship between $\mathbf{e}_\theta$ and $\mathbf{e}_{\tilde{\theta}}$ deserves special attention. We calculate both vectors by deriving the respective $\mathbf{e}_r$ with respect to θ or $\tilde{\theta}$. Then, however, we have that $\mathbf{e}_\theta = -\mathbf{e}_{\tilde{\theta}}$. While on the sphere the single components of $\mathbf{e}_{\tilde{\theta}}$ are oriented from the South Pole to North Pole, for $\mathbf{e}_\theta$ it is the other way round. As a consequence, the velocity field that is described in American papers by $\mathbf{v} = u\mathbf{e}_\varphi + v\mathbf{e}_{\tilde{\theta}}$ needs to be changed to $\mathbf{v} = u\mathbf{e}_\varphi - v\mathbf{e}_\theta$ (!).

Some more formulas. In what follows, we stick to the physical convention used throughout this work.
With the definitions above, the Cartesian coordinates x, y and z are easily obtained from the angles θ and φ. Sometimes, however, the opposite direction is needed as well; for example, when the mesh nodes are given in Cartesian coordinates, but the

flow field is described by spherical coordinates, see Table 10.3. Then the expressions to convert read

$$\theta = \arccos(z),$$
$$\varphi = \operatorname{atan2}(y,x) = \begin{cases} \arctan(\frac{y}{x}) & \text{if } x > 0 \\ \frac{\pi}{2}\operatorname{sgn}(y) & \text{if } x = 0 \\ \arctan(\frac{y}{x}) + \pi & \text{if } x < 0 \wedge y \geq 0 \\ \arctan(\frac{y}{x}) - \pi & \text{if } x < 0 \wedge y < 0, \end{cases}$$
$$\varphi = \varphi + 2\pi, \quad \text{if } \varphi < 0.$$

Within the proof of Theorem 2.4 derivatives of the basis functions are needed. They can be easily calculated as [Qua93, A.4.1]

$$\begin{aligned} \frac{\partial \mathbf{e}_r}{\partial \theta} &= \mathbf{e}_\theta; & \frac{\partial \mathbf{e}_\theta}{\partial \theta} &= -\mathbf{e}_r; & \frac{\partial \mathbf{e}_\varphi}{\partial \theta} &= \mathbf{0}; \\ \frac{\partial \mathbf{e}_r}{\partial \varphi} &= \sin\theta\, \mathbf{e}_\varphi; & \frac{\partial \mathbf{e}_\theta}{\partial \varphi} &= \cos\theta\, \mathbf{e}_\varphi; & \frac{\partial \mathbf{e}_\varphi}{\partial \varphi} &= -\sin\theta\, \mathbf{e}_r - \cos\theta\, \mathbf{e}_\theta. \end{aligned}$$

Finally, we list the spherical gradient, divergence and Laplacian as they can be found in the literature [Qua93, A.4.2+3]:

$$\nabla_{\mathbf{p}} = \mathbf{e}_\theta \frac{\partial}{\partial \theta} + \mathbf{e}_\varphi \frac{1}{\sin\theta} \frac{\partial}{\partial \varphi} \tag{10.1}$$
$$\nabla_{\mathbf{p}} \cdot V = \begin{cases} \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta\, V_\theta) + \frac{\partial V_\varphi}{\partial \varphi} & \text{if } V = V_\theta \frac{\partial \tau}{\partial \theta} + V_\varphi \frac{\partial \tau}{\partial \varphi}, \\ \frac{1}{\sin\theta}\frac{\partial}{\partial\theta}(\sin\theta\, V_\theta) + \frac{1}{\sin\theta}\frac{\partial V_\varphi}{\partial \varphi} & \text{if } V = V_\theta \mathbf{e}_\theta + V_\varphi \mathbf{e}_\varphi. \end{cases} \tag{10.2}$$
$$\Delta_{\mathbf{p}} = \frac{1}{\sin\theta} \frac{\partial}{\partial \theta} \left(\sin\theta \frac{\partial}{\partial \theta} \right) + \frac{1}{\sin^2\theta} \frac{\partial^2}{\partial \varphi^2} \tag{10.3}$$

Remark 10.1 (Derivation of the divergence)**.** Taking formula (10.1) as given, the divergence can be calculated as the dot product of $\nabla_{\mathbf{p}}$ and a vector field V. Also using the above definitions of the basis functions we obtain

$$\begin{aligned} \nabla_{\mathbf{p}} \cdot V = & \left(\mathbf{e}_\theta \frac{\partial}{\partial \theta} + \mathbf{e}_\varphi \frac{1}{\sin\theta} \frac{\partial}{\partial \varphi} \right) \cdot (V_\theta \mathbf{e}_\theta + V_\varphi \sin\theta\, \mathbf{e}_\varphi) \\ = & \ \mathbf{e}_\theta \left(\frac{\partial V_\theta}{\partial \theta} \mathbf{e}_\theta + V_\theta \underbrace{\frac{\partial \mathbf{e}_\theta}{\partial \theta}}_{=-\mathbf{e}_r} + \frac{\partial (V_\varphi \sin\theta)}{\partial \theta} \mathbf{e}_\varphi + V_\varphi \sin\theta \underbrace{\frac{\partial \mathbf{e}_\varphi}{\partial \theta}}_{=0} \right) \\ & + \mathbf{e}_\varphi \frac{1}{\sin\theta} \left(\frac{\partial V_\theta}{\partial \varphi} \mathbf{e}_\theta + V_\theta \underbrace{\frac{\partial \mathbf{e}_\theta}{\partial \varphi}}_{\cos\theta\, \mathbf{e}_\varphi} + \frac{\partial (V_\varphi \sin\theta)}{\partial \varphi} \mathbf{e}_\varphi + V_\varphi \sin\theta \underbrace{\frac{\partial \mathbf{e}_\varphi}{\partial \varphi}}_{=c_1 \mathbf{e}_r + c_2 \mathbf{e}_\theta} \right) \\ = & \ \frac{\partial V_\theta}{\partial \theta} + \frac{\cos\theta}{\sin\theta} V_\theta + \frac{\partial V_\varphi}{\partial \varphi} \quad = \ \frac{1}{\sin\theta} \frac{\partial}{\partial \theta} (\sin\theta\, V_\theta) + \frac{\partial V_\varphi}{\partial \varphi}. \end{aligned}$$

Concrete calculations. To perform the numerical tests in the Sections 5.6.1 and 5.6.2, the spherical Laplacian, the surface gradient and where necessary the integral value for our two functions are needed. To calculate the gradient and the Laplacian, the formulas (5.14) and (5.15) can be used after a conversion from Cartesian to spherical coordinates. For the Laplacian, we can alternatively use the relationship between the Laplacian for $\mathbb{R}^n$ and for $\mathbb{S}^{n-1}$ as described by Lemma 5.18.

u=2z. Reformulating the function $u(z) = 2z$ to $u(\theta) = 2\cos\theta$, we have

$$\begin{aligned}\Delta_{\mathbf{p}} u &= \frac{\partial^2}{\partial\theta^2}(2\cos\theta) + \frac{\cos\theta}{\sin\theta}\frac{\partial}{\partial\theta}(2\cos\theta)\\ &= -2\cos\theta - 2\cos\theta\\ &= -4\cos\theta\\ &= -4z.\end{aligned}$$

We obtain the same result by applying the Cartesian Laplace operator, normalizing each argument of the function $u(x, y, z)$ by $\|\mathbf{x}\|_2 = \sqrt{x^2+y^2+z^2}$. The normalization is necessary to obtain the right coefficients even though we omit $\|\mathbf{x}\|_2$ after deriving since it is equal to one. Hence,

$$\begin{aligned}\Delta\left(\frac{2z}{\sqrt{x^2+y^2+z^2}}\right) &= \frac{\partial^2}{\partial x^2}\left(\frac{2z}{\sqrt{\dots}}\right) + \frac{\partial^2}{\partial y^2}\left(\frac{2z}{\sqrt{\dots}}\right) + \frac{\partial^2}{\partial z^2}\left(\frac{2z}{\sqrt{\dots}}\right)\\ &= -2z + 6x^2z - 2z + 6y^2z - 6z(x^2+y^2)\\ &= -4z.\end{aligned}$$

Applying the formula for the surface gradient to $u(\theta)$, we find that

$$\nabla_{\mathbf{p}} u = \mathbf{e}_\theta \frac{\partial}{\partial\theta}(2\cos\theta) = -2\begin{pmatrix}\cos\theta\cos\varphi\\ \cos\theta\sin\varphi\\ -\sin\theta\end{pmatrix}\sin\theta = \begin{pmatrix}-2xy\\ -2yz\\ 2(x^2+y^2)\end{pmatrix}.$$

Finally, applying the integration formula given in Remark 5.23 verifies that the integral over the sphere is zero:

$$\int_{\mathbb{S}^2} u\,\mathrm{d}\mathbf{p} = 2\int_0^{2\pi}\int_0^{\pi}\cos\theta\sin\theta\,\mathrm{d}\theta\,\mathrm{d}\varphi = 4\pi\int_0^{\pi}\cos\theta\,\sin\theta\,\mathrm{d}\theta = 0.$$

u=xy. Laplacian and gradient on the sphere have to be calculated for the second function $u = xy = \sin^2\theta\cos\varphi\sin\varphi$ as well. For the Laplacian we obtain either

$$\begin{aligned}\Delta_{\mathbf{p}} u &= \frac{\partial^2}{\partial\theta^2}u + \frac{\cos\theta}{\sin\theta}\frac{\partial}{\partial\theta}u + \frac{1}{\sin^2\theta}\frac{\partial^2}{\partial\varphi^2}u\\ &= 2(\cos^2\theta - \sin^2\theta)\cos\varphi\sin\varphi + 2\cos^2\theta\cos\varphi\sin\varphi - 4\cos\varphi\sin\varphi\\ &= -6\sin^2\theta\cos\varphi\sin\varphi\\ &= -6xy\end{aligned}$$

or

$$\Delta\left(\frac{xy}{x^2+y^2+z^2}\right) \;=\; \ldots \;=\; -\frac{6xy}{(x^2+y^2+z^2)^2} \;=\; -6xy.$$

For the gradient we have

$$\begin{aligned}\nabla_{\mathbf{p}} u \;&=\; \mathbf{e}_\theta \,\frac{\partial}{\partial\theta}\,(\sin^2\theta\,\cos\varphi\,\sin\varphi) \;+\; \mathbf{e}_\varphi\,\frac{1}{\sin\theta}\,\frac{\partial}{\partial\varphi}\,(\sin^2\theta\,\cos\varphi\,\sin\varphi)\\ &=\; 2\,\mathbf{e}_\theta\,\sin\theta\cos\theta\sin\varphi\cos\varphi \;-\; \mathbf{e}_\varphi\,\sin\theta\,(\cos^2\varphi-\sin^2\varphi).\end{aligned}$$

For the x-component this means that

$$\begin{aligned}&2\cos^2\theta\sin\theta\cos^2\varphi\sin\varphi \;-\; \sin\theta\sin\varphi\,(\cos^2\varphi-\sin^2\varphi)\\ &=(2\cos^2\theta\,\cos^2\varphi \;-\; \cos^2\varphi \;+\; \sin^2\varphi)\,\underbrace{\sin\theta\,\sin\varphi}_{=y}\\ &=((2\cos^2\theta-2+1)\cos^2\varphi+\sin^2\varphi)\,y\\ &=(\,-2\sin^2\theta\cos^2\varphi+1)\,y\\ &=(-x^2+y^2+z^2)\,y.\end{aligned}$$

In a similar vein the y-component can be written as

$$\begin{aligned}&2\cos^2\theta\sin\theta\cos\varphi\sin^2\varphi \;+\; \sin\theta\cos\varphi\,(\cos^2\varphi-\sin^2\varphi)\\ &=(2\cos^2\theta\sin^2\varphi+\cos^2\varphi-\sin^2\varphi)\,\underbrace{\sin\theta\cos\varphi}_{=x}\\ &=((2\cos^2\theta-2+1)\sin^2\varphi+\cos^2\varphi)\,x\\ &=(-2\sin^2\theta\sin^2\varphi+1)\,x\\ &=(x^2-y^2+z^2)\,x.\end{aligned}$$

The last equations are valid because

$$\begin{aligned}-x^2+y^2+z^2 &= -\sin^2\theta\cos^2\varphi+\sin^2\theta\sin^2\varphi+\cos^2\theta\\ &=\; \sin^2\theta\,(1-2\cos^2\varphi) \;+\; \cos^2\theta\\ &=-2\,\sin^2\theta\,\cos^2\varphi \;+\; 1.\end{aligned}$$

and

$$\begin{aligned}x^2-y^2+z^2 &= \sin^2\theta\cos^2\varphi-\sin^2\theta\,\sin^2\varphi+\cos^2\theta\\ &=\sin^2\theta\,(1-2\sin^2\varphi) \;+\cos^2\theta\\ &=\; -2\sin^2\theta\,\sin^2\varphi \;+\; 1.\end{aligned}$$

Finally, the z-component of the gradient can be rewritten as

$$-2\cos\theta\sin^2\theta\cos\varphi\sin\varphi+0 \;=\; -2xyz.$$

Consequently, the spherical gradient expressed in Cartesian coordinates reads

$$\nabla_{\mathbf{p}} u = \begin{pmatrix} y(-x^2+y^2+z^2)\\ x(\;\; x^2-y^2+z^2)\\ -2xyz \end{pmatrix}.$$

Remark 10.2 (Reusing code)**.** During implementation, it was possible to reuse code for the error calculation or to obtain (parts of) the right hand side. Note, however, that this only works if the relevant derivatives are explicitly provided in the background, not if we let the code automatically derive the functions as if we were in the Cartesian space.

10.2 The sphere as submanifold

A prime example of a submanifold is the $\mathbb{S}^2$, the surface of the unit sphere in $\mathbb{R}^3$. First of all, with $k = 2$ and $n = 3$ it provides an illustrative example for a submanifold. It is intuitive that small pieces of $\mathbb{S}^2$ can be identified with areas in $\mathbb{R}^2$. Moreover, the sphere is of interest because of its applications such as our earth. Last but not least, within this work the Fokker-Planck equation is partially defined on $\mathbb{S}^2$.

To get used to the definitions of a submanifold in Section 5.2, we consider them for the $\mathbb{S}^2$, which can be described as

$$M = \mathbb{S}^2 = \left\{ (X, Y, Z)^T \in \mathbb{R}^3 \;\middle|\; X^2 + Y^2 + Z^2 = 1 \right\}. \tag{10.4}$$

The variables for $\mathbb{S}^2 \subset \mathbb{R}^3$ are denoted by capital letters, the variables for $\Omega \subset \mathbb{R}^2$ by lowercase letters. First, we illustrate Definition 5.5. Let $U = \mathbb{R}^3$. With regard to expression (10.4) we choose $f_1(X, Y, Z) = X^2 + Y^2 + Z^2 - 1$. As Jacobian we obtain $Df_1 = 2(X\ Y\ Z)$, where the rank is $n - k = 1$. This already proves that the $\mathbb{S}^2$ is a submanifold.

The focus of Definition 5.3 is on the parametrization. Since the parametrization τ we are looking for is not unique, we introduce different possibilities to verify that $\mathbb{S}^2$ is a submanifold.

Hemisphere. As first example, [Bär10], let us consider

$$\Omega = \left\{ (x, y)^T \in \mathbb{R}^2 \;\middle|\; x^2 + y^2 < 1 \right\}, \quad U = \left\{ (X, Y, Z)^T \in \mathbb{R}^3 \;\middle|\; Z > 0 \right\},$$
$$\text{and} \quad \tau : \Omega \to M \cap U, \ \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \begin{pmatrix} x \\ y \\ \sqrt{1 - (x^2 + y^2)} \end{pmatrix} =: \begin{pmatrix} X \\ Y \\ Z \end{pmatrix},$$

where the third component of τ results from the formula $X^2 + Y^2 + Z^2 = 1$ for $\mathbb{S}^2$. As required by the definition, Ω and U are open sets of correct dimension. It holds true that $\tau(\Omega) = M \cap U$. Parametrization τ maps the interior of a unit circle onto a hemisphere, see Figure 10.4. It is intuitively clear that the geometries are homeomorphic, which includes that

$$\tau^{-1} : M \cap U \to \Omega : \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \mapsto \begin{pmatrix} X \\ Y \end{pmatrix} =: \begin{pmatrix} x \\ y \end{pmatrix}$$

as the inverse function to τ is continuous.

The Jacobian reads

$$D\tau = \begin{pmatrix} \frac{\partial \tau_1}{\partial x} & \frac{\partial \tau_1}{\partial y} \\ \frac{\partial \tau_2}{\partial x} & \frac{\partial \tau_2}{\partial y} \\ \frac{\partial \tau_3}{\partial x} & \frac{\partial \tau_3}{\partial y} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ \frac{\partial \tau_3}{\partial x} & \frac{\partial \tau_3}{\partial y} \end{pmatrix}.$$

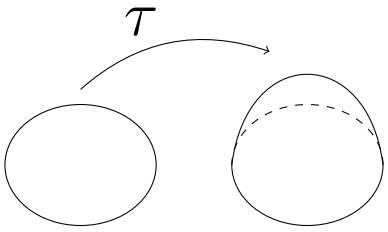


Figure 10.4: *Mapping to the hemisphere*

Because of the number of columns, $\text{rank}(D\tau) \leq 2$. Since the first two rows are linearly independent, in total $\text{rank}(D\tau) = 2$, which proves that τ is an immersion. By defining a set U for the cases $Z < 0, X < 0, X > 0, Y < 0$ and $Y > 0$ as well, every point on the sphere is covered at least one time, and with these six charts it is shown that $\mathbb{S}^2$ is a submanifold. It is possible to reduce the number of charts to two, as shown in the following example.

Stereographic projection. Our second parametrization is based on the stereographic projection [Bär10]. With this projection each point $P = (X, Y, Z)^T$ of the sphere, except the north pole $N = (0, 0, 1)^T$, is projected to a point $p = (x, y)^T$ in the $\mathbb{R}^2$-plane as visualized in Figure 10.5, where $g(t)$ symbolizes the projection.

The line $g(t)$ can be described mathematically as linear combination of N and P, that is,

$$g(t) = tN + (1-t)P = t \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} + (1-t) \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}.$$

Since the third component disappears in the projection plane, we find that

$$t + (1-t)Z \stackrel{!}{=} 0 \quad \Longrightarrow \quad 1 - t = \frac{1}{1-Z}.$$

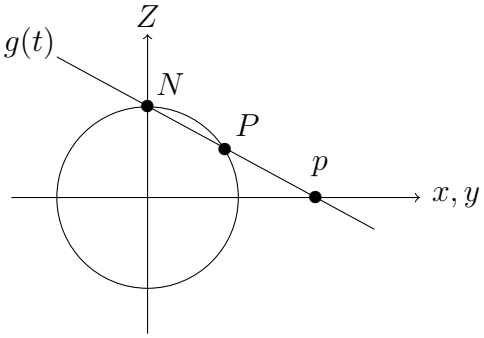


Figure 10.5: *Stereographic projection through N from P to p.*

Applying this to the first and the second component of $g(t)$, the projection to p reads

$$\tau^{-1}: \quad \mathbb{S}^2 \backslash (0,0,1)^T \to \mathbb{R}^2, \quad \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \mapsto \frac{1}{1-Z} \begin{pmatrix} X \\ Y \end{pmatrix} =: \begin{pmatrix} x \\ y \end{pmatrix}$$

The inverse function corresponding to τ^{-1} is given by

$$\tau: \quad \mathbb{R}^2 \to \mathbb{S}^2 \backslash (0,0,1)^T, \quad \begin{pmatrix} x \\ y \end{pmatrix} \mapsto \frac{1}{x^2+y^2+1} \begin{pmatrix} 2x \\ 2y \\ x^2+y^2-1 \end{pmatrix} =: \begin{pmatrix} X \\ Y \\ Z \end{pmatrix}$$

as it can be verified through

$$\tau\left(\tau^{-1}\begin{pmatrix} X \\ Y \\ Z \end{pmatrix}\right) = \tau\left(\frac{1}{1-Z}\begin{pmatrix} X \\ Y \end{pmatrix}\right) = \tau\left(\frac{1}{1-\frac{x^2+y^2-1}{x^2+y^2+1}} \frac{2}{x^2+y^2-1} \begin{pmatrix} x \\ y \end{pmatrix}\right)$$

$$= \tau\left(\frac{x^2+y^2+1}{2}\frac{2}{x^2+y^2-1}\begin{pmatrix}x\\y\end{pmatrix}\right) = \tau\left(\begin{pmatrix}x\\y\end{pmatrix}\right) = \begin{pmatrix}X\\Y\\Z\end{pmatrix},$$

and similarly

$$\begin{aligned}\tau^{-1}\left(\tau\begin{pmatrix}x\\y\end{pmatrix}\right) &= \tau^{-1}\left(\frac{1}{x^2+y^2+1}\begin{pmatrix}2x\\2y\\x^2+y^2-1\end{pmatrix}\right) = \tau^{-1}\left(\frac{1}{\frac{X^2+Y^2}{(1-Z)^2}+1}\begin{pmatrix}\frac{2X}{1-Z}\\\frac{2Y}{1-Z}\\\frac{X^2+Y^2}{(1-Z^2)-1}\end{pmatrix}\right)\\ &= \tau^{-1}\left(\frac{1}{X^2+Y^2+(1-Z)^2}\begin{pmatrix}2X(1-Z)\\2Y(1-Z)\\X^2+Y^2-(1-Z)^2\end{pmatrix}\right)\\ &\overset{(*)}{=} \tau^{-1}\left(\frac{1}{2(1-Z)}\begin{pmatrix}2X(1-Z)\\2Y(1-Z)\\2Z(1-Z)\end{pmatrix}\right) \quad = \tau^{-1}\begin{pmatrix}X\\Y\\Z\end{pmatrix} \quad = \begin{pmatrix}x\\y\end{pmatrix},\end{aligned}$$

where $X^2+Y^2+Z^2=1$ was used for $(*)$. Both $\Omega=\mathbb{R}^2$ and $U=\mathbb{R}^3\backslash(0,0,1)^T$ are open sets; and as composition of continuous functions, τ and τ^{-1} are continuous. Let us finally consider

$$D\tau = \begin{pmatrix}\frac{2(-x^2+y^2+1)}{(x^2+y^2+z^2)^2} & -\frac{4xy}{(x^2+y^2+1)^2}\\ \frac{-4xy}{(x^2+y^2+1)^2} & \frac{2(x^2-y^2+1)}{(x^2+y^2+1)^2}\\ \frac{\partial\tau_3}{\partial x} & \frac{\partial\tau_3}{\partial y}\end{pmatrix}.$$

It is easily verified that the first two rows are linearly independent, that is,

$$\alpha(\nabla\tau_1)+\beta(\nabla\tau_2)=\mathbf{0} \quad \Longrightarrow \alpha=\beta=0.$$

The quadratic terms cannot be canceled out by the (bi)linear terms. Setting $x=y$ or even $x=y=0$ does not change the situation. Consequently, $\nabla\tau$ has rank 2 and τ is an immersion. A second chart is needed to cover the north pole. An obvious choice is the stereographic projection from the south pole.

Spherical coordinates. Spherical coordinates are the most common way to parameterize the $\mathbb{S}^2$. They were used repeatedly in the thesis. Relevant formulas are summarized in Appendix 10.1. Here we consider

$$\begin{aligned}\Omega &= \left\{(\theta,\varphi)\in\mathbb{R}^2 \;\middle|\; \theta\in(0,\pi),\varphi\in(0,2\pi)\right\};\\ U &= \mathbb{R}^3\backslash\left\{(0,0,1)^T,(0,0,-1)^T, \text{longitude described by } \varphi=0\right\};\\ \tau:\ \Omega\to M\cap U,&\quad \begin{pmatrix}\theta\\\varphi\end{pmatrix}\mapsto\begin{pmatrix}\sin\theta\cos\varphi\\\sin\theta\sin\varphi\\\cos\theta\end{pmatrix};\\ D\tau &= \begin{pmatrix}\frac{\partial\tau}{\partial\theta} & \frac{\partial\tau}{\partial\varphi}\end{pmatrix} = \begin{pmatrix}\cos\theta\cos\varphi & -\sin\theta\sin\varphi\\ \cos\theta\sin\varphi & \sin\theta\cos\varphi\\ -\sin\theta & 0\end{pmatrix}.\end{aligned}$$

We choose open intervals for Ω. By removing the interval boundaries, north pole, south pole and prime meridian are not covered. However, Ω has to be an open set by definition. The restriction to the respective intervals is important for τ to be bijective, which is a prerequisite for τ to be a homeomorphism. With this restriction to the intervals the inverse function τ^{-1} is also continuous. To prove that τ is an immersion, we show that $\operatorname{rank}(D\tau) = 2$, where $\operatorname{rank}(D\tau) \leq 2$ is clear because of the number of columns. Since $\theta \in (0, \pi)$, it is always true that $\sin(\theta) = 0$. Furthermore, $\sin(\varphi)$ and $\cos(\varphi)$ never become zero at the same time. In consequence, always either the second and the third or the first and the third line of $D\tau$ are linearly independent, which finishes the proof.

10.3 Aspects of implementation

The key to make finite element theory fruitful is implementation. In Chapter 3 about FEM, we did not cover the implementation. On the one hand, much literature can be found on the subject, see, e.g., [KA13]. On the other hand, a suitable finite element software package was already available. However, a basic understanding is important in any case.

FEAT3. There is a wide variety of commercial or open source FE software. For this work, LS3's own software package FEAT3 written in C++11 was used. The development of its predecessors started in the 1980s, the origin of FEAT3 itself goes back to 2010. It provides different components, see Figure 10.6, which can be combined according to the modular design principle to solve several PDEs for varying applications. In particular, FEAT3 features hardware-related fast solvers, multigrid and different possibilities for parallelization.

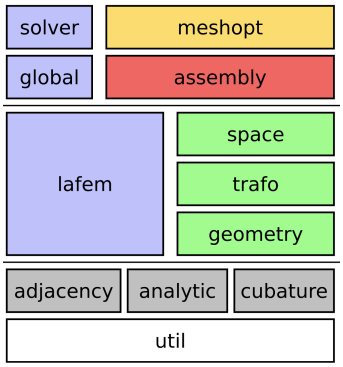


Figure 10.6: *kernel structure (by Peter Zajac)*

Mesh. During the phase of pre-processing, mesh generation has to be faced. In the case of more complex geometries, external software might be used for this task. The main computational issue is the keeping track of the local and global indices of the nodes and the elements [ESW14]. The numbering does not only describe the adjacency of the mesh but it is also necessary to place the respective contributions correctly in the FE matrices and the right hand side vector for the algebraic system.

Assembly. In Section 3.4, mathematical properties of the FE matrices were investigated. The structure of the matrix plays an important role. The specific FE matrix is defined by the bilinear form, the chosen space, e.g., $\mathbf{P}_1$ or $\mathbf{Q}_1$, and the formula for numerical integration. This formula must be of sufficiently high order to guarantee the optimal order of convergence.

CSR format. The structure and the sparsity pattern of an FE matrix have a huge impact on the functionality of the applied numerical methods. FEAT3 uses the compressed sparse row (CSR) format, which is a common and beneficial way to store sparse matrices. It is based on three vectors. In the first vector 'val' the non-zero entries of the matrix are stored row by row, while the second vector 'col_ind' contains the corresponding column indices. Accordingly, the length of both vectors is equal to the number of the non-zero entries. Finally, the vector 'row_ptr' determines, which value belongs to which row by indicating after how many non-zero entries a new row begins. This counting includes all the elements from previous rows. For technical reasons the third vector always starts with a zero, so that its length is one more than the number of rows. Every matrix can be unambiguously described by these vectors. As an example consider

$$A = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 5 & 8 & 0 & 0 \\ 0 & 0 & 3 & 0 \\ 0 & 6 & 0 & 0 \end{pmatrix},$$

where val = (5 8 3 6), col_ind = (0 1 2 1) and row_ptr = (0 0 2 3 4).

Sometimes we have to loop through the elements of a matrix. In the case of a sparse $n \times m$-matrix the simple approach 'for i = 1:n { for j = 1:m {...}}' that loops over each individual element is utterly inefficient and leads to a prohibitively large execution time. Instead, using the CSR format we only loop over the non-zero entries, as described in Figure 10.7.

```
for(Index i=0; i<n; ++i)
{
  const Index start = row_ptr[i];
  const Index end = row_ptr[i+1];
  for (Index k=start; k<end; ++k)
  {
        Index j=col_ind[k];
    (...)
  }
}
```

Figure 10.7: *Loop over a CSR matrix.*

Node-based vs. element-based implementation. In the context of finite elements we distinguish between node- and element-based implementations. Each row of an FE matrix belongs to a node identified by an index. When using the node-based implementation, the associated elements must be determined for each node. While this approach is often used for model problems to get familiar with FEM, in practice it consumes unnecessary computational time [Bra13].
Software packages usually employ the element-based approach, where we sequentially pass through the elements. For each element the nodal data is 'gathered', processed and saved in local matrices, that are finally inserted in the global matrix, when the results are 'scattered' back to the nodes [Kuz10].

Last but not least, to avoid confusion, we mention that in the context of limiting there is also a distinction between 'edge-based' and 'element-based' algorithms. In this thesis, we stuck to so-called edge-based methods throughout, since they have the advantage to work in a black box manner. Element-based methods promise several advantages but also require a deeper knowledge of the underlying data structures since they need access to the element matrices, see [KH23].