Numerical Methods for the Simulation of Taxis–Diffusion–Reaction Systems

Dissertation

zur Erlangung des akademischen Grades

doctor rerum naturalium (Dr. rer. nat.)

vorgelegt der

Mathematisch–Naturwissenschaftlich–Technischen Fakultät
(mathematisch–naturwissenschaftlicher Bereich)
der Martin–Luther–Universität Halle–Wittenberg

von Herrn Dipl.-Math. Alf Gerisch

geb. am 2. September 1973 in Merseburg

Gutachter:

1. Prof. Dr. M. A. J. Chaplain (University of Dundee)

2. Prof. Dr. J. G. Verwer (CWI Amsterdam)

3. Prof. Dr. R. Weiner (Martin-Luther-Universität Halle-Wittenberg)

Halle (Saale), den 23. August 2001

Dank

An dieser Stelle möchte ich mich herzlichst bei Prof. Dr. Rüdiger Weiner für die kontinuierliche Betreuung und allumfassende Unterstützung während meiner Promotionszeit bedanken. Seine Ratschläge sowie die gemeinsamen Diskussionen waren mir eine große Hilfe.
Mein besonderer Dank gilt Helmut Podhaisky, Dr. Jörg Wensch, Dr. Eberhard Scholz und Dr. Lothar Boltze für ihre wertvollen Anregungen, sowie allen MitarbeiterInnen am Institut für Numerische Mathematik für das freundliche Arbeitsklima.
Bedanken möchte ich mich bei den Organisatoren des durch die DFG geförderten Graduiertenkollegs „Transport von Wirkstoffen in biologischen Systemen“, welches mir durch Bereitstellung eines Stipendiums die Promotion finanziell ermöglichte. Die gemeinsamen Veranstaltungen des Graduiertenkollegs werden mir in guter Erinnerung bleiben.
Ebenso danke ich Prof. Dr. Jan G. Verwer für die Einladungen zu Forschungsaufenthalten am CWI in Amsterdam. Die dort gesammelten wertvollen Erfahrungen flossen in die vorliegende Arbeit ein. Weiterhin danke ich Dr. David F. Griffiths und Prof. Dr. Mark A.J. Chaplain (University of Dundee, Schottland), die mein Interesse für mathematische Modelle aus den Biowissenschaften weckten.
Meinen Eltern bin ich für ihre uneingeschränkte Unterstützung in allen erdenklichen Situationen außerordentlich dankbar.

…, nicht zu vergessen: die vielen guten Freunde. Danke.

Abstract We describe and evaluate a method of lines (MOL) technique for the simulation of taxis–diffusion–reaction (TDR) systems. These time-dependent PDE systems arise when modelling the spatio-temporal evolution of a population of organisms which migrate in direct response to e.g. concentration differences of a diffusible chemical in their surrounding (*chemo*taxis). Examples include pattern formation and different processes in cancer development. The effect of taxis is modelled by a nonlinear advection term in the TDR system (the taxis term).
The MOL-ODE is obtained by replacing the spatial derivatives in the TDR system by finite volume approximations. These respect the conservation of mass property of the TDR system, and are constructed such that the MOL-ODE has a nonnegative analytic solution (positivity). The latter property is natural (because densities/concentrations are modelled) and highly desirable (because negative solution values might turn stable reaction terms into unstable ones). Diffusion and reaction terms can be replaced by standard approximations to ensure positivity, and we employ upwinding in combination with limiter functions in the discretization of the taxis term to ensure positivity of the MOL-ODE. The discretization near the boundary of the spatial domain is discussed. The appropriateness of the spatial discretization is demonstrated for a simple taxis problem (we provide the exact PDE solution).
The MOL-ODE is stiff and of large dimension. We develop integration schemes which treat the discretization of taxis and diffusion/reaction differently (splitting). We employ operator (Strang-) splitting and/or the approximate matrix factorization technique. The splitting schemes are based on explicit Runge-Kutta (ERK) and linearly-implicit W-methods. Positivity and stability of the integration schemes are investigated. We identify an ERK method with favourable positivity properties. A corresponding W-method is constructed. Numerical experiments with a variety of splitting schemes applied to some semi-discretized TDR systems confirm the broad applicability of the splitting schemes and lead to a selection of efficient methods for the class of TDR systems. These methods are more efficient than (suitable) standard ODE solvers in the lower and moderate accuracy range.
Altogether, the numerical technique developed is appropriate and efficient for the simulation of TDR systems.

Zusammenfassung Wir entwickeln und evaluieren eine Linienmethode (MOL) für die Simulation von Taxis–Diffusions–Reaktions (TDR)-Systemen. Diese zeitabhängigen PDE-Systeme treten bei der Modellierung der räumlich-zeitlichen Entwicklung von Populationen von Organismen auf, die sich in direkter Antwort auf z.B. Konzentrationsunterschiede in diffundierenden Chemikalien in ihrer Umgebung bewegen (*Chemo*taxis). Beispiele sind Musterbildungsvorgänge und verschiedene Prozesse in der Tumorentwicklung. Der Taxiseffekt wird durch einen nichtlinearen Advektionsterm im TDR-System modelliert (Taxisterm).
Die MOL-ODE erhalten wir durch Ersetzen der Ortsableitungen im TDR-System mit Finite-Volumen-Approximationen. Diese beachten die Massenerhaltungseigenschaft des TDR-Systems und sind so konstruiert, daß die MOL-ODE eine nichtnegative analytische Lösung besitzt (Positivität). Letztere Eigenschaft ist natürlich (da Dichten/Konzentrationen modelliert werden) und sehr wünschenswert (da negative Lösungswerte stabile Reaktionsterme in instabile verwandeln können). Diffusions- und Reaktionsterme können durch Standardapproximationen ersetzt werden, um die Positivität zu sichern. Wir verwenden *Upwinding* in Kombination mit *Limiterfunktionen* in der Diskretisierung des Taxisterms, um die Positivität der MOL-ODE zu erzielen. Die Diskretisierung in der Nähe des Randes des räumlichen Gebiets wird diskutiert. Die Angemessenheit der räumlichen Diskretisierung wird anhand eines einfachen Taxisproblems demonstriert (wir geben die exakte PDE-Lösung an).
Die MOL-ODE ist steif und hochdimensional. Wir entwickeln Integrationsverfahren, welche die Diskretisierung des Taxisterms und der Diffusions-/Reaktionsterme unterschiedlich behandeln (Splitting). Wir verwenden Operator-(Strang-) Splitting und/oder die Technik der approximierenden Matrixfaktorisierung. Die Splittingmethoden basieren auf expliziten Runge-Kutta (ERK) und linear-impliziten W-Methoden. Positivität und Stabilität der Integrationsverfahren werden untersucht. Wir identifizieren eine ERK-Methode mit vorteilhaften Positivitätseigenschaften. Eine zugehörige W-Methode wird konstruiert. Numerische Experimente mit einer Vielzahl von Splittingmethoden angewendet auf einige semidiskretisierte TDR-Systeme bestätigen die breite Anwendbarkeit der Splittingmethoden und führen zu einer Auswahl effizienter Methoden für die betrachtete Klasse von TDR-Systemen. Diese Methoden sind effizienter als (geeignete) Standard-ODE-Integratoren im unteren und mittleren Genauigkeitsbereich.
Insgesamt wurde eine geeignete und effiziente numerische Technik zur Simulation von TDR-Systemen entwickelt.

[This page is empty.]

Contents

Abbreviations, Symbols, and Notation

AMF	approximate matrix factorization
BC	boundary condition
EC	endothelial cell
ECM	extracellular matrix
ERK method	explicit Runge-Kutta method
FVM	finite volume method
IVP	initial value problem
MDE	matrix degradative enzymes
MOL	method of lines
MOL-ODE	the first step (spatial discretization) of the MOL results in this ODE system
ODE	ordinary differential equation
OPS	operator splitting
PDE	partial differential equation
RD	reaction–diffusion
RK method	Runge-Kutta method
TAF	tumour angiogenesis factors
TDR	taxis–diffusion–reaction

Chap.	Chapter
Sec.	Section
Eq.	Equation
Tab.	Table
Fig.	Figure

$\mathbb{R},\ \mathbb{R}^m,\ \mathbb{R}^{m,m}$	real numbers, real vectors of dimension m, real $m \times m$ matrices
$\mathbb{R}_+,\ \mathbb{R}_{+,0}$	positive real numbers and nonnegative real numbers
$\mathbb{C}, \Re z, \Im z$	complex numbers, real and imaginary part of $z \in \mathbb{C}$
$\mathbb{C}_-, \mathbb{C}_{-,0}$	$\mathbb{C}_- := \{z \in \mathbb{C},\ \Re z < 0\}$, $\mathbb{C}_{-,0} := \{z \in \mathbb{C},\ \Re z \le 0\}$
W_α	$:= \{z \in \mathbb{C} \ : \ \lvert\arg(-z)\rvert \le \alpha\}$ for $\alpha \in [0, \pi/2]$ (closed wedge in the left complex half plane)
$\mathbb{N}$	$= \{1, 2, 3, \ldots\}$
d	spatial dimension
$\Omega \subset \mathbb{R}^d$	bounded, nonempty (spatial) domain with piecewise smooth boundary
I_T, T	$:= (0, T)$ time domain, $T \in \mathbb{R}_+$ final time
$\mathbf{x} = (x_j)_{j=1}^d, t$	space and time variable
$\mathcal{I}$	finite index set of partition of Ω
$\mathbf{i} = (i_j)_{j=1}^d$	d-dimensional multi-index (referring to an element of a partition of Ω)
$\mathfrak{n}(\mathbf{x})$	outer unit normal vector in the boundary point $\mathbf{x}$ of some spatial domain $\subset \mathbb{R}^d$
$\mathbf{e}_j$	jth unit vector (of appropriate dimension)
$\mathbb{1}$	$= (1, 1, \ldots, 1)$-vector (of appropriate dimension)
h, τ, τ_k	spatial grid size and time step sizes (generic and in step k)

n	cell density function
l	number of chemical concentrations
$\mathbf{c} = (c_j)_{j=1}^{l}$	vector function of the concentration of the chemicals
$p_1, p_2, \ldots, p_l$	taxis functions
$p_0, \mathbf{g}_0$	reaction term functions
ε	random motility coefficient in the taxis equation
u	PDE solution
$\mathbf{U}(t)$	$= (U_{\mathbf{i}}(t))_{\mathbf{i}\in\mathcal{I}}$ time-continuous approximation of u on a partition of Ω
H	right-hand side of the exact cell average evolution equation
$\mathcal{H}$	right-hand side of approximate (FVM) cell average evolution equation
$\mathcal{F}$	flux approximation
$\mathcal{D}, \mathcal{T}$	diffusion and taxis flux approximations
$\mathcal{S}^{\pm}, \Phi$	state interpolants and limiter function in the taxis flux approximation
Ψ, Ψ_0, Ψ_1	approximate evolution operators (OPS methods)
$\mathcal{P}$	class of right-hand side functions of positive ODEs
C^k	space of k times differentiable functions
$L^k(\Omega)$	space of kth power Lebesgue-integrable functions over Ω

We write in short $i(k)j$ for the sequence $i, i+k, i+2k, \ldots, j$ of integer numbers.
The inner product $\sum_{j=1}^{m} a_j b_j$ of two vectors $a, b \in \mathbb{R}^m$ is denoted by $a \cdot b$.
Relation symbols are to understand in a componentwise sense in this work.
We denote the total derivative of a vector valued function $y : \mathbb{R} \to \mathbb{R}^m$ with respect to t as either $\frac{\mathrm{d}\, y(t)}{\mathrm{d}\, t}$ or $y'(t)$. The partial derivative of a vector valued function $f : \mathbb{R} \times \mathbb{R}^d \to \mathbb{R}^m$ with respect to t or x_j ($j = 1(1)d$) is denoted by $\partial_t f(t, \mathbf{x})$ and $\partial_{x_j} f(t, \mathbf{x})$, respectively. The Jacobian matrix of the function $f(t, \mathbf{x})$ is given and denoted by

$$\frac{\partial f(t, \mathbf{x})}{\partial \mathbf{x}} \equiv [\partial_{x_1} f(t, \mathbf{x}), \cdots, \partial_{x_d} f(t, \mathbf{x})] \ .$$

The Nabla operator ∇ and the Laplace operator Δ (in d dimensions) are defined by

$$\nabla := (\partial_{x_1}, \partial_{x_2}, \cdots, \partial_{x_d})^{\mathsf{T}} \quad \text{and} \quad \Delta := \nabla \cdot \nabla = \partial_{x_1}^2 + \partial_{x_2}^2 + \cdots + \partial_{x_d}^2 \ .$$

In Chapters 2 and 3 we use **bold** face letters to denote vectors and matrices. Thereafter, we switch to normal (math mode) *italic* face letters except in some places when referring to material discussed in Chapters 2 or 3.

Chapter 1

Introduction

Mathematical models are an important tool in most areas of science and research. They form the basis for the control of many technical systems (chemical engineering, space crafts, airbags, ...), they can improve the efficiency of such systems and hence, for example, reduce environmental stress. They are also used for short and long term prediction of weather and climate changes. The application of mathematical models in the life sciences is another, rapidly growing research activity. Models are used here to understand biological systems (e.g. pattern formation and growth processes) and to gain insight into otherwise nonobservable mechanisms. Further, mathematical models describing phenomena from the life sciences can be used for educational purposes. There are three main tasks of mathematical models:

- reproduction of real life processes,
- prediction of results under the variation of internal and external parameters,
- discovery of new results about the model which then must be validated in reality.

Models cover the reality partially only in order to be manageable and not to draw away attention from major processes to minor processes. The art of modelling is to carefully select variables and processes which are significant for the modelling goal and to neglect unnecessary details. Models of small, understandable systems are then combined to produce more and more complex models.
In this thesis we are concerned with the development and evaluation of numerical techniques for the simulation of taxis–diffusion–reaction (TDR) systems from mathematical biology. Hence, we already assume that the model exists; new models are not derived in this work.
TDR systems are time-dependent partial differential equation (PDE) systems composed of a taxis equation describing the evolution of the density function of a population of organisms and a reaction–diffusion (RD) subsystem describing the evolution of concentrations of substances in the surrounding of the organisms. The existence of a density function for the population of organisms implies that this population is sufficiently large within the considered spatial domain. An important characteristic is that the organisms can sense spatial differences in the concentration of the surrounding substances and migrate in direct response to this signal – a process known as taxis. If the migration depends on the concentration field of a soluble (diffusible) chemical then the process is termed *chemotaxis*; if this chemical is bound to some underlying substratum (e.g. extracellular matrix) then we talk about *haptotaxis* (other forms of taxis are also possible e.g. galvanotaxis,

phototaxis, gravitaxis [44]). A substance may have either an attracting or a repelling influence on the migration of the organisms depending on the interactions between them. In both cases, the migration due to taxis is proportional to the gradient (in space) of the substance. Beside this migration due to taxis, the organisms might also relocate by random motility (modelled the same way as molecular diffusion). In this work we assume that the effect of random motility is small compared to the effect of taxis. In Chap. 2 we derive a TDR conservation equation and make precise the class of TDR systems which are to be considered in this thesis. Following this we present some TDR systems from the literature describing biological processes. Later, Chap. 5, we present simulation results of these models. There we consider the models with random motility and, additionally, also with the random motility term in the taxis equation switched off. There exist also processes which rely on taxis but cannot be described by the TDR systems considered here. The green turtle, for instance, is supposed to travel over 1000 kilometres to reach its breeding place through detection of an unknown chemical source originating there [34]. Obviously, we cannot define a sensible density function for the organisms, i.e. the turtles, in this case because their number is too low to justify this. Here we would have to take into account the effect of taxis on each single turtle and trace their movements individually. This case is not considered here but the coupling of discrete and continuous (RD subsystem) structures is an interesting area for future research.

In Chap. 3 and Chap. 4 we develop and describe the numerical technique for the simulation of TDR systems. We follow the Method of Lines (MOL) which decouples the discretization of spatial and temporal derivatives in the equations. The first step of the MOL is the spatial discretization (semi-discretization) leading to an initial value problem (IVP) for an ordinary differential equation (ODE) system, the so-called MOL-ODE. This part is described in Chap. 3. The space- and time-dependent density and concentrations in the TDR system are naturally nonnegative (we, in fact, require this property from the model). Therefore, we also require that the semi-discretization leads to a MOL-ODE with a nonnegative analytic solution (positive ODE system). We cite from the literature conditions on the ODE system which guarantee that this requirement is satisfied. Thereafter, we detail the discretization of the taxis, diffusion, and reaction terms of the TDR system. We follow the finite volume methodology to derive the semi-discretization. Whereas diffusion and reaction terms are replaced by standard approximations, the taxis term deserves special attention. The taxis term is present in the taxis equation only and the solution of this equation (density function of the organism population) generally contains steep moving fronts. A simple central or even upwinding (taking the flow direction into account) discretization would introduce oscillations and subsequently negative solution values into the solution of the MOL-ODE. Firstly, this would contradict our requirement that we want a nonnegative solution of the MOL-ODE, and secondly, negative solution values might turn a stable reaction term into an unstable one, and this in turn gives rise to numerical problems when solving the MOL-ODE. Therefore we use limiter functions in an upwinding discretization of the taxis term such that a nonnegative solution is enforced. This approach is widely used in the numerical solution of nonlinear conservation laws and is here applied to the taxis term in an adapted form. This way we combine second-order accuracy with nonnegativity of the solution. Nonnegativity can also be achieved by first-order upwinding without limiter function but this would require an excessive amount of spatial grid points in order to attain a reasonable accuracy and is therefore considered to be no option. Some numerical evidence for this statement is provided in Sec. 3.4, where the taxis discretization is evaluated for a simple model problem (Model 1). We also detail the semi-discretization of the TDR system near the boundary.

Altogether we arrive at an IVP for the MOL-ODE which is guaranteed to have a nonnegative solution. The MOL-ODE is (in general) a second-order consistent approximation of the TDR system (in a finite difference sense). This system is of very large dimension (at least if the spatial dimension is greater than one and this is the case for all biomathematical models considered here) and it is stiff due to the diffusion (and possibly also due to the reaction) terms.

Chap. 4 is devoted to the development of appropriate integration schemes for the solution of the MOL-ODE. Stiffness requires the application of implicit (or linearly-implicit) schemes because otherwise we would face a severe time step size restriction and hence unacceptable computational costs. On the other hand, an ODE system with the taxis discretization as right-hand side function is efficiently solved by explicit methods. This also avoids problems with the possible nonexistence of derivatives (Jacobian matrix) of the taxis discretization (due to the non-differentiability of the limiter function) which are required in implicit integration methods. We try to combine both demands by employing splitting methods for the solution of the MOL-ODE. The first approach (AMF – approximate matrix factorization) is based on linearly-implicit Rosenbrock-type W-methods (henceforth in short W-methods). These schemes are applied to the full MOL-ODE and they use an inexact Jacobian matrix of the right-hand side function. This matrix is obtained by, firstly, neglecting the taxis discretization in the Jacobian computation, and, secondly, approximately factorizing the matrix in the stage equations such that linear systems with this matrix can be solved efficiently (banded matrices). The second approach is operator (or Strang-) splitting (OPS). OPS splits the right-hand side F of the MOL-ODE into a sum of two parts: the discretization of the taxis term F_0 and the discretization of the diffusion–reaction terms F_1 (each with corresponding boundary treatment). Then ODEs with either part as right-hand side are solved in turn. If the right-hand side is F_0 then an explicit Runge-Kutta (ERK) method is used. If the right-hand side is F_1 then a W-method with AMF is employed. The splitting techniques AMF and OPS are introduced in Sec. 4.2 (following a general introduction to Runge-Kutta (RK) and Rosenbrock-type methods in Sec. 4.1) and detailed with specific methods in Sec. 4.5 and Sec. 4.6. All methods derived are accurate of order two. This is a suitable compromise for the class of problems under consideration: first-order methods are too inefficient because they require too many time steps to reach a certain level of accuracy and higher order methods might fail to be efficient because of a lack of smoothness in the solution of the MOL-ODE. Sec. 4.3 and Sec. 4.4 discuss methods which are applied or are fundamental in the AMF and OPS schemes from the point of positivity and stability, see the next paragraph for a more detailed description. Finally, Sec. 4.7 describes alternative methods for the solution of the MOL-ODE and different splitting approaches.

The spatial discretization of the TDR system results in a MOL-ODE which is guaranteed to have a nonnegative analytic solution (positive ODE system). Our aim is to have this property also for the numerical solution of the MOL-ODE. The most troublesome part of the MOL-ODE with respect to this is the taxis discretization. In Sec. 4.3 we discuss the positivity of numerical schemes if applied to positive ODE systems. We start with the positivity of RK and Rosenbrock-type methods applied to problem classes of linear, positive ODEs. The foundations of this theory are already given by Bolley and Crouzeix [6] in 1978. We give a characterization of the class of $\bar{M}$-matrices which are important in their theory. After presenting the main results of their theory, we give relaxed conditions on the problem class such that the results of Bolley and Crouzeix are still valid if we consider ERK schemes only. This theory is then applied to lower order ERK methods (especially three-stage, second-order methods) and lower order Rosenbrock-type methods. Next we consider

the positivity of RK methods applied to subclasses of positive, dissipative problems. The respective positivity theory is developed by Horvath [27]. We identify a unique three-stage, second-order ERK method (we refer to this method as RK32) with optimal positivity properties on this problem class (i.e. RK32 is the unique method from this method class which can take the largest time steps without violating the nonnegativity of the numerical solution for all problems from the problem class). Finally, we consider the positivity of ERK methods applied to general nonlinear, positive ODEs based on work by Shu and Osher [52] and Hundsdorfer et al. [31]. We again identify RK32 as an optimal method. This approach is directly applicable to the spatial discretization of the taxis term, see Sec. 4.4. In Sec. 4.4 we investigate the specific positivity and stability properties of RK32 applied to the ODE system arising from the discretization of the taxis term. We compare our findings with results obtained for standard second- and third-order ERK methods.
In Chap. 5 we apply the splitting methods devised to the biomathematical models described in Chap. 2, and discuss the results obtained. We also compare with two general purpose integration schemes for large, stiff ODE systems (VODPK and ROWMAP). We also describe the dynamics of the solution for each of the biomathematical models. The most important quantity in all models is the density function of the organisms and this function is depicted for different output times.
The main findings and conclusions are finally summarized in Chap. 6. There we also give possible future research directions in the field of numerical simulation of TDR systems.

Chapter 2

Taxis–Diffusion–Reaction Systems

In this chapter we define the class of taxis–diffusion–reaction (TDR) systems which are the subject of this thesis. To this end, we start in Sec. 2.1 with the derivation of a conservation equation which contains all the important terms to model taxis, diffusion, and reaction. Whereas diffusion and reaction are often discussed in the literature, taxis terms came into the focus of numerical interest just recently. However, they form important ingredients of many models from mathematical biology. The problem class of this work is made precise in Sec. 2.2. Finally, in Sec. 2.3, we give a collection of TDR models from mathematical biology. The purpose of this collection is to illustrate the importance of TDR models in mathematical biology on one hand, and to have a few examples for testing and evaluation of the numerical schemes which are developed in the following chapters on the other hand. The collection contains also a simple taxis test problem (Model 1), where we can provide an analytic solution.

2.1 Derivation of a TDR conservation equation

Let $I \subset \mathbb{R}_{+,0}$ be a time interval and $\Omega \subset \mathbb{R}^d, d \in \mathbb{N}$, be a bounded, nonempty (spatial) domain with piecewise smooth boundary such that the Gauß integral theorem can be applied. For any subdomain of Ω we assume the same properties.
We describe the derivation of a conservation equation for a scalar quantity Q, see e.g. [41, p. 14], [1, p. 54]. The resulting equation contains terms which model the effects of taxis, diffusion and reaction. We denote with $u(t, \mathbf{x})$ the density of Q at $\mathbf{x} \in \Omega$ in space and at $t \in I$ in time such that $\int_{\widetilde{\Omega}} u(t, \mathbf{x}) d\mathbf{x}$ is its total mass in any subdomain $\widetilde{\Omega} \subset \Omega$ at time $t \in I$. The total mass of Q in $\widetilde{\Omega}$ can only change in time by production or destruction of Q within $\widetilde{\Omega}$ or by a flow of Q through the boundary $\partial\widetilde{\Omega}$ of $\widetilde{\Omega}$. Let $s(t, \mathbf{x}) \in \mathbb{R}$ denote the source density of Q at $(t, \mathbf{x})$ (positive for production and negative for destruction) and $\mathbf{v}(t, \mathbf{x}) \in \mathbb{R}^d$ be the velocity field associated with Q at $(t, \mathbf{x})$. (The functions s and $\mathbf{v}$ may also depend on $u(t, \mathbf{x})$ or its spatial derivatives.) The rate of mass flow or mass flux of Q at $(t, \mathbf{x})$ is given by $u(t, \mathbf{x})\mathbf{v}(t, \mathbf{x})$ and the function

$$\mathbf{f}(t, \mathbf{x}) := u(t, \mathbf{x})\mathbf{v}(t, \mathbf{x})$$

is the flux function. The change of total mass of Q in any subdomain $\widetilde{\Omega} \subset \Omega$ is hence given by

$$\frac{\mathrm{d}}{\mathrm{d}\, t} \int_{\widetilde{\Omega}} u(t, \mathbf{x}) d\mathbf{x} = - \oint_{\partial\widetilde{\Omega}} \mathbf{f}(t, \mathbf{x}) \cdot \mathfrak{n}(\mathbf{x}) d\partial\widetilde{\Omega} + \int_{\widetilde{\Omega}} s(t, \mathbf{x}) d\mathbf{x}\,. \tag{2.1}$$

Here, $\mathfrak{n}(\mathbf{x})$ is the outer unit normal at the point $\mathbf{x} \in \partial\widetilde{\Omega}$. The minus sign in front of the surface integral in the equation ensures that an inflow into $\widetilde{\Omega}$ leads to an increase of the mass of u in $\widetilde{\Omega}$. We remark that the surface integral in Eq. (2.1) reduces in one spatial dimension ($\widetilde{\Omega} = (\mathbf{x}_l, \mathbf{x}_r) \subset \mathbb{R}$) to a point evaluation of the flux function: $\oint_{\partial\widetilde{\Omega}} \mathbf{f}(t,\mathbf{x})\mathfrak{n}(\mathbf{x})d\partial\widetilde{\Omega} = \mathbf{f}(t,\mathbf{x}_r) - \mathbf{f}(t,\mathbf{x}_l)$. Eq. (2.1) is an integral form of the conservation equation for Q. If, for instance, the source $s \equiv 0$ in $\widetilde{\Omega}$ and the flux $\mathbf{f} \equiv 0$ on $\partial\widetilde{\Omega}$ then the total mass of Q in $\widetilde{\Omega}$ is conserved. This conservation property should carry over to numerical approximations of u.
If $u, \mathbf{v} \in C^1(I \times \Omega)$ then, under the smoothness assumption on the boundary of $\widetilde{\Omega}$, we can apply the integral theorem of Gauß to the surface integral in Eq. (2.1) and obtain

$$\frac{\mathrm{d}}{\mathrm{d}\,t}\int_{\widetilde{\Omega}} u(t,\mathbf{x})d\mathbf{x} = -\int_{\widetilde{\Omega}} \nabla\cdot\mathbf{f}(t,\mathbf{x})d\mathbf{x} + \int_{\widetilde{\Omega}} s(t,\mathbf{x})d\mathbf{x}\,. \tag{2.2}$$

We can differentiate under the integral, and if we further assume that $s \in C^0(I\times\Omega)$ then it follows that u satisfies the transport equation

$$\partial_t u(t,\mathbf{x}) + \nabla\cdot\mathbf{f}(t,\mathbf{x}) = s(t,\mathbf{x})\,, \quad \text{for all } (t,\mathbf{x}) \in I\times\Omega\,, \tag{2.3}$$

the differential form of the conservation equation for Q.
For sufficiently smooth functions u, s and $\mathbf{v}$, both the integral and the differential form of a conservation law are equivalent. However, we will make use of the convenient notation of the differential form even if the assumptions on u, s and $\mathbf{v}$ are not satisfied. We will understand Eq. (2.3) in the sense of Eq. (2.1) for arbitrary subdomains $\widetilde{\Omega} \subset \Omega$ in this case.
Eq. (2.3) becomes a scalar conservation equation for u if $\mathbf{v}$ and s are *a priori* known functions or if they are functions of u itself. In this case, Eq. (2.3), together with suitable initial and boundary conditions, can be solved on its own. More frequently we encounter the situation that the velocity $\mathbf{v}$ or the sources s depend on other conserved quantities—the conservation equations for all these quantities constitute a system of conservation equations. The models from mathematical biology considered in this work are such systems.
We consider models with two different flux types. First, there is diffusive flux which we assume, according to Fick's law, to be proportional to the gradient of u itself ($D > 0$ is the diffusion coefficient of u)

$$\mathbf{f}_D(t,\mathbf{x}) = -D\nabla u(t,\mathbf{x})\,.$$

This definition is based on the assumption that the quantity Q is transported (diffuses) from regions of high density to regions of low density.
A second type of flux appears if the velocity field depends on the gradient of the density or concentration $c_1(t,\mathbf{x})$ of some other quantity. This so-called tactic flux is given by

$$\mathbf{f}_{T_1}(t,\mathbf{x}) = u(t,\mathbf{x})p_1(\mathbf{c}(t,\mathbf{x}))\nabla c_1(t,\mathbf{x})\,.$$

We include the function p_1 which may depend on various quantities (with concentrations collected in the vector $\mathbf{c} = (c_1, c_2, \ldots, c_l)^\mathsf{T}$) to model the strength of the tactic response of Q to the quantity described by c_1 and also to model whether higher densities c_1 attract ($p_1 > 0$) or repel ($p_1 < 0$) the quantity Q. Tactic fluxes feature in a broad range of models from mathematical biology, e.g. pattern formation and growth processes. A class of such models is considered in this work. Othmer

and Stevens [46] derive different forms of tactic flux functions based on continuous time, discrete space random walk models. The PDE models are obtained in the so-called diffusion limit. Of course, diffusion and taxis may happen at the same time and Q may be under the tactic influence of several substances. Such an extension of the work in [46] with multiple tactic cues is treated in [47]. We assume that the total flux is the sum of the individual contributions and arrive altogether at the flux function

$$\mathbf{f}(t,\mathbf{x}) = \mathbf{f}_D(t,\mathbf{x}) + \sum_{j=1}^{l} \mathbf{f}_{T_j}(t,\mathbf{x}) = -D\nabla u(t,\mathbf{x}) + u(t,\mathbf{x}) \sum_{j=1}^{l} p_j(\mathbf{c}(t,\mathbf{x}))\nabla c_j(t,\mathbf{x})\,.$$

The source term $s(t,\mathbf{x})$ often represents chemical reactions of the quantities described by u and $\mathbf{c}$ of a system beside explicitly modelling sinks and sources. Therefore $s(t,\mathbf{x})$ depends most often nonlinearly on the density of the quantities of the system, i.e. $s(t,\mathbf{x}) = p_0(t,\mathbf{x},u(t,\mathbf{x}),\mathbf{c}(t,\mathbf{x}))$. Finally, we arrive at the following integral form of the taxis–diffusion–reaction equation for u which follows from Eq. (2.1) by inserting the derived flux expression $\mathbf{f}(t,\mathbf{x})$ and the source function (we neglect the dependence of $u, \mathbf{c}$ and the normal vector $\mathfrak{n}$ on t and $\mathbf{x}$ in the notation):

$$\frac{\mathrm{d}}{\mathrm{d}\,t}\int_{\widetilde{\Omega}} u d\mathbf{x} = -\oint_{\partial\widetilde{\Omega}} \left[-D\nabla u + u\sum_{j=1}^{l} p_j(\mathbf{c})\ \nabla c_j\right]\cdot\mathfrak{n}\ d\partial\widetilde{\Omega} + \int_{\widetilde{\Omega}} p_0(t,\mathbf{x},u,\mathbf{c})d\mathbf{x}\,. \tag{2.4}$$

2.2 Problem class

In this section we specify and describe the class of problems which we want to solve numerically. Let $I_T := (0,T), T \in \mathbb{R}_+$, be a time interval and $\Omega \subset \mathbb{R}^d, d \in \mathbb{N}$, a bounded, nonempty spatial domain with piecewise smooth boundary $\partial\Omega =: \Gamma$.
Consider real-valued, time- and space-dependent functions $n(t,\mathbf{x})$ and $\mathbf{c}(t,\mathbf{x})$

$$n : \bar{I}_T \times \bar{\Omega} \to \mathbb{R} \ \text{ and } \mathbf{c} : \bar{I}_T \times \bar{\Omega} \to \mathbb{R}^l\,,$$

which denote the density of a population of organisms and a vector of l concentrations of certain substances (e.g. chemicals), respectively. We study the taxis–diffusion–reaction system for $n(t,\mathbf{x})$ and $\mathbf{c}(t,\mathbf{x})$

$$\partial_t n = \varepsilon\Delta n - \nabla\cdot\left(n\ \sum_{j=1}^{l} p_j(\mathbf{c})\nabla c_j\right) + p_0(t,\mathbf{x},n,\mathbf{c})\,, \qquad (t,\mathbf{x}) \in I_T\times\Omega\,, \tag{2.5a}$$

$$\partial_t \mathbf{c} = \mathbf{D}\Delta\mathbf{c} + \mathbf{g}_0(t,\mathbf{x},n,\mathbf{c})\,, \tag{2.5b}$$

with initial conditions

$$n(0,\mathbf{x}) = n_0(\mathbf{x})\,, \quad \mathbf{c}(0,\mathbf{x}) = \mathbf{c}_0(\mathbf{x})\,, \qquad \mathbf{x} \in \bar{\Omega}\,, \tag{2.5c}$$

and boundary conditions (for n and for $c_j\,, j = 1(1)l$, with $D_j > 0$)

$$\begin{aligned}
n(t,\mathbf{x}) &= \alpha_D^{(0)}(t,\mathbf{x}) \geq 0\,, \quad (t,\mathbf{x}) \in I_T\times\Gamma_D^{(0)}\,,\\
c_j(t,\mathbf{x}) &= \alpha_D^{(j)}(t,\mathbf{x}) \geq 0\,, \quad (t,\mathbf{x}) \in I_T\times\Gamma_D^{(j)}\,,\\
\left(-\varepsilon\nabla n + n\left(\textstyle\sum_{j=1}^{l} p_j(\mathbf{c})\nabla c_j\right)\right)\cdot\mathfrak{n}(\mathbf{x}) &= \alpha_F^{(0)}(t,\mathbf{x}) \leq 0\,, \quad (t,\mathbf{x}) \in I_T\times\Gamma_F^{(0)}\,,\\
-D_j\nabla c_j(t,\mathbf{x})\cdot\mathfrak{n}(\mathbf{x}) &= \alpha_F^{(j)}(t,\mathbf{x}) \leq 0\,, \quad (t,\mathbf{x}) \in I_T\times\Gamma_F^{(j)}\,.
\end{aligned} \tag{2.5d}$$

Here $\varepsilon \in \mathbb{R}_{+,0}$ is a constant, $\mathbf{D} = \mathrm{diag}(D_1, D_2, \ldots, D_l) \in \mathbb{R}^{l,l}$ is a constant, diagonal matrix with nonnegative entries, and $p_j : \mathbb{R}^l \to \mathbb{R}$ for $j = 1(1)l$, $p_0 : I_T \times \Omega \times \mathbb{R}^{l+1} \to \mathbb{R}$ and $\mathbf{g}_0 : I_T \times \Omega \times \mathbb{R}^{l+1} \to \mathbb{R}^l$ are given functions. The prescribed functions n_0 and $\mathbf{c}_0$ define the initial data of the system and the functions $\alpha_D^{(j)}$ and $\alpha_F^{(j)}$, $j = 0(1)l$, its boundary data. No boundary data is prescribed for c_j if $D_j = 0$. We consider two different kinds of boundary conditions (BCs) for all other c_j and for n: Dirichlet BCs on $\Gamma_D^{(j)} \subset \Gamma$ with prescribed (nonnegative) state and flux BCs on $\Gamma_F^{(j)} \subset \Gamma$ with prescribed (inflow) flux (where $\Gamma_D^{(j)} \cap \Gamma_F^{(j)} = \emptyset$). Further, $\Gamma_D^{(j)} \cup \Gamma_F^{(j)} = \Gamma$ for $j > 0$ and $\Gamma_D^{(0)} \cup \Gamma_F^{(0)} \subset \Gamma$ (if $\varepsilon > 0$ then equality is required also for $j = 0$).
Some general remarks on the meaning of the parameters and functions in Eqs. (2.5) are in order.

- The population density n diffuses with diffusion constant $\varepsilon > 0$, or exhibits no diffusion if $\varepsilon = 0$. This models the random motility of the organisms described by the density n.

- The chemical concentrations in $\mathbf{c}$ can also change by diffusion, or they can be non-diffusible (then the corresponding diagonal entry in $\mathbf{D}$ is zero).

- A characteristic property is that the evolution of n depends on gradients ∇c_j of the chemical concentrations—a process known as taxis which adds (nonlinear) advection terms to the population equation. The strength and the sign of the tactic influence of each chemical concentration c_j on the population density n is described by the function $p_j(\mathbf{c})$. If $p_j(\mathbf{c}) > 0$ then c_j acts as an attractant (the population migrates up gradients, i.e. towards higher concentrations c_j); c_j is a repellant for $p_j(\mathbf{c}) < 0$.

- We focus on systems where the speed of migration of the organisms in the population induced by diffusion is much smaller than the speed of migration induced by the taxis term, or where there is no random motility (diffusion) in the population at all.

- The reaction term $p_0(t, \mathbf{x}, n, \mathbf{c})$ accounts for creation or loss of entities in the population due to interaction with themselves or with the chemicals. The reactions between the chemicals and the population are modelled through the function $\mathbf{g}_0(t, \mathbf{x}, n, \mathbf{c})$.

Eqs. (2.5) are valid on general domains Ω in space. However, in this work we will restrict our attention to d-dimensional unit cubes Ω in space and finite time intervals I_T,

$$\Omega := (0,1)^d\,, d \in \mathbb{N}\,, \text{ and } I_T := (0,T)\,,\ T \in \mathbb{R}_+\,. \tag{A1}$$

The numerical schemes are described for $d \in \mathbb{N}$ but in the numerical experiments we restrict attention to $d = 2$.
Only nonnegative solutions of the system (2.5) make sense from a modelling point of view because the functions n and $\mathbf{c}$ describe densities or concentrations and as such they are naturally nonnegative. Therefore, any model about the temporal and spatial development of n and $\mathbf{c}$ should respect this property and allow only nonnegative solutions. We assume henceforth:

The problem (2.5), together with the functions, parameters, and initial and boundary data prescribed has a unique, nonnegative solution $(n(t,\mathbf{x}), \mathbf{c}(t,\mathbf{x}))$ for all $(t,\mathbf{x}) \in \bar{I}_T \times \bar{\Omega}$. (A2)

2.3 Collection of TDR models

We describe four mathematical models which are of the general form of a TDR system given in (2.5). Model 1 (taken from [58]) is actually not really a TDR system but a scalar taxis equation. We derive an analytic solution for this problem. The availability of this solution makes this model ideal for evaluating the taxis discretization algorithm which is developed in the following chapters. The other three models arise from the study of certain aspects of tumour growth. They describe the processes of tumour induced angiogenesis (Models 2 and 3) and tumour invasion (Model 4).
There exist other mathematical models describing biological processes which fit into the framework of TDR systems. Pattern formation of bacterial populations often relies on chemotactic cues, see e.g. [59], and the study of the aggregation phase of the social amoeba *Dictyostelium discoideum*, e.g. [26, 25], is an example from developmental biology. Further, the onset of capillary sprout formation is also modelled as a TDR system in [45] and recently in [42].

2.3.1 A simple taxis test model (Model 1)

This model is taken from [58] and will be used to evaluate our taxis discretization algorithm. In the model, a scalar quantity (density n) is simply advected up the gradient of an attractant with fixed concentration profile. The problem is posed on the unit square, $\Omega = (0,1)^2$, and for the time interval I_T with $T = 0.021$. The attractant concentration is radially symmetric with centre $\left(\frac{1}{2}, \frac{1}{2}\right)$

$$c_1(t, \mathbf{x}) := \widetilde{c}(r(\mathbf{x})) = 1 - \cos(4\pi r(\mathbf{x})) \quad \text{for all } (t, \mathbf{x}) \in \bar{I}_T \times \bar{\Omega}\,,$$

where

$$r(\mathbf{x}) := \left(\left(x_1 - \tfrac{1}{2}\right)^2 + \left(x_2 - \tfrac{1}{2}\right)^2\right)^{\frac{1}{2}}\,,$$

is the distance of $\mathbf{x}$ from the centre of Ω. This corresponds to a ring of chemoattractant with maximum at $r = \frac{1}{4}$, see Fig. 2.1 (left).
The model equation is given by

$$\partial_t n = -\nabla \cdot (n \nabla c_1)\,, \quad \text{for } (t, \mathbf{x}) \in I_T \times \Omega. \tag{2.6}$$

We use parametrized initial data with parameter $0 \le \kappa < 0.1$

$$n(0, \mathbf{x}) = n_\kappa(r(\mathbf{x})) = \begin{cases} 1 & : \ r \le 0.4 - \kappa \\ \frac{1}{2}\left(1 + \cos\left(\frac{r - 0.4 + \kappa}{2\kappa}\pi\right)\right) & : \ 0.4 - \kappa < r \le 0.4 + \kappa \\ 0 & : \ r > 0.4 + \kappa \end{cases}\,. \tag{2.7}$$

This initial data has continuous first derivatives in space if $\kappa > 0$. The parameter κ controls the steepness of the front in the initial data. We use two different values in our experiments: $\kappa = 0.09$ for a fairly smooth initial condition, see Fig. 2.1 (right), and $\kappa = 0$ which is a jump initial condition with jump at $r = 0.4$. The initial condition and its gradient are zero for all $r \ge \frac{1}{2}$ if $\kappa < 0.1$. We assume no-flux boundary conditions for n. This is consistent with the initial data, and, together with the given attractant concentration, implies that the boundary has no influence on the solution in Ω. As time proceeds, the population n moves up the gradient of c_1 and tends to cluster into a ridge at $r = \frac{1}{4}$ where c_1 has its maximum value.

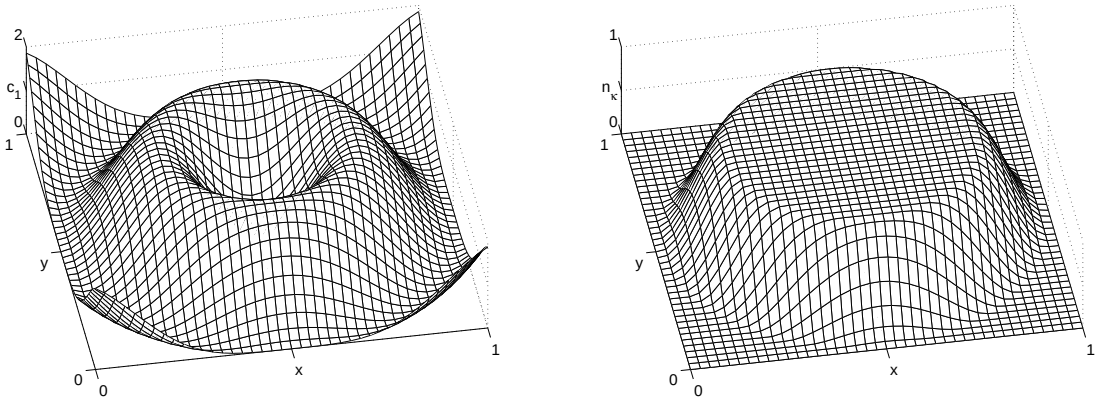

Figure 2.1: Concentration c_1 of attractant chemical (left) and initial population density n_κ for $\kappa = 0.09$ (right) of Model 1.

The solution n of this problem is radially symmetric and we define $\widetilde{n}(t,r)$ such that $\widetilde{n}(t,r(\mathbf{x})) = n(t,\mathbf{x})$ for all $\mathbf{x} \in \bar{\Omega}$. Further denote $v(r) := \widetilde{c}\,'(r) = 4\pi\sin(4\pi r)$. Then we obtain

$$\begin{aligned}\nabla\cdot(n\nabla c_1) &= \nabla\cdot(\widetilde{n}v\nabla r)\\ &= (v\partial_r\widetilde{n} + \widetilde{n}v')\,\nabla r\cdot\nabla r + \widetilde{n}v\Delta r\\ &= v\partial_r\widetilde{n} + \widetilde{n}v' + \widetilde{n}vr^{-1}\,,\end{aligned}$$

and hence Eq. (2.6) is equivalent to $\partial_t\widetilde{n} + v\partial_r\widetilde{n} = -\left(v' + vr^{-1}\right)\widetilde{n}$ for $r > 0$. This equation can be solved by the Method of Characteristics (see Chap. A.1 in the appendix; we can restrict our attention to $r \in \left[0, \frac{1}{2}\right]$) and we obtain for smooth initial data $\widetilde{n}_0(r) := \widetilde{n}(0,r)$ and $r \neq 0, \frac{1}{4}, \frac{1}{2}$

$$\widetilde{n}(t,r) = \widetilde{n}_0(s(t,r))\frac{s(t,r)}{r}\frac{\sin(4\pi s(t,r))}{\sin(4\pi r)}\,, \tag{2.8}$$

with

$$s(t,r) = \frac{1}{2\pi}\arctan\left(\frac{\tan(2\pi r)}{\exp(16\pi^2 t)}\right) + \frac{\mathrm{int}(4r) + (\mathrm{int}(4r) \mod 2)}{4}\,,$$

where $\mathrm{int}(z)$ is the integer part of $z \in \mathbb{R}_{+,0}$. For $r = 0, \frac{1}{4}, \frac{1}{2}$ we obtain by continuity

$$\widetilde{n}(t,0) = \widetilde{n}_0(0)\exp(-16\pi^2 t),\ \widetilde{n}\left(t,\tfrac{1}{4}\right) = \widetilde{n}_0\left(\tfrac{1}{4}\right)\exp(16\pi^2 t),\ \widetilde{n}\left(t,\tfrac{1}{2}\right) = \widetilde{n}_0\left(\tfrac{1}{2}\right)\exp(-16\pi^2 t)\,.$$

Hence we have a smooth (classical) solution for all $t > 0$ whenever the initial data is differentiable. If the initial data is non-differentiable at some points then Eq. (2.8) can still be evaluated and is a generalized solution, see e.g. [41, p. 21].

2.3.2 Mathematical models related to tumour growth processes

The development of a primary solid tumour begins with a single normal cell becoming transformed as a result of mutations in certain key genes. This transformed cell differs from a normal one in several ways, one of the most notable being its escape from the body's homeostatic mechanisms,

leading to inappropriate proliferation. An individual tumour cell has the potential, over successive divisions, to develop into a cluster (or nodule) of tumour cells. Further growth and proliferation leads to the development of an avascular tumour consisting of approximately 10^6 cells. This cannot grow any further, owing to its dependence on diffusion as the only means of receiving nutrients and removing waste products. For any further development to occur the tumour must initiate angiogenesis – the recruitment of blood vessels. The tumour cells first secrete angiogenic factors which in turn induce endothelial cells in a neighbouring blood vessel to degrade their basal lamina and begin to migrate towards the tumour. As it migrates, the endothelium begins to form sprouts which can then form loops and branches through which blood circulates. From these branches more sprouts form and the whole process repeats forming a capillary network which eventually connects with the tumour, completing angiogenesis and supplying the tumour with the nutrients it needs to grow further. During the process of vessel formation, the endothelial cells may secrete enzymes which degrade the local tissue (extracellular matrix) thus facilitating the whole process. Indeed this process of matrix degradation through enzyme secretion is also carried out by the tumour cells themselves. This enables active migration by the tumour cells into the tissue to take place. The combination of tumour vascularization (i.e. the blood vessels connect with the tumour) and active tissue invasion means that there is now the possibility of tumour cells finding their way into the circulation and/or lymph system, and subsequently being deposited in distant sites in the body, resulting in metastasis.
The complete process of metastasis involves several sequential steps, each of which must be successfully completed by cells of the primary tumour before a secondary tumour (a metastasis) is formed. The mathematical models which we will present in the following two subsections focus specifically on the processes of tumour-induced angiogenesis (endothelial cell migration in response to external stimuli) and tumour invasion of tissue (cancer cell migration).

2.3.2.1 Mathematical models of tumour-induced angiogenesis (Models 2 and 3)

Angiogenesis, the formation of blood vessels from a pre-existing vasculature, is a crucial component of many mammalian growth processes, including embryogenesis and wound healing. It is also a key component in the metastatic cascade enabling a solid tumour to progress from the relatively harmless avascular growth phase to the potentially lethal vascular growth phase. In recent times a variety of models have appeared focusing on different aspects of the process. A comprehensive account of the complete angiogenic process may be found in [2] and references therein. We summarise the main events of angiogenesis as:

- the secretion of chemicals known as tumour angiogenic factors (TAF) by cancer cells,
- the response of endothelial cells (EC) in any neighbouring blood vessels to these chemicals through migration and proliferation,
- interaction between the ECs and the extracellular matrix (ECM),
- the formation of new individual capillary sprouts, which in turn connect up to form a new vasculature.

We describe below two models of tumour-induced angiogenesis developed by Chaplain and Stuart [8] and Anderson and Chaplain [2].

Model 2:
The model of Chaplain and Stuart [8] focused on two key variables, namely EC density n and TAF concentration c_1. The model assumed that TAF was secreted by tumour cells (located on one edge of the domain boundary), diffused into the surrounding tissue, was taken up by the ECs via cell-surface receptors and underwent some natural decay. The motion of the ECs was assumed to be influenced by two factors only: random motility (analogous to molecular diffusion) and chemotaxis in response to TAF gradients.
The random motility of the ECs is described by a diffusion term with constant $\varepsilon > 0$, the cell random motility coefficient. The chemotactic flux function was taken to be of the simple form $p_1 \equiv \chi_0$, where χ_0 is the (constant) chemotactic coefficient. The ECs were also assumed to undergo death at rate β and proliferation in a logistic manner, with proliferation constant μ. The latter was assumed to be governed by a threshold TAF concentration c_1^*, i.e. there was no proliferation for $c_1 < c_1^*$ and logistic proliferation for $c_1 > c_1^*$. The non-dimensionalized model equations are then given by:

$$\begin{aligned} \partial_t n &= \overbrace{\varepsilon\Delta n}^{\text{random motility}} - \overbrace{\nabla\cdot(\chi_0\, n\, \nabla c_1)}^{\text{chemotaxis}} + \overbrace{\max\{0, c_1 - c_1^*\}\mu n(1-n)}^{\text{proliferation}} - \overbrace{\beta n}^{\text{cell death}}, \\ \partial_t c_1 &= \overbrace{\Delta c_1}^{\text{diffusion}} - \overbrace{\frac{\alpha n c_1}{\gamma + c_1}}^{\text{uptake by cells}} - \overbrace{\lambda c_1}^{\text{decay}}. \end{aligned} \tag{2.9}$$

Chaplain and Stuart estimated the model parameter from experimental data and they are as follows:

$$\varepsilon = 0.001\,,\ \alpha = 10\,,\ \gamma = 1\,,\ \lambda = 1\,,\ \chi_0 = 0.75\,,\ \mu = 100\,,\ \beta = 4\,,\ c_1^* = 0.2\,.$$

The above parameter estimation shows that the cell random motility is much smaller than the chemotaxis. Hence, we also consider the system without cell random motility, i.e. $\varepsilon = 0$.
We consider this model on the unit square $\Omega = (0,1)^2$ in space. The initial TAF concentration is given by

$$c_1(0,\mathbf{x}) = \tfrac{1}{2}\cos\left(\tfrac{\pi}{2}x_1\right)\left(4 - 2x_1 + \cos\left(2\pi\left(\tfrac{1}{2} - x_2\right)\right)\right)\exp\left(-\left(1 - \cos\left(\tfrac{\pi}{2}x_1\right)\right)\right)\ .$$

This assumes a single tumour (TAF source) located on the left edge of the spatial domain ($x_1 = 0$). The boundary condition of c_1 is of no-flux type on the upper and lower boundary and of Dirichlet-type (time-independent and consistent with the initial data) on the left and right boundary. The initial EC density is zero in Ω except in five blocks near the right boundary $x_1 = 1$, where the initial values are one. These blocks have a width of 0.05 in the x_1-direction and a width 0.07 in the x_2-direction. Their centres are the points $(0.975, 0.2), (0.975, 0.36), (0.975, 0.5), (0.975, 0.64)$, and $(0.975, 0.8)$. This initial EC density assumes a parent blood vessel along the right boundary with five initial capillary sprouts developed already. The boundary condition of n on the right boundary is of Dirichlet-type (time-independent and consistent with the initial data). We assume a boundary condition of no-flux type for n on the remaining part of the boundary if $\varepsilon > 0$. If $\varepsilon = 0$ then no boundary condition is prescribed on the remaining boundary (outflow boundary). Fig. 2.2 gives plots of the initial data for EC density and TAF concentration.

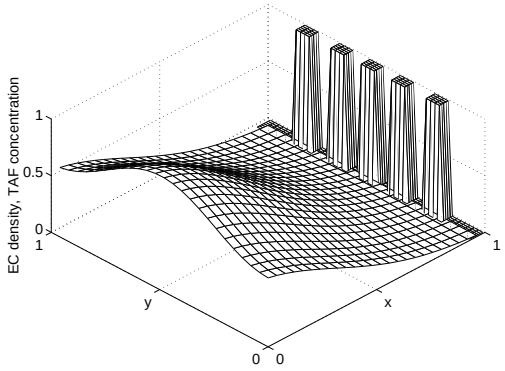

Figure 2.2: Initial conditions for Model 2. The smooth function is the initial TAF concentration c_1 and the function which is zero except for the blocks on the right boundary is the initial EC density n.

The final simulation time for the described setups are $T = 1.2$ for $\varepsilon = 0$ and $T = 1$ for $\varepsilon = 0.001$. Thereafter the assumptions underlying the model do not hold anymore because the blood vessels have reached the tumour and other processes take over.

Model 3:

The model of Anderson and Chaplain [2] extended and developed the model of Chaplain and Stuart [8] by including the interactions between the ECs and the ECM through the matrix macromolecule fibronectin. The model therefore consists of three equations governing the evolution of the three variables, EC density n, TAF concentration c_1, and fibronectin concentration c_2. Once again the model assumed that the motion of the ECs (at or near a capillary sprout-tip) is influenced by random motility and chemotaxis (in response to TAF gradients), but also by haptotaxis in response to fibronectin gradients.

Once again, the random motility of the ECs at or near the sprout-tips is described by a diffusion term with constant $\varepsilon > 0$, the cell random motility coefficient. However the chemotactic flux function p_1 was taken to depend on the TAF concentration, $p_1(c_1) = \frac{\chi_0}{1+\alpha c_1}$, (in contrast to being constant in Model 2), reflecting the saturation of TAF receptors on the cell surface. Finally, the influence of fibronectin on the ECs was modelled by the simple form of a constant haptotactic flux function $p_2 \equiv \rho_0$, where $\rho_0 > 0$ is the (constant) haptotactic coefficient. The model omitted any proliferation terms for the cells since in this case attention was focused on the ECs at the sprout-tips (where there is no proliferation).

The equation for fibronectin contained a degradation term (the ECs degrade the fibronectin via enzymes) and a production term since the ECs themselves produce and secrete fibronectin which then becomes bound to the ECM and does not diffuse. Therefore the equation for fibronectin contains no diffusion term ($D_2 = 0$).

Following the results of Chaplain and Stuart [8] where it was noted that the TAF diffusion occurred on such a fast timescale so as to set up a quasi-steady state concentration profile, the TAF equation contains only one term—that of uptake or binding of the TAF to the EC surface receptors. The initial quasi-steady state concentration profile is provided through the initial conditions for the TAF.

Hence the complete, non-dimensionalized system of equations describing the interactions of the ECs, TAF and fibronectin as modelled by Anderson and Chaplain [2] is

$$\begin{aligned}
\partial_t n &= \overbrace{\varepsilon\Delta n}^{\text{random motility}} - \overbrace{\nabla\cdot\left(\frac{\chi_0 n}{1+\alpha c_1}\nabla c_1\right)}^{\text{chemotaxis}} - \overbrace{\nabla\cdot(\rho_0 n\,\nabla c_2)}^{\text{haptotaxis}}, \\
\partial_t c_1 &= -\overbrace{\eta\, n c_1}^{\text{uptake}}, \\
\partial_t c_2 &= \overbrace{\beta n}^{\text{production}} - \overbrace{\gamma\, n c_2}^{\text{degradation}}.
\end{aligned} \tag{2.10}$$

Anderson and Chaplain [2] estimated as many parameter values as possible from experimental data and used the following set in their simulations

$$\varepsilon = 0.00035\,,\ \chi_0 = 0.38\,,\ \rho_0 = 0.34\,,\ \alpha = 0.6\,,\ \beta = 0.05\,,\ \gamma = 0.1\,,\ \eta = 0.1\,.$$

Although Anderson and Chaplain [2] considered random migration of the ECs, here we also consider the system without this random motion, i.e. $\varepsilon = 0$. This can be justified biologically: prior to stimulation by the TAF, the ECs are migrationally inert and are simply attached to one another while lining their parent vessel. Also we can see from the estimated parameter values ($\varepsilon = 0.00035$, $\chi_0 = 0.38$, $\rho_0 = 0.34$) that the (scaled) random migration coefficient of the ECs is several orders of magnitude smaller than the taxis coefficients.
We consider the model on the unit square $\Omega = (0,1)^2$ in space with the parent vessel located along the left edge, $x_1 = 0$, and the (circular) tumour located on the opposite edge, $x_1 = 1$. We assume that three (initially separated) capillary sprouts have formed already near $x_1 = 0$. Let $r^2 = (x_1 - 1)^2 + \left(x_2 - \frac{1}{2}\right)^2$. The initial conditions are depicted in Fig. 2.3 and are given by

$$\begin{aligned}
n(0,\mathbf{x}) &= \exp\left(-\frac{x_1^2}{0.001}\right)\max\left\{0, \sin\left(\pi\left(6x_2 - \frac{1}{2}\right)\right)\right\}^2, \\
c_1(0,\mathbf{x}) &= \begin{cases} 1\,, & 0 \le r \le 0.1\,, \\ \left(\frac{\nu - r}{\nu - 0.1}\right)^2, & 0.1 \le r \le 1, \end{cases} \text{, where } \nu := \frac{\sqrt{5} - 0.1}{\sqrt{5} - 1}\,, \\
c_2(0,\mathbf{x}) &= \frac{3}{4}\exp\left(-\frac{x_1^2}{0.45}\right).
\end{aligned}$$

It is assumed that the cells, and consequently the capillary sprouts, remain within the domain Ω and therefore no-flux boundary conditions for n are imposed on the boundaries of the Ω. We consider a final time $T = 10$ for this model.

2.3.2.2 A mathematical model of tumour invasion (Model 4)

A crucial part of the metastatic process is the ability of the cancer cells to degrade the surrounding tissue or extracellular matrix (ECM). The matrix is highly dynamic, at any one time being actively secreted and degraded. A number of specific matrix degradative enzymes (MDEs) have been described and have been repeatedly implicated in all of the key steps of tumour invasion and

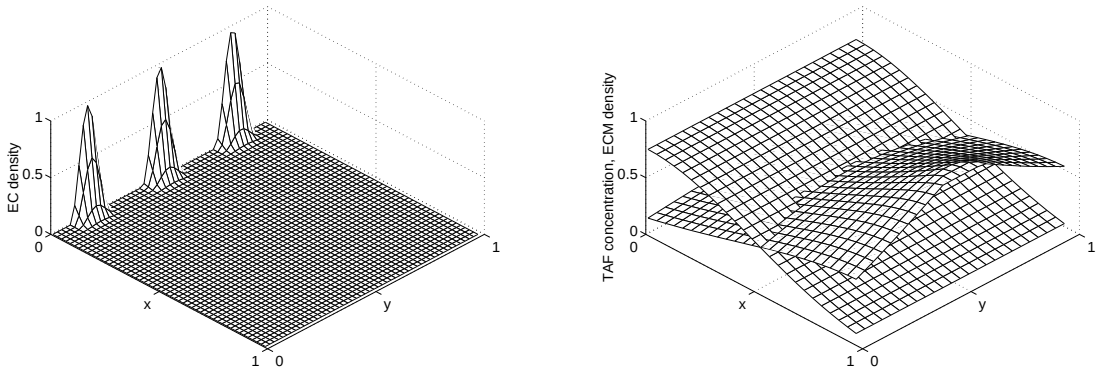

Figure 2.3: Initial conditions for Model 3. The initial EC density n is depicted in the left plot and both, the initial TAF concentration c_1 and the the initial fibronectin concentration c_2, are in the right plot.

metastasis. A comprehensive description of the invasive process and its place in the metastatic cascade is given in Anderson et al. [3] and references therein. We describe below the recent model of Anderson et al. [3] who developed a mathematical model of tumour invasion based on generic solid tumour growth, which for simplicity was assumed to be in an avascular state.

In the model three variables were considered: tumour cell density n, ECM density c_1, and MDE concentration c_2. The main assumptions of the model were that the tumour cells produce MDEs which degrade the ECM locally and that ECM degradation results in the production of molecules which are actively attractive to tumour cells (e.g. fibronectin) and which then aid in directed tumour cell motility (haptotaxis).

The model considered tumour cell motion to be driven only by random motility and haptotaxis in response to adhesive and/or attractive gradients created by degradation of the matrix. To describe the random motility of the tumour cells a diffusion term with random motility coefficient $\varepsilon > 0$ is assumed. (Anderson et al. [3] additionally considered nonlinear diffusion but here we only consider the linear case.) The haptotactic flux functions p_1 was taken to be of the simple form $p_1 \equiv \rho_0$, where $\rho_0 > 0$ is the (constant) haptotactic coefficient. The model did not consider any proliferation of the tumour cells.

Active MDEs were assumed to be produced by the tumour cells, diffuse throughout the tissue and undergo some form of decay (either passive or active). The ECM was assumed to have no motility and was degraded by the MDEs upon contact.

Hence the complete system of equations from the model of Anderson et al. [3] describing the interactions of the tumour cells, ECM and MDEs is

$$\begin{aligned} \partial_t n &= \overbrace{\varepsilon \Delta n}^{\text{random motility}} - \overbrace{\nabla \cdot (n \rho_0 \nabla c_1)}^{\text{haptotaxis}}, \\ \partial_t c_1 &= - \overbrace{\eta\, c_2 c_1}^{\text{degradation}}, \\ \partial_t c_2 &= \overbrace{d_2 \Delta c_2}^{\text{diffusion}} + \overbrace{\alpha n}^{\text{production}} - \overbrace{\beta c_2}^{\text{decay}}. \end{aligned} \tag{2.11}$$

Anderson et al. [3] undertook a range of simulation experiments based around the following set of

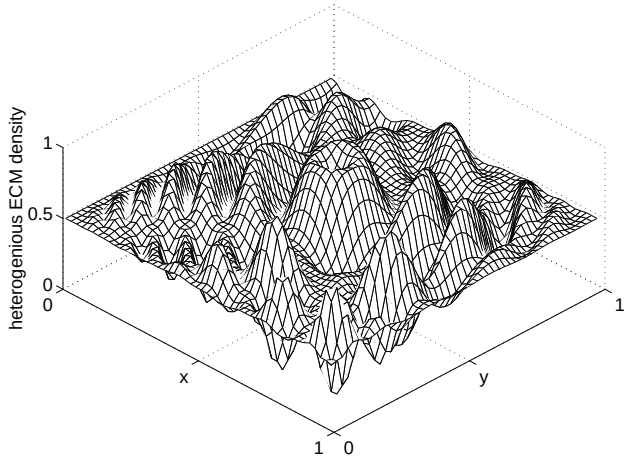

Figure 2.4: Initial heterogeneous ECM density c_1 used in Model 4.

parameter values:

$$\varepsilon = 0.001\,,\ d_2 = 0.001\,,\ \rho_0 = 0.005\,,\ \eta = 10\,,\ \alpha = 0.1\,,\ \beta = 0.5\,.$$

Prior to invasion, the tumour is a compact mass of cells with little or no local migration. Once invasion is triggered, the migration of the cells is very focused and direct. Hence we also consider the above system without random motion of the tumour cells i.e. $\varepsilon = 0$.
We study the tumour invasion model in two spatial dimensions on the unit square $\Omega = (0,1)^2$ and assume that a tumour is situated in the centre of the domain. In contrast to the case of only one spatial dimension, this enables us to consider the effect of spatial heterogeneity explicitly. In particular we can consider a heterogeneous ECM density which is more representative of real tissue. To this end a hypothetical heterogeneous initial ECM density c_1 is used as depicted in Fig. 2.4, [3]. The initial condition for tumour cell density n and MDE concentration c_2 are chosen as

$$n(0,\mathbf{x}) = \begin{cases} \exp\left(-\frac{r(\mathbf{x})^2}{0.0025}\right)\,, & r \in [0,0.1] \\ 0\,, & r > 0.1 \end{cases}\,,$$
$$c_2(0,\mathbf{x}) = \frac{1}{2}n(0,\mathbf{x})\,,$$

where $r(\mathbf{x})^2 = (x_1 - 0.5)^2 + (x_2 - 0.5)^2$.
We assume that tumour cells and MDEs remain within Ω and therefore impose no-flux boundary conditions for n and c_2 on the boundary. We consider a final simulation time $T = 15$ in the numerical simulations.

Chapter 3

The Method of Lines and Space Discretization

We obtain numerical approximations of the solution of problem (2.5) by following the *Method of Lines* (MOL). This approach is widely used in the numerical solution of PDEs and means that we consider the discretization of the spatial operators and the time integration separately. The space domain Ω is covered by a grid which defines computational cells (grid cells). We emphasize that these grid cells should not be confused with biological cells. The spatial derivatives in the system (2.5) are then discretized on the grid by using approximate average values of n and $\mathbf{c}$ in the grid cells (following the finite volume methodology). We describe the grid in Sec. 3.1 and the discretization of the various terms in the right-hand side of Eq. (2.5) in Sec. 3.3. The result of this procedure is an initial value problem (IVP) for a huge system of stiff, nonlinear ODEs (one ODE for each grid cell and component of Eqs. (2.5a,2.5b)), the so-called MOL-ODE. It is the aim of this chapter to obtain a MOL-ODE which is a suitable approximation of the PDE model (2.5). The time integration of this ODE will be discussed in the next chapter.

For our application it is important that the solution methods preserve the positivity of an exact solution in its numerical approximation. This means that a numerical solution remains nonnegative for all $t \in I_T$ whenever the initial data is nonnegative and the exact solution is nonnegative in I_T (this is ensured by our assumption (A2)). Violating positivity is highly undesirable because it may turn stable reactions $(p_0, \mathbf{g}_0)$ into unstable ones which in turn may lead to numerical instabilities. This especially shows up with the logistic source term in Model 2, and we discuss this problem in the presentation of our numerical results in Chap. 5. For this reason, we will require that the spatial discretization results in a MOL-ODE with a nonnegative analytical solution whenever the initial values are nonnegative. In Sec. 3.2 we give conditions on the right-hand side of an ODE system which guarantee positivity of the exact solution (*positive ODE systems*).

In Sec. 3.3 we describe an appropriate *Finite Volume Method* (FVM) to discretize the system (2.5) in space. Finite volume schemes appear to be very suitable for the problem class under consideration because they are based on the conservation form Eq. (2.4) and the numerical schemes obtained with this approach are naturally conservative, i.e. no mass is produced or used up "by the scheme". We also pay special attention to the discretization near the domain boundary and show that the resulting discretization in space results in a positive ODE system. Finally, in Sec. 3.4, we evaluate the spatial discretization of the taxis term in Eq. (2.5) by considering the solution of Model 1.

3.1 Spatial grid

Let $\Omega = (0,1)^d$ be the spatial domain of our problem class as defined in Sec. 2.2. We use an equidistant grid having grid width $h := \frac{1}{M}$, $M \in \mathbb{N}$, in each spatial direction and define grid points $\mathbf{x_i}$, where $\mathbf{i} := (i_1, \ldots, i_d) \in \mathcal{I} := \{1, 2, \ldots, M\}^d$ is a d-dimensional multi-index, by

$$\mathbf{x_i} := \left(\left(i_1 - \frac{1}{2} \right) h, \ldots, \left(i_d - \frac{1}{2} \right) h \right)^{\mathsf{T}} .$$

Each grid point $\mathbf{x_i}$ is the centre of an associated control volume $\Omega_\mathbf{i}$ defined by

$$\Omega_\mathbf{i} := ((i_1 - 1)\, h, i_1 h) \times ((i_2 - 1)\, h, i_2 h) \times \cdots \times ((i_d - 1)\, h, i_d h) \ .$$

The set of all control volumes, $\{\Omega_\mathbf{i}\}_{\mathbf{i} \in \mathcal{I}}$, forms a partition of the domain Ω, [4]. For notational convenience it is useful to assume that we have also control volumes outside of Ω which we define and denote in an analogous fashion as the elements of the partition itself.

3.2 Positivity of the spatial discretization

We stated in the introduction that the result of the spatial discretization of (2.5) is an IVP for a huge system of stiff, nonlinear ODEs in $\mathbb{R}^m$. We denote this IVP in the same form as later used in Chap. 4:

$$y'(t) = F(t, y(t)), \quad t \geq t_0 \in \mathbb{R}, \quad y(t_0) = y_0 \in \mathbb{R}^m \, . \tag{3.1}$$

This IVP represents the semi-discretization of problem (2.5) as a result of the first step of the MOL. The vector $y(t)$ contains the (time-continuous) approximations to the averages of the solution of system (2.5) in all grid cells. We have already emphasized that we are seeking approximate solutions which are nonnegative. We make this requirement precise. Let F have the property (see [27])

$$F \text{ is continuous and (3.1) has a unique uncontinuable solution for all } t_0 \in \mathbb{R} \text{ and all } y_0 \in \mathbb{R}^m. \tag{3.2}$$

We can now define the terms positive ODE system and positive semi-discretization.

Definition 1 *The ODE system in (3.1) as well as the IVP (3.1) are called* positive *if F has the property (3.2) and $y(t) \geq 0$ holds for all $t \geq t_0$ whenever $t_0 \in \mathbb{R}$ and $y_0 \geq 0$.*
If a semi-discretization of a given PDE results in a positive MOL-ODE then this semi-discretization is called positive.

The following theorem from [27] characterizes positive ODE systems (see also [31]).

Theorem 1 *Let F satisfy condition (3.2). The IVP (3.1) corresponding to F is positive if and only if for all $t \in \mathbb{R}$ and any vector $v \in \mathbb{R}^m_{+,0}$ and all $i = 1(1)m$ holds*

$$v_i = 0 \quad \Rightarrow \quad F_i(t, v) \geq 0 \, .$$

We denote with $\mathcal{P}$ the class of functions F for which the corresponding IVP (3.1) is positive.
The right-hand side of the MOL-ODE of our model system (2.5) will be a sum of functions. Therefore the following corollary is useful.

Corollary 1 *If $F, G \in \mathcal{P}$ then for all $\alpha, \beta \in \mathbb{R}_{+,0}$ holds $\alpha F + \beta G \in \mathcal{P}$.*

If the right-hand side F of an ODE is linear then we can easily check positivity of this ODE.

Corollary 2 *Let $F(t, y) = Py + g(t)$ with a continuous function g satisfying $g(t) \geq 0$ for all $t \in \mathbb{R}$ and a matrix $P \in \mathbb{R}^{m,m}$. Then holds $F \in \mathcal{P}$ if and only if the off-diagonal elements of P are nonnegative.*

3.3 A semi-discrete finite volume method

After the preparations in the previous two sections we are now going to describe the finite volume discretization in space of problem (2.5). However, to avoid too difficult notation, the description will not be given for the problem (2.5) itself but for a scalar equation. Consider for a scalar function $u : \bar{I}_T \times \bar{\Omega} \to \mathbb{R}$ the PDE

$$\partial_t u = D\Delta u - \nabla \cdot \left(u \sum_{j=1}^{l} p_j(\mathbf{c}) \nabla c_j \right) + p_0(t, \mathbf{x}, u, \mathbf{c}) \,, \quad \text{for } (t, \mathbf{x}) \in I_T \times \Omega \,, \tag{3.3}$$

where $\mathbf{c} : \bar{I}_T \times \bar{\Omega} \to \mathbb{R}^l$, $p_j : \mathbb{R}^l \to \mathbb{R}$ for $j = 1(1)l$, and $p_0 : I_T \times \Omega \times \mathbb{R}^{l+1} \to \mathbb{R}$ are given functions and $D \geq 0$ is a constant. We describe the discretization of the right-hand side of this problem on the spatial grid defined in Sec. 3.1. The scalar equation (3.3) can be regarded as a prototype equation for the model system (2.5) because it contains all the relevant terms of (2.5) – taxis, diffusion, and reaction. The application of the method to problem (2.5) is then straightforward.
Let Ω_i be an element of an partition $\{\Omega_i\}_{i \in \mathcal{I}}$ of Ω (not necessarily the partition described in Sec. 3.1), and denote the (time-continuous) cell average of u over Ω_i by $\bar{u}_i(t)$,

$$\bar{u}_i(t) := \frac{1}{|\Omega_i|} \int_{\Omega_i} u(t, \mathbf{x}) d\mathbf{x} \,, \quad i \in \mathcal{I} \,.$$

The integral form of the conservation law (3.3) is the starting point for the finite volume scheme. We know from Eq. (2.4) that the evolution of the averages $\bar{u}_i(t)$ is exactly governed by

$$\frac{\mathrm{d}\,\bar{u}_i}{\mathrm{d}\,t} = \underbrace{\frac{1}{|\Omega_i|} \oint_{\partial\Omega_i} D\nabla u \cdot \mathfrak{n} d\partial\Omega_i}_{=:\, H_D(u(t,\cdot); i)} \underbrace{- \frac{1}{|\Omega_i|} \oint_{\partial\Omega_i} u \left(\sum_{j=1}^{l} p_j(\mathbf{c}) \nabla c_j \right) \cdot \mathfrak{n} d\partial\Omega_i}_{=:\, H_T(u(t,\cdot); i)} + \underbrace{\frac{1}{|\Omega_i|} \int_{\Omega_i} p_0(t, \mathbf{x}, u, \mathbf{c}) d\mathbf{x}}_{=:\, H_R(u(t,\cdot); i)} . \tag{3.4}$$

We denote with $H(u(t,\cdot); i)$ the right-hand side of the exact cell average equation (3.4). It may depend on u at time t in the whole domain Ω (we denote this by $u(t,\cdot)$), and also on $\mathbf{c}(t,\cdot), t$, and $\mathbf{x} \in \Omega_i$. Further, H_D, H_T, and H_R are the parts of H corresponding to diffusion, taxis, and reaction, respectively. In short we write for the above equation

$$\frac{\mathrm{d}}{\mathrm{d}\,t} \bar{u}_i(t) = H(u(t,\cdot); i) \equiv H_D(u(t,\cdot); i) + H_T(u(t,\cdot); i) + H_R(u(t,\cdot); i) \,, \; i \in \mathcal{I} \,. \tag{3.5}$$

Henceforth we use the following notations. With $U_i(t)$ we denote an approximation to the cell average $\bar{u}_i(t)$ and collect the approximations for all cells of the partition in the vector $\mathbf{U}(t)$. Similarly, $\bar{\mathbf{u}}(t)$ denotes the vector of the exact cell averages $\bar{u}_i(t)$. Also, for each concentration $c_j(t,\mathbf{x}), j = 1(1)l$, we denote with $\bar{c}_{j,i}(t), C_{j,i}(t), \bar{\mathbf{c}}_j(t)$, and $\mathbf{C}_j(t)$ the exact and approximate cell average in grid cell i, and the vectors of these quantities for all elements of the partition. Finally, denote with $\mathbf{C}_{\cdot,i} := [C_{1,i}, C_{2,i}, \ldots, C_{l,i}]^\mathsf{T}$ and $\bar{\mathbf{c}}_{\cdot,i} := [\bar{c}_{1,i}, \bar{c}_{2,i}, \ldots, \bar{c}_{l,i}]^\mathsf{T}$ the vector of all approximate and exact concentration averages in grid cell i.

The idea of the finite volume approach is to approximate the right-hand side of Eq. (3.5) by using cell averages of u in neighbouring cells of Ω_i. Let $\mathcal{H}(\mathbf{U}(t); i)$ be an approximation to $H(u(t,\cdot); i)$ which depends on a *finite* number of elements of $\mathbf{U}(t)$ (and possibly on the time t, space points $\mathbf{x} \in \Omega_i$, and on a finite number of components of $\mathbf{C}_j(t), j = 1(1)l$). Then we obtain an ODE system for the evolution of the approximate cell averages

$$\frac{\mathrm{d}}{\mathrm{d}\,t} U_i(t) = \mathcal{H}(\mathbf{U}(t); i)\,, \quad i \in \mathcal{I}\,, \tag{3.6}$$

the so-called MOL-ODE. The initial values for this ODE are provided as approximations $\mathbf{U}_0$ to the cell averages of a given initial condition for u in Ω.

In the next subsections we describe the construction of appropriate approximations $\mathcal{H}_D$, $\mathcal{H}_T$, and $\mathcal{H}_R$ to H_D, H_T, and H_R, respectively. Finally, Sec. 3.3.4 deals with the special requirements for the approximations in cells Ω_i close to the boundary of Ω. However, before going into details, we shortly review some important concepts for the analysis and evaluation of the discretization in space.

In Sec. 3.2 we have already given a condition which the ODE system (3.6) must satisfy to be a positive ODE system. We will ensure that this condition holds for the approximations $\mathcal{H}_D$, $\mathcal{H}_T$, and $\mathcal{H}_R$ and hence also for the sum $\mathcal{H}$.

We now look at the spatial accuracy of (3.6) with respect to (3.5). Therefore we define for each grid cell $\Omega_i, i \in \mathcal{I}$, of the partition the global error $e_i(t)$ in the exact solution of the ODE system (3.6) with respect to the average over Ω_i of the exact solution of (3.5), i.e.

$$e_i(t) := U_i(t) - \bar{u}_i(t)\,,$$

and further the local truncation error in cell i, $lte(t; i)$, as the difference between the discrete operator $\mathcal{H}$ and the exact operator H applied to a smooth solution $u(t,\mathbf{x})$ of (3.5), i.e.

$$lte(t; i) := \mathcal{H}(\bar{\mathbf{u}}(t); i) - H(u(t,\cdot); i)\,.$$

Subtracting (3.5) from (3.6) and adding $0 = \mathcal{H}(\bar{\mathbf{u}}(t); i) - \mathcal{H}(\bar{\mathbf{u}}(t); i)$ yields the error equation

$$\frac{\mathrm{d}}{\mathrm{d}\,t} e_i(t) = lte(t; i) + \left(\mathcal{H}(\mathbf{U}(t); i) - \mathcal{H}(\bar{\mathbf{u}}(t); i)\right)\,. \tag{3.7}$$

We have the following estimate for the norm of the global error vector $\mathbf{e}(t)$, see also [41, p. 196]. We consider the discrete L^1-norm of vectors $\mathbf{v} = (v_i)_{i\in\mathcal{I}}$ (on a partition $\{\Omega_i\}_{i\in\mathcal{I}}$ of Ω) defined by $\|\mathbf{v}\|_1 := \sum_{i\in\mathcal{I}} |\Omega_i||v_i|$. Here $|\Omega_i|$ is the Lebesgue measure of Ω_i (and $|\Omega|$ the measure of Ω).

Theorem 2 *Let h be the maximum diameter of the grid cells Ω_i of a partition $\{\Omega_i\}_{i\in\mathcal{I}}$ of Ω. If the approximation $\mathcal{H}(\mathbf{U}(t);i)$ is Lipschitz continuous, i.e. for each $i\in\mathcal{I}$ we have for all $\mathbf{U}_1,\mathbf{U}_2$ that*

$$|\mathcal{H}(\mathbf{U}_1;i)-\mathcal{H}(\mathbf{U}_2;i)|\le L\|\mathbf{U}_1-\mathbf{U}_2\|_1,$$

with constant $L\in\mathbb{R}_+$ independent of $i\in\mathcal{I}$ and $t\in\bar{I}_T$, and if the local truncation error is of order p in h, i.e. there exists a constant $K\in\mathbb{R}_+$ (independent of $i\in\mathcal{I}$ and $t\in\bar{I}_T$) such that for all $i\in\mathcal{I}$

$$|lte(t;i)|\le Kh^p,$$

then the global error satisfies for $t\in\bar{I}_T$

$$\|\mathbf{e}(t)\|_1\le\|\mathbf{e}(0)\|_1\exp(t|\Omega|L)+\frac{K}{L}\left(\exp(t|\Omega|L)-1\right)\cdot h^p.$$

This implies, if the initial error satisfies $\|\mathbf{e}(0)\|_1=\mathcal{O}\left(h^p\right)$, that the semi-discrete approximation (3.6) is pth order accurate, i.e. $\|\mathbf{e}(t)\|_1=\mathcal{O}\left(h^p\right)$ on $\bar{I}_T$.

For the proof of this theorem we use a Gronwall lemma.

Lemma 1 *[1, p. 99] Let $I:=[t_0,T]\subset\mathbb{R}$ be an interval, and $y(t),h(t)$ and $M(t)$ be scalar, continuous and nonnegative functions on I. If $y(t)$ satisfies*

$$y(t)\le h(t)+\int_{t_0}^{t}M(\tau)y(\tau)d\tau\quad\textit{for all } t\in I$$

then

$$y(t)\le h(t)+\int_{t_0}^{t}\exp\left(\int_{\tau}^{t}M(\sigma)d\sigma\right)M(\tau)h(\tau)d\tau\quad\textit{for all } t\in I.$$

Proof (**of Theorem 2**) All relations in this proof hold for all $t\in\bar{I}_T$. We obtain from the error equation (3.7) the equality

$$e_i(t)=e_i(0)+\int_0^t\left[lte(\tau;i)+(\mathcal{H}(\mathbf{U}(\tau);i)-\mathcal{H}(\bar{\mathbf{u}}(\tau);i))\right]d\tau,$$

and using the assumptions of the theorem

$$|e_i(t)|\le|e_i(0)|+Kh^pt+\int_0^t L\|\mathbf{e}(\tau)\|_1\,d\tau.$$

Multiplying this equation with $|\Omega_i|$ and summing over all $i\in\mathcal{I}$ yields

$$\|\mathbf{e}(t)\|_1\le\|\mathbf{e}(0)\|_1+|\Omega|Kh^pt+\int_0^t|\Omega|L\|\mathbf{e}(\tau)\|_1\,d\tau.$$

Now we can apply the Gronwall Lemma 1 with $y(t):=\|\mathbf{e}(t)\|_1$, $h(t):=\|\mathbf{e}(0)\|_1+|\Omega|Kh^pt$ and $M(t):=|\Omega|L$. The statement of the theorem follows by evaluating the integral in the inequality obtained from this lemma. □

Whereas the discussion in this section up to here is valid for any partition $\{\Omega_i\}_{i\in\mathcal{I}}$ of a bounded, nonempty spatial domain Ω, we consider henceforth the domain and partition defined in Sec. 3.1. Hence Ω is a d-dimensional unit cube and we refer to the elements of the partition $\{\Omega_{\mathbf{i}}\}_{\mathbf{i}\in\mathcal{I}}$ with d-dimensional multi-indices $\mathbf{i}$.
If $p_0 \equiv 0$ in Eq. (3.3) then we obtain that the total mass of the quantity with density u in Ω changes only through boundary fluxes and if these fluxes are zero then this total mass is conserved. We would like to have that for the total mass of the solution of (3.6) a discrete conservation property is true. Therefore we consider for $\mathcal{H} = \mathcal{H}_D$ or $\mathcal{H} = \mathcal{H}_T$ discretizations in *conservation form*, that is

$$\mathcal{H}(\mathbf{U}(t);\mathbf{i}) := -\frac{1}{h}\sum_{j=1}^{d}\left(\mathcal{F}_j(\mathbf{U}(t);\mathbf{i}) - \mathcal{F}_j(\mathbf{U}(t);\mathbf{i}-\mathbf{e}_j)\right), \tag{3.8}$$

where $\mathcal{F}_j(\mathbf{U}(t);\mathbf{i})$ approximates the average of the (diffusive or tactic) flux from $\Omega_{\mathbf{i}}$ to $\Omega_{\mathbf{i}+\mathbf{e}_j}$ through the common cell face of $\Omega_{\mathbf{i}}$ and $\Omega_{\mathbf{i}+\mathbf{e}_j}$. From the definition of $\mathcal{H}$ we see that the d-dimensional problem is essentially broken down to d one-dimensional problems due to the special structure of our partition of Ω. We note that we make use of the auxiliary grid cells outside of Ω in the notation of Eq. (3.8). Now, summing Eq. (3.8) over all $\mathbf{i} \in \mathcal{I}$, we see that on the right-hand side all terms cancel except for those $\mathcal{F}_j(\mathbf{U}(t);\mathbf{i})$ which approximate fluxes through the boundary of Ω. If these boundary fluxes are zero then we obtain with Eq. (3.6)

$$\frac{\mathrm{d}}{\mathrm{d}\,t}\sum_{\mathbf{i}\in\mathcal{I}} U_{\mathbf{i}}(t) = 0,$$

and this means that the total mass of the initial data is conserved.
In the following we require that the flux approximations $\mathcal{F}_j(\mathbf{U}(t);\mathbf{i})$ are Lipschitz continuous. Further, we omit the time-dependence of approximations and simply write e.g. $U_{\mathbf{i}}$ instead of $U_{\mathbf{i}}(t)$. We now discuss the approximation of taxis, diffusion, and reaction part of Eq. (3.5) in turn.

3.3.1 Taxis

We give an approximation $\mathcal{H}_T(\mathbf{U};\mathbf{i})$ in conservation form to the taxis part $H_T(u(t,\cdot);\mathbf{i})$ in Eq. (3.5) in this section and mainly follow the ideas of Hundsdorfer et al. [31], see also the paper by Sweby [57] and the book by LeVeque [41]. We start with the conservative formula

$$\mathcal{H}_T(\mathbf{U};\mathbf{i}) := -\frac{1}{h}\sum_{j=1}^{d}\left(\mathcal{T}_j(\mathbf{U};\mathbf{i}) - \mathcal{T}_j(\mathbf{U};\mathbf{i}-\mathbf{e}_j)\right), \tag{3.9}$$

where the function $\mathcal{T}_j(\mathbf{U};\mathbf{i})$ approximates the average of the tactic flux $u\left(\sum_{k=1}^{l} p_k(\mathbf{c})\partial_{x_j} c_k\right)$ from grid cell $\Omega_{\mathbf{i}}$ to $\Omega_{\mathbf{i}+\mathbf{e}_j}$ through their common cell face. We follow the state interpolation approach to define the approximations $\mathcal{T}_j(\mathbf{U};\mathbf{i})$. A possible flux interpolation approach (for a specific TDR system) is described in [14].
We make the approximation $\mathcal{T}_j(\mathbf{U};\mathbf{i})$ on a given cell face dependent on the sign of the local velocity perpendicular to this face, that is dependent on the flow direction (upwinding). Upwinding is a standard technique in the discretization of advection terms and the taxis term in our problem

class can be regarded as advection. In our case, the velocity in spatial direction j is given by $\sum_{k=1}^{l} p_k(\mathbf{c})\partial_{x_j} c_k$. Let $v_{\mathbf{i},j}$ denote an approximate average of this velocity on the common cell face of $\Omega_{\mathbf{i}}$ and $\Omega_{\mathbf{i}+\mathbf{e}_j}$. We set

$$v_{\mathbf{i},j} := \sum_{k=1}^{l} p_k \left(\frac{\mathbf{C}_{\cdot,\mathbf{i}} + \mathbf{C}_{\cdot,\mathbf{i}+\mathbf{e}_j}}{2} \right) \frac{C_{k,\mathbf{i}+\mathbf{e}_j} - C_{k,\mathbf{i}}}{h}\,, \tag{3.10}$$

and define the sign-dependent, approximate tactic flux by

$$\mathcal{T}_j(\mathbf{U};\mathbf{i}) := \max\{0, v_{\mathbf{i},j}\}\mathcal{S}_j^+(\mathbf{U};\mathbf{i}) + \min\{0, v_{\mathbf{i},j}\}\mathcal{S}_j^-(\mathbf{U};\mathbf{i}). \tag{3.11}$$

Here, $\mathcal{S}_j^+(\mathbf{U};\mathbf{i})$ and $\mathcal{S}_j^-(\mathbf{U};\mathbf{i})$ are the state interpolants. They approximate the average value (state) of u on the common cell face of $\Omega_{\mathbf{i}}$ and $\Omega_{\mathbf{i}+\mathbf{e}_j}$. If we choose the state interpolants to be linear combinations of components of $\mathbf{U}$ then we can achieve approximation order greater than one but the resulting discretizations would not be positive and oscillations are introduced into the solution, see [31, 13]. We want to combine a higher approximation order with positivity and therefore use so-called limiter functions $\Phi(r)$ in the definition of the state interpolants:

$$\mathcal{S}_j^+(\mathbf{U};\mathbf{i}) := \begin{cases} U_{\mathbf{i}} + \frac{1}{2}\Phi(r_{\mathbf{i},j})(U_{\mathbf{i}} - U_{\mathbf{i}-\mathbf{e}_j}) & \text{for } U_{\mathbf{i}} - U_{\mathbf{i}-\mathbf{e}_j} \neq 0 \\ U_{\mathbf{i}} & \text{otherwise,} \end{cases} \tag{3.12a}$$

$$\mathcal{S}_j^-(\mathbf{U};\mathbf{i}) := \begin{cases} U_{\mathbf{i}+\mathbf{e}_j} + \frac{1}{2}\Phi(r^{-1}_{\mathbf{i}+\mathbf{e}_j,j})(U_{\mathbf{i}+\mathbf{e}_j} - U_{\mathbf{i}+2\mathbf{e}_j}) & \text{for } U_{\mathbf{i}+\mathbf{e}_j} - U_{\mathbf{i}+2\mathbf{e}_j} \neq 0 \\ U_{\mathbf{i}+\mathbf{e}_j} & \text{otherwise.} \end{cases} \tag{3.12b}$$

The limiter function Φ depends on a *smoothness monitor function* r. We define this smoothness monitor for our grid function $\mathbf{U}$ by

$$r_{\mathbf{i},j} := \frac{U_{\mathbf{i}+\mathbf{e}_j} - U_{\mathbf{i}}}{U_{\mathbf{i}} - U_{\mathbf{i}-\mathbf{e}_j}}\,. \tag{3.13}$$

We see that $r_{\mathbf{i},j} \approx 1$ in smooth, monotone regions of $\mathbf{U}$ along the jth coordinate direction and $r_{\mathbf{i},j} < 0$ if $U_{\mathbf{i}}$ is a local extrema of $\mathbf{U}$ in the jth coordinate direction. The stencils of the approximate taxis flux $\mathcal{T}_j(\mathbf{U};\mathbf{i})$ depending on the local velocity are depicted in Fig. 3.1.

Figure 3.1: Stencils of $\mathcal{T}_j(\mathbf{U};\mathbf{i})$ for positive (left) and negative (right) local velocity $v_{\mathbf{i},j}$.

We require that the limiter function Φ is Lipschitz continuous (such that the resulting taxis discretization will also be Lipschitz continuous) and has the following properties (with $\delta > 0$, a free parameter):

$$\Phi(1) = 1\,, \tag{3.14a}$$

$$\Phi(r) = 0 \text{ for } r \leq 0\,, \quad 0 \leq \Phi(r) \leq \delta\,, \quad \text{and} \quad \Phi(r) \leq 2r \text{ for } r > 0\,. \tag{3.14b}$$

The property (3.14a) is important for the order of the discretization and the properties in (3.14b) are sufficient for positivity, see Lemmas 2 and 3. Increasing the value of δ improves the accuracy of the discretization near peaks in the solution, see [31]. A good choice is $\delta = 2$.
In order to make the definition of the discretization complete, we need a limiter function Φ and also a strategy for the treatment of cells close to the domain boundary. The first topic will be discussed in the end of this subsection after the presentation of the following two lemmas on accuracy and positivity of the state interpolation approach. The second topic is discussed for the problem class (2.5) in Sec. 3.3.4.

Lemma 2 *Let $u(t, \mathbf{x})$ and $c_j(t, \mathbf{x}), j = 1(1)l$, be smooth functions and regard $\mathbf{U}$ and $\mathbf{C}_j$ as point approximations to u and c_j in the centres $\mathbf{x}_\mathbf{i}$ of the grid cells $\Omega_\mathbf{i}$, i.e. $U_\mathbf{i} = u(t, \mathbf{x}_\mathbf{i})$ and $C_{j,\mathbf{i}} = c_j(t, \mathbf{x}_\mathbf{i})$. Consider grid cells $\Omega_\mathbf{i}$, $\mathbf{i} \in \{3, 4, \ldots, M-2\}^d$, (sufficiently far away from the boundary $\partial\Omega$ so that we can apply the state interpolation approach).*
If the limiter function Φ is Lipschitz continuous and if $\Phi(1) = 1$ then the local truncation error (in a pointwise sense) of the the state interpolation approach is second-order in the grid width h, i.e.

$$\mathcal{H}_T(\mathbf{U}; \mathbf{i}) + \nabla \cdot \left(u(t, \mathbf{x}_\mathbf{i}) \sum_{k=1}^{l} p_k(\mathbf{c}(t, \mathbf{x}_\mathbf{i})) \nabla c_k(t, \mathbf{x}_\mathbf{i}) \right) = \mathcal{O}\left(h^2\right),$$

in all grid cells $\Omega_\mathbf{i}$, where for each $j \in \{1, 2, \ldots, d\}$

1. *the local velocities $v_{\mathbf{i},j}$ and $v_{\mathbf{i}-\mathbf{e}_j,j}$ have the same sign (uniform flow regions), and*
2. *$U_\mathbf{i}, U_{\mathbf{i}+\mathbf{e}_j}$ are no local extrema in the jth coordinate direction if $v_{\mathbf{i},j}, v_{\mathbf{i}-\mathbf{e}_j,j} \leq 0$ and $U_\mathbf{i}, U_{\mathbf{i}-\mathbf{e}_j}$ are no local extrema in the jth coordinate direction if $v_{\mathbf{i},j}, v_{\mathbf{i}-\mathbf{e}_j,j} \geq 0$.*

In cells $\Omega_\mathbf{i}$ where these two conditions are not satisfied we have a local truncation error $\mathcal{O}(h)$.

Proof Taylor expansion. □

Lemma 3 *Let the limiter function Φ satisfy the condition (3.14b). Then the ODE system*

$$\frac{\mathrm{d}}{\mathrm{d}\,t} U_\mathbf{i}(t) = \mathcal{H}_T(\mathbf{U}(t); \mathbf{i}), \quad \mathbf{i} \in \mathcal{I},$$

obtained with the state interpolation approach is a positive ODE system.

Proof We restrict our attention to a single value of $j \in \{1, 2, \ldots, d\}$ in (3.9). If we prove positivity for this case then we also have positivity of the full system by Corollary 1.
We assume in the following that $\mathbf{U} \geq \mathbf{0}$. The application of the formulas (3.12) in cells close to the boundary requires values $U_\mathbf{i}$, where the multi-index $\mathbf{i} \notin \mathcal{I}$ because they correspond to grid cells outside the partition of Ω (and hence these values are not contained in $\mathbf{U}$). We assume in the following that these values are computed from $\mathbf{U}$ and given boundary data and that the resulting values are nonnegative. Then, without loss of generality, we need to consider cells $\mathbf{i}$ sufficiently far away from the domain boundary only.

Formula (3.11) generates four cases depending on the signs of $v_{\mathbf{i},j}$ and $v_{\mathbf{i}-\mathbf{e}_j,j}$. We prove the result only for *Case (I)* $v_{\mathbf{i},j}, v_{\mathbf{i}-\mathbf{e}_j,j} \geq 0$; the other three cases follow similarly.
Let $v_{\mathbf{i},j}, v_{\mathbf{i}-\mathbf{e}_j,j} \geq 0$ (Case (I)). This leads to

$$h\mathcal{H}_T(\mathbf{U};\mathbf{i}) = -\left(v_{\mathbf{i},j}\mathcal{S}_j^+(\mathbf{U};\mathbf{i}) - v_{\mathbf{i}-\mathbf{e}_j,j}\mathcal{S}_j^+(\mathbf{U};\mathbf{i}-\mathbf{e}_j)\right). \tag{3.15}$$

This results in four different cases again which we consider in turn now.
Case (Ia) $U_{\mathbf{i}} - U_{\mathbf{i}-\mathbf{e}_j} \neq 0$ and $U_{\mathbf{i}-\mathbf{e}_j} - U_{\mathbf{i}-2\mathbf{e}_j} \neq 0$
This leads to

$$h\mathcal{H}_T(\mathbf{U};\mathbf{i}) = -\left[\frac{v_{\mathbf{i},j}U_{\mathbf{i}} - v_{\mathbf{i}-\mathbf{e}_j,j}U_{\mathbf{i}-\mathbf{e}_j}}{U_{\mathbf{i}} - U_{\mathbf{i}-\mathbf{e}_j}} + \frac{v_{\mathbf{i},j}}{2}\Phi(r_{\mathbf{i},j}) - \frac{v_{\mathbf{i}-\mathbf{e}_j,j}}{2}\frac{\Phi(r_{\mathbf{i}-\mathbf{e}_j,j})}{r_{\mathbf{i}-\mathbf{e}_j,j}}\right](U_{\mathbf{i}} - U_{\mathbf{i}-\mathbf{e}_j}),$$

and for $U_{\mathbf{i}} = 0$ to

$$h\mathcal{H}_T(\mathbf{U};\mathbf{i}) = -\left[v_{\mathbf{i}-\mathbf{e}_j,j} + \frac{v_{\mathbf{i},j}}{2}\Phi(r_{\mathbf{i},j}) - \frac{v_{\mathbf{i}-\mathbf{e}_j,j}}{2}\frac{\Phi(r_{\mathbf{i}-\mathbf{e}_j,j})}{r_{\mathbf{i}-\mathbf{e}_j,j}}\right](-U_{\mathbf{i}-\mathbf{e}_j}).$$

Hence the condition of Theorem 1 is fulfilled if the expression in the square brackets is nonnegative and this is ensured by the limiter properties (3.14b).
Case (Ib) $U_{\mathbf{i}} - U_{\mathbf{i}-\mathbf{e}_j} = 0$ and $U_{\mathbf{i}-\mathbf{e}_j} - U_{\mathbf{i}-2\mathbf{e}_j} \neq 0$
This gives $r_{\mathbf{i}-\mathbf{e}_j,j} = 0$ and hence, for $U_{\mathbf{i}} = 0$, $h\mathcal{H}_T(\mathbf{U};\mathbf{i}) = v_{\mathbf{i}-\mathbf{e}_j,j}U_{\mathbf{i}-\mathbf{e}_j}$, and the condition of Theorem 1 is satisfied.
Case (Ic) $U_{\mathbf{i}} - U_{\mathbf{i}-\mathbf{e}_j} \neq 0$ and $U_{\mathbf{i}-\mathbf{e}_j} - U_{\mathbf{i}-2\mathbf{e}_j} = 0$
This, together with $U_{\mathbf{i}} = 0$, results in

$$h\mathcal{H}_T(\mathbf{U};\mathbf{i}) = -\left[v_{\mathbf{i}-\mathbf{e}_j,j} + \frac{v_{\mathbf{i},j}}{2}\Phi(r_{\mathbf{i},j})\right](-U_{\mathbf{i}-\mathbf{e}_j}),$$

and the expression in the square brackets is again nonnegative because of (3.14b).
Case (Id) $U_{\mathbf{i}} - U_{\mathbf{i}-\mathbf{e}_j} = 0$ and $U_{\mathbf{i}-\mathbf{e}_j} - U_{\mathbf{i}-2\mathbf{e}_j} = 0$
Here we obtain, for $U_{\mathbf{i}} = 0$, $h\mathcal{H}_T(\mathbf{U};\mathbf{i}) = v_{\mathbf{i}-\mathbf{e}_j,j}U_{\mathbf{i}-\mathbf{e}_j}$, and the condition of Theorem 1 is satisfied.
Altogether we have that $\mathcal{H}_T \in \mathcal{P}$ and the corresponding ODE system is positive. □
We now give a few limiter functions which we will use in our numerical tests. There are more functions available in the literature, see e.g. [57, 31, 35].

- Van Leer limiter $\Phi_{VL}(r)$:

$$\Phi_{VL}(r) := \frac{r + |r|}{1 + |r|}.$$

  This limiter function satisfies the conditions (3.14a), and (3.14b) with $\delta = 2$. $\Phi_{VL}(r)$ is a smooth function except in the origin $r = 0$.

- Koren limiter $\Phi_K(r)$:

$$\Phi_K(r) := \max\left\{0, \min\left\{2r, \delta, K_{1/3}(r)\right\}\right\},\ \delta = 2,$$

  where

$$K_\kappa(r) := \frac{1-\kappa}{2} + \frac{1+\kappa}{2}r\,.$$

This limiter function also satisfies the conditions (3.14a) and (3.14b) but is less smooth than the van Leer limiter. For the choice $\Phi(r) := K(r)$ we obtain the so-called κ-methods which include the second-order upwind discretization ($\kappa = -1$), second-order central discretization ($\kappa = 1$), and the third-order upwind biased discretization ($\kappa = \frac{1}{3}$). However, these schemes do not satisfy the conditions (3.14b) and are prone to introduce wiggles and negative solution values in the numerical approximations. Therefore we only consider the limited version Φ_K with $\kappa = \frac{1}{3}$.

- First-order upwind $\Phi_1(r)$:

$$\Phi_1(r) := 0$$

This leads to the standard first-order upwind discretization and a positive scheme ((3.14b) is satisfied). However, the approximation order is only one ((3.14a) is not satisfied).

3.3.2 Diffusion

In this section we present an approximation $\mathcal{H}_D(\mathbf{U};\mathbf{i})$ in conservation form to the diffusion part $H_D(u(t,\cdot);\mathbf{i})$ in Eq. (3.5). We start with the conservative formula

$$\mathcal{H}_D(\mathbf{U};\mathbf{i}) := \frac{1}{h}\sum_{j=1}^{d}\left(\mathcal{D}_j(\mathbf{U};\mathbf{i}) - \mathcal{D}_j(\mathbf{U};\mathbf{i}-\mathbf{e}_j)\right), \tag{3.16}$$

where the function $\mathcal{D}_j(\mathbf{U};\mathbf{i})$ approximates the average of the negative diffusive flux from grid cell $\Omega_{\mathbf{i}}$ to $\Omega_{\mathbf{i}+\mathbf{e}_j}$ through their common cell face Γ, i.e.

$$\mathcal{D}_j(\mathbf{U};\mathbf{i}) \approx \frac{1}{h^{d-1}}\int_\Gamma D\nabla u\cdot\mathfrak{n}d\Gamma = \frac{1}{h^{d-1}}\int_\Gamma D\partial_{x_j}u d\Gamma, \quad \Gamma := \bar{\Omega}_{\mathbf{i}}\cap\bar{\Omega}_{\mathbf{i}+\mathbf{e}_j}.$$

We define

$$\mathcal{D}_j(\mathbf{U};\mathbf{i}) := \frac{D}{h}\left(U_{\mathbf{i}+\mathbf{e}_j} - U_{\mathbf{i}}\right). \tag{3.17}$$

Substituting this into Eq. (3.16) leads, in a pointwise interpretation, to the standard second-order central difference approximation of the diffusion operator. Further, we also obtain a second-order approximation of the evolving cell averages.

Lemma 4 *Let $u(t,\mathbf{x})$ be a smooth function. Then the local truncation error of the approximation (3.16, 3.17) to the exact diffusion term $H_D(u(t,\cdot);\mathbf{i})$ is second-order,*

$$lte(t,\mathbf{i}) := \mathcal{H}_D(\bar{u}(t);\mathbf{i}) - H_D(u(t,\cdot);\mathbf{i}) = \mathcal{O}\left(h^2\right) \quad \textit{for all } \mathbf{i}\in\{2,3,\ldots,M-1\}^d.$$

Proof We only show that the local truncation error in one particular coordinate direction j is second-order. The statement of the lemma follows then immediately.
We denote $X^k := [-h/2,h/2]^k$ for $k\in\mathbb{N}$, and, for $\mathbf{x}\in X^{k-1}$, denotes $\widehat{\mathbf{x}}\in X^k$ the vector $\mathbf{x}$ with an additional zero inserted in the jth position. Further, for each $\mathbf{x}\in X^k$, there exist unique

$\mathbf{x}^*(\mathbf{x}) \in X^{k-1}$ and $r(\mathbf{x}) \in X$ such that $\mathbf{x} = \widehat{\mathbf{x}}^*(\mathbf{x}) + r(\mathbf{x})e_j$. Then we obtain:

$$
\begin{aligned}
& \frac{1}{h}\left(\mathcal{D}_j(\bar{u}(t);\mathbf{i}) - \mathcal{D}_j(\bar{u}(t);\mathbf{i}-\mathbf{e}_j)\right) - \frac{1}{h^d}\oint_{\partial\Omega_i} D(\partial_{x_j}u)\mathbf{e}_j \, \mathfrak{n} d\partial\Omega_{\mathbf{i}} \\
= & \frac{D}{h^{d+2}}\int_{X^d} u(t,\mathbf{x}_{\mathbf{i}}+\mathbf{x}+h\mathbf{e}_j) - 2u(t,\mathbf{x}_{\mathbf{i}}+\mathbf{x}) + u(t,\mathbf{x}_{\mathbf{i}}+\mathbf{x}-h\mathbf{e}_j)d\mathbf{x} \\
& - \frac{D}{h^d}\int_{X^{d-1}} \left(\partial_{x_j}u\left(t,\mathbf{x}_{\mathbf{i}}+\frac{h}{2}\mathbf{e}_j+\widehat{\mathbf{x}}\right) - \partial_{x_j}u\left(t,\mathbf{x}_{\mathbf{i}}-\frac{h}{2}\mathbf{e}_j+\widehat{\mathbf{x}}\right)\right)d\mathbf{x} \\
= & \frac{D}{h^d}\int_{X^d}\partial^2_{x_j}u(t,\mathbf{x}_{\mathbf{i}}+\mathbf{x})d\mathbf{x} - \frac{D}{h^d}h\int_{X^{d-1}}\partial^2_{x_j}u(t,\mathbf{x}_{\mathbf{i}}+\widehat{\mathbf{x}})d\mathbf{x} + \mathcal{O}\left(h^2\right) \\
= & \frac{D}{h^d}\left(\int_{X^d}\partial^2_{x_j}u(t,\mathbf{x}_{\mathbf{i}}+\widehat{\mathbf{x}}^*) + \partial^3_{x_j}u(t,\mathbf{x}_{\mathbf{i}}+\widehat{\mathbf{x}}^*)rd\mathbf{x} - h\int_{X^{d-1}}\partial^2_{x_j}u(t,\mathbf{x}_{\mathbf{i}}+\widehat{\mathbf{x}})d\mathbf{x}\right) + \mathcal{O}\left(h^2\right) \\
= & \frac{D}{h^d}\left(h\int_{X^{d-1}}\partial^2_{x_j}u(t,\mathbf{x}_{\mathbf{i}}+\widehat{\mathbf{x}})d\mathbf{x} + 0 - h\int_{X^{d-1}}\partial^2_{x_j}u(t,\mathbf{x}_{\mathbf{i}}+\widehat{\mathbf{x}})d\mathbf{x}\right) + \mathcal{O}\left(h^2\right) \\
= & \mathcal{O}\left(h^2\right).
\end{aligned}
$$

The integral over $r\partial^3_{x_j}u$ vanishes because of symmetry reasons. This proves second-order accuracy of the discretization. □

Lemma 5 *The ODE system*

$$\frac{\mathrm{d}}{\mathrm{d}\,t}U_{\mathbf{i}}(t) = \mathcal{H}_D(\mathbf{U}(t);\mathbf{i}), \quad \mathbf{i} \in \mathcal{I},$$

is a positive ODE system.

Proof The ODE system is linear and the system matrix has nonnegative off-diagonal entries. Hence positivity follows with Corollary 2. □

3.3.3 Reaction

We approximate the reaction part H_R of Eq. (3.5) with $\mathcal{H}_R(\mathbf{U};\mathbf{i})$ defined by

$$\mathcal{H}_R(\mathbf{U};\mathbf{i}) := p_0(t,\mathbf{x}_{\mathbf{i}},U_{\mathbf{i}},\mathbf{C}_{\cdot,\mathbf{i}}). \tag{3.18}$$

The local truncation error of this approximation computes as

$$
\begin{aligned}
lte(t,\mathbf{i}) =& \mathcal{H}_R(\bar{u}(t);\mathbf{i}) - H_R(u(t,\cdot);\mathbf{i}) \\
=& p_0(t,\mathbf{x}_{\mathbf{i}},\bar{u}_{\mathbf{i}}(t),\bar{\mathbf{c}}_{\cdot,\mathbf{i}}(t)) - \frac{1}{|\Omega_{\mathbf{i}}|}\int_{\Omega_{\mathbf{i}}} p_0(t,\mathbf{x},u(t,\mathbf{x}),\mathbf{c}(t,\mathbf{x}))d\mathbf{x} \\
=& \frac{1}{|\Omega_{\mathbf{i}}|}\int_{\Omega_{\mathbf{i}}}[p_0(t,\mathbf{x}_{\mathbf{i}},\bar{u}_{\mathbf{i}}(t),\bar{\mathbf{c}}_{\cdot,\mathbf{i}}(t)) - p_0(t,\mathbf{x},u(t,\mathbf{x}),\mathbf{c}(t,\mathbf{x}))]d\mathbf{x} \\
=& \mathcal{O}\left(h\right),
\end{aligned}
$$

for continuously differentiable $u(t,\cdot)$, $\mathbf{c}(t,\cdot)$ and $p_0(t,\cdot,\cdot,\cdot)$ because for $\mathbf{x} \in \Omega_\mathbf{i}$ we have

$$p_0(t,\mathbf{x},u(t,\mathbf{x}),\mathbf{c}(t,\mathbf{x})) = p_0(t,\mathbf{x}_\mathbf{i},u(t,\mathbf{x}_\mathbf{i}),\mathbf{c}(t,\mathbf{x}_\mathbf{i})) + \mathcal{O}(h) = p_0(t,\mathbf{x}_\mathbf{i},\bar{u}_\mathbf{i}(t),\bar{\mathbf{c}}_\mathbf{i}(t)) + \mathcal{O}(h)\,.$$

We see that this only leads to a first-order approximation in a finite volume interpretation. We are satisfied with this approximation for two reasons.

- If we regard the average quantities $U_\mathbf{i}(t)$ and $\mathbf{C}_{\cdot,\mathbf{i}}$ as point approximations of $u(t,\mathbf{x}_\mathbf{i})$ and $\mathbf{c}(t,\mathbf{x}_\mathbf{i})$ then (3.18) corresponds just to the correct evaluation of the source term at $(t,\mathbf{x}_\mathbf{i})$ and is exact.

- The order of the approximation (3.18) could be improved by making it dependent on average values in neighbouring cells. If we assume that a wave with steep front travels across the domain then using other values than $U_\mathbf{i}$ and $\mathbf{C}_{\cdot,\mathbf{i}}$ in (3.18) could trigger a reaction in cell $\Omega_\mathbf{i}$ although the wave has not reached the cell $\Omega_\mathbf{i}$ yet and hence lead to wrong solutions or wave speeds.

The positivity of the ODE system $\frac{\mathrm{d}}{\mathrm{d}\,t}U_\mathbf{i}(t) = \mathcal{H}_R(\mathbf{U}(t);\mathbf{i})$ depends strongly on the properties of the function p_0 and can be characterized by the following lemma.

Lemma 6 *The ODE system*

$$\frac{\mathrm{d}}{\mathrm{d}\,t}U_\mathbf{i}(t) = \mathcal{H}_R(\mathbf{U}(t);\mathbf{i})\,, \quad \mathbf{i} \in \mathcal{I}\,,$$

is a positive ODE system if and only if $p_0(t,\mathbf{x},0,\mathbf{c}) \geq 0$ *for all* $(t,\mathbf{x}) \in I_T \times \Omega$ *and all possible values of* $\mathbf{c}$*.*

Proof The statement follows immediately with Theorem 1. □

3.3.4 Spatial discretization of problem class (2.5) in boundary cells

We consider here computational cells $\Omega_\mathbf{i}$ adjacent to the boundary of Ω. Specifically, we fix a spatial direction $j \in \{1,2,\ldots,d\}$ and consider a cell $\mathbf{i} \in \mathcal{I}$ with $i_j = 1$ (*left* boundary cell) or $i_j = M$ (*right* boundary cell). The boundary face of $\Omega_\mathbf{i}$ in the jth coordinate direction which is part of $\partial\Omega$ is denoted by Γ. We define $\nu = \pm 1$ such that $\nu\mathbf{e}_j$ is the outer normal vector on Γ with respect to $\Omega_\mathbf{i}$, and $\mathbf{i}^B$ such that $\Omega_{\mathbf{i}^B}$ is the grid cell to the *left* of Γ, i.e. $\nu = -1$, $\mathbf{i}^B = \mathbf{i} - \mathbf{e}_j$ for $i_j = 1$ and $\nu = 1$, $\mathbf{i}^B = \mathbf{i}$ for $i_j = M$. Finally, $\mathbf{x}^*_\mathbf{i} := \mathbf{x}_\mathbf{i} + \nu\frac{h}{2}\mathbf{e}_j$ is the centre of Γ. We discuss the spatial discretization of problem class (2.5) in such cells $\Omega_\mathbf{i}$. We do not consider the prototype equation (3.3) here because we take some advantage of the special structure of (2.5).
We start with the equation for the chemical c_k, $k \in \{1,2,\ldots,l\}$. If the diffusion coefficient $D_k = 0$ then we have no boundary conditions (BCs) for c_k and the spatial discretization of the corresponding equation is well defined. Let $D_k > 0$ in the following. We assume that we have either Dirichlet or flux BCs prescribed on Γ. In order to evaluate the diffusion discretization (3.16) in the cell $\Omega_\mathbf{i}$ we must provide $\mathcal{D}_j(\mathbf{C}_k;\mathbf{i}^B) \approx \frac{1}{h^{d-1}}\int_\Gamma D_k\partial_{x_j}c_k(t,\mathbf{x})d\Gamma$.

(i) Dirichlet BC for c_k: $c_k(t,\mathbf{x}) = \alpha_D^{(k)}(t,\mathbf{x}) \geq 0$ for $\mathbf{x} \in \Gamma$.
Here we have the state of c_k on Γ prescribed. Let $\bar{\alpha}_D^{(k)} := \frac{1}{h^{d-1}} \int_\Gamma \alpha_D^{(k)}(t,\mathbf{x})\, d\Gamma$. We only want to use $\bar{\alpha}_D^{(k)}$ and the values $C_{k,\mathbf{i}}$ and $C_{k,\mathbf{i}-\nu\mathbf{e}_j}$ to approximate the average of the negative diffusive flux on Γ because then the stencil of the diffusion discretization in cell $\mathbf{i}$ is the same as for any interior cell of the partition (this will become important in the time integration process, see the next chapter). The best what we can achieve under this restriction is to set

$$\mathcal{D}_j(\mathbf{C}_k;\mathbf{i}^B) := -\nu\frac{D_k}{3h}\left(-8\bar{\alpha}_D^{(k)} + 9C_{k,\mathbf{i}} - C_{k,\mathbf{i}-\nu\mathbf{e}_j}\right).$$

This results in a first-order accurate discretization of the diffusion part in $\Omega_\mathbf{i}$. For the spatially independent BCs of the models from Sec. 2.3 we simply have $\bar{\alpha}_D^{(k)} = \alpha_D^{(k)}(t,\mathbf{x}_\mathbf{i}^*)$. We see that this definition results in a positive semi-discretization of the diffusion part because $\alpha_D^{(k)} \geq 0$.
(ii) Flux BC for c_k: $-D_k\nabla c_k(t,\mathbf{x})\cdot\mathfrak{n}(\mathbf{x}) = \alpha_F^{(k)}(t,\mathbf{x}) \leq 0$ for $\mathbf{x}\in\Gamma$.
In this case we have the diffusive flux of c_k through Γ prescribed (inflow because $\alpha_F^{(k)} \leq 0$). Hence the average of the negative diffusive flux over Γ is

$$\mathcal{D}_j(\mathbf{C}_k;\mathbf{i}^B) := -\frac{\nu}{h^{d-1}}\int_\Gamma \alpha_F^{(k)}(t,\mathbf{x})\, d\Gamma =: -\nu\bar{\alpha}_F^{(k)}.$$

Again we can approximate by the point value in the centre and set $\mathcal{D}_j(\mathbf{C}_k;\mathbf{i}^B) := -\nu\alpha_F^{(k)}(t,\mathbf{x}_\mathbf{i}^*)$. Both definitions are equivalent if $\alpha_F^{(k)}$ is independent of $\mathbf{x}$ and this is the case for all models described in Sec. 2.3. The result is a positive semi-discretization because $\alpha_F^{(k)} \leq 0$.
Let us turn the attention to the population density equation for n now. If there is diffusion, $\varepsilon > 0$, then we need an approximation $\mathcal{D}_j(\mathbf{N};\mathbf{i}^B) \approx \frac{1}{h^{d-1}}\int_\Gamma \varepsilon\partial_{x_j} n(t,\mathbf{x})d\Gamma$ so that we can apply the diffusion discretization (3.16) in $\Omega_\mathbf{i}$. Further in order to apply the taxis discretization (3.9) in grid cell $\mathbf{i}$, we require an approximate of the average of the taxis flux through Γ, i.e. $\mathcal{T}_j(\mathbf{N};\mathbf{i}^B) \approx \frac{1}{h^{d-1}}\int_\Gamma n\left(\sum_{k=1}^l p_k(\mathbf{c})\partial_{x_j}c_k\right)d\Gamma$, and also possibly (depending on the upwind direction) an approximate value $N_{\mathbf{i}+\nu\mathbf{e}_j}$ of the average of the state of n in the (outside of the domain Ω) grid cell $\Omega_{\mathbf{i}+\nu\mathbf{e}_j}$ for the computation of the state interpolation $\mathcal{S}_j^\pm(\mathbf{N};\mathbf{i}^B - \nu\mathbf{e}_j)$, see Eq. (3.12). We again assume that we have exactly one type of BC prescribed on Γ.
(iii) Dirichlet BC for n: $n(t,\mathbf{x}) = \alpha_D^{(0)}(t,\mathbf{x}) \geq 0$ for $\mathbf{x}\in\Gamma$.
Let again $\bar{\alpha}_D^{(0)} := \frac{1}{h^{d-1}}\int_\Gamma \alpha_D^{(0)}(t,\mathbf{x})\, d\Gamma$. We define $\mathcal{D}_j(\mathbf{N};\mathbf{i}^B)$ by the same approach as in (i). Hence we set (leading to a positive semi-discretization of the diffusion part in $\Omega_\mathbf{i}$)

$$\mathcal{D}_j(\mathbf{N};\mathbf{i}^B) := -\nu\frac{\varepsilon}{3h}\left(-8\bar{\alpha}_D^{(0)} + 9N_\mathbf{i} - N_{\mathbf{i}-\nu\mathbf{e}_j}\right).$$

For the definition of $\mathcal{T}_j(\mathbf{N};\mathbf{i}^B)$ we approximate

$$\frac{1}{h^{d-1}}\int_\Gamma n\left(\sum_{k=1}^l p_k(\mathbf{c})\partial_{x_j}c_k\right)d\Gamma \approx v_{av}\frac{1}{h^{d-1}}\int_\Gamma n d\Gamma = v_{av}\bar{\alpha}_D^{(0)} =: \mathcal{T}_j(\mathbf{N};\mathbf{i}^B),$$

where v_{av} is a suitable approximation of the velocity $\sum_{k=1}^l p_k(\mathbf{c})\partial_{x_j}c_k$ on Γ. The definition of v_{av} depends on the BCs of the c_k on Γ again and we proceed as follows.

- If c_k satisfies a Dirichlet BC on Γ, i.e. $c_k(t,\mathbf{x}) = \alpha_D^{(k)}(t,\mathbf{x})$, then we define $\widetilde{c}_k := \bar{\alpha}_D^{(k)}$ and approximate $\partial_{x_j} c_k$ (as in (i) but with $D_k = 1$) by $\widetilde{c}_{k,x_j} := -\nu\frac{1}{3h}\left(-8\widetilde{c}_k + 9C_{k,\mathbf{i}} - C_{k,\mathbf{i}-\nu\mathbf{e}_j}\right)$.
- If c_k satisfies a flux BC on Γ, i.e. $-D_k\nabla c_k(t,\mathbf{x})\cdot\mathfrak{n}(\mathbf{x}) = \alpha_F^{(k)}(t,\mathbf{x})$, then we define the approximation of $\partial_{x_j} c_k$ as $\widetilde{c}_{k,x_j} := -\nu\frac{1}{D_k}\bar{\alpha}_F^{(k)}$ and set $\widetilde{c}_k := \max\{0, C_{k,\mathbf{i}} + \nu\frac{h}{2}\widetilde{c}_{k,x_j}\}$.
- If no BCs for c_k on Γ are prescribed (and if we cannot deduce values of c_k or its derivative on the boundary otherwise) then we simply use linear (positive) extrapolation and set $\widetilde{c}_k := \max\{0, C_{k,\mathbf{i}} - \frac{1}{2}(C_{k,\mathbf{i}-\nu\mathbf{e}_j} - C_{k,\mathbf{i}})\}$ and $\widetilde{c}_{k,x_j} := -\nu\frac{2}{h}(C_{k,\mathbf{i}} - \widetilde{c}_k)$.

Let $\widetilde{\mathbf{c}} := [\widetilde{c}_1, \widetilde{c}_2, \ldots, \widetilde{c}_l]^\mathsf{T}$. Then we define the average velocity v_{av} by

$$v_{av} := \sum_{k=1}^{l} p_k(\widetilde{\mathbf{c}})\widetilde{c}_{k,x_j},$$

which completes the definition of $\mathcal{T}_j(\mathbf{N};\mathbf{i}^B)$. Finally, we assume that the state of n in $\Omega_{\mathbf{i}+\nu\mathbf{e}_j}$ is the same as on the Dirichlet boundary part Γ and therefore approximate $N_{\mathbf{i}+\nu\mathbf{e}_j}$ by

$$N_{\mathbf{i}+\nu\mathbf{e}_j} := \bar{\alpha}_D^{(0)}.$$

It can be shown (by a tedious calculation using the definition of the state interpolants and properties of the limiter) that if $\nu v_{av}\bar{\alpha}_D^{(0)} \le 0$ then, with the given definitions, the semi-discretization of the taxis part in $\Omega_\mathbf{i}$ is positive. The condition roughly states that Γ is a no outflow boundary for n.

(iv) Flux BC for n: $\left(-\varepsilon\nabla n + n\left(\sum_{k=1}^{l} p_k(\mathbf{c})\nabla c_k\right)\right)\cdot\mathfrak{n}(\mathbf{x}) = \alpha_F^{(0)}(t,\mathbf{x}) \le 0$ for $\mathbf{x}\in\Gamma$.

This implies on Γ the relation $-\varepsilon\partial_{x_j} n + n\left(\sum_{k=1}^{l} p_k(\mathbf{c})\partial_{x_j} c_k\right) = \nu\alpha_F^{(0)}(t,\mathbf{x})$.

For $\varepsilon = 0$ we can simply set

$$\mathcal{T}_j(\mathbf{N};\mathbf{i}^B) := \nu\frac{1}{h^{d-1}}\int_\Gamma \alpha_F^{(0)}(t,\mathbf{x})d\Gamma = \nu\bar{\alpha}_F^{(0)} \quad \text{and} \quad N_{\mathbf{i}+\nu\mathbf{e}_j} := N_\mathbf{i}.$$

For $\varepsilon > 0$ we compute an average velocity v_{av} as in (iii) and an average value $\bar{\alpha}_D^{(0)}$ of n on Γ by first-order extrapolation (so as not to enlarge the stencil of the diffusion discretization) with enforced positivity,

$$\bar{\alpha}_D^{(0)} := \max\left\{0, N_\mathbf{i} - \frac{1}{2}\left(N_{\mathbf{i}-\nu\mathbf{e}_j} - N_\mathbf{i}\right)\right\}.$$

Then we approximate the tactic and the negative diffusive flux by

$$\mathcal{T}_j(\mathbf{N};\mathbf{i}^B) := v_{av}\bar{\alpha}_D^{(0)} \quad \text{and} \quad \mathcal{D}_j(\mathbf{N};\mathbf{i}^B) := \mathcal{T}_j(\mathbf{N};\mathbf{i}^B) - \nu\bar{\alpha}_F^{(0)}.$$

Finally, we set $N_{\mathbf{i}+\nu\mathbf{e}_j} := \max\{0, N_\mathbf{i} + \frac{\nu h}{\varepsilon}\mathcal{D}_j(\mathbf{N};\mathbf{i}^B)\}$.

Altogether we obtain a positive semi-discretization of the diffusion and the taxis part of the population density equation also in this case because $\alpha_F^{(0)} \le 0$ on Γ.

(v) No BC for n:

This can only be the case if $\varepsilon = 0$ and we need approximations $\mathcal{T}_j(\mathbf{N};\mathbf{i}^B)$ and $N_{\mathbf{i}+\nu\mathbf{e}_j}$. We define $\bar{\alpha}_D^{(0)} = N_\mathbf{i}$ and then apply the formulas from case **(iii)**.

3.4 Evaluation of the spatial discretization of the taxis part

In this section we present numerical confirmation that the spatial discretization described is appropriate for the taxis part of TDR systems. We investigate whether the expected order of convergence is attained numerically, and we also discuss the different results obtained with different limiter functions Φ. Model 1 is a suitable test model for this purpose because we have an analytic solution of this problem. Further, the solution of this problem is radial symmetric and we want that the numerical approximation shares (approximately) the same qualitative property. Other qualitative tests are the conservation of mass and the nonnegativity property of the solution of Model 1.
We discretize the taxis term in Model 1 with the state interpolation approach and use three limiter functions: van Leer Φ_{VL}, Koren Φ_K, and first-order Φ_1. The result of this spatial discretization on the partition $\{\Omega_{\mathbf{i}}\}_{\mathbf{i}\in\mathcal{I}}$ is the MOL-ODE. (Note that the computations of the model are executed on the unit square – no advantage is taken by assuming that the solution is radial symmetric.)
We are concerned with comparing the exact solution $u(t,\mathbf{x})$ of Model 1 with the exact solution $\mathbf{U}(t)$ of the MOL-ODE in some norm. We do not know the exact solution of the MOL-ODE and therefore it has to be obtained numerically. For this purpose we employ the standard ODE solver DOPRI5 [22] with sufficiently high accuracy so that the errors of the time integration become negligible compared to the spatial errors introduced by the discretization in space. We regard the result as exact solution of the MOL-ODE and denote it with $\mathbf{U}(t)$. On the other hand, we know the exact (PDE) solution of Model 1 and define a reference solution $\mathbf{U}_{ref}(t)$ by $U_{ref,\mathbf{i}} := n(t,\mathbf{x}_{\mathbf{i}})$ for all $i \in \mathcal{I}$. We measure the difference $\mathbf{E} := \mathbf{U}(t) - \mathbf{U}_{ref}(t)$ between both vectors in two different norms: the maximum norm $\|\cdot\|_\infty$ and the discrete L^1-norm $\|\cdot\|_1$,

$$\|\mathbf{E}\|_\infty := \max_{\mathbf{i}\in\mathcal{I}} |E_{\mathbf{i}}| \quad \text{and} \quad \|\mathbf{E}\|_1 := \sum_{\mathbf{i}\in\mathcal{I}} |\Omega_{\mathbf{i}}||E_{\mathbf{i}}|. \tag{3.19}$$

We start with assessing the numerical order of convergence of our discretization in space and therefore choose the smooth initial function (2.6) with $\kappa = 0.09$. Then the solution of Model 1 is also smooth. We note, however, that the gradients in the solution become steeper with increasing time. For this reason, we consider three final times, $T_1 = 0.007, T_2 = 0.014$, and $T_3 = 0.021$. Tyson et al. [58] consider the same final times for this model (but with parameter $\kappa = 0$ in the initial condition as we will also do later in this section). We compute the solution on the sequence of partitions of Ω with grid widths $h = h_k := \frac{1}{50k},\ k = 2,3,\ldots,12$. In Fig. 3.2 we plot the logarithm of the measured $\|\cdot\|_1$-error (top row plots) and the logarithm of the measured $\|\cdot\|_\infty$-error (bottom row plots) obtained with the van Leer Φ_{VL}, Koren Φ_K, and first-order Φ_1 limiter functions (see end of Sec. 3.3.1) vs. the cell width h for three final times T (left to right). The corresponding (numerical) orders of convergence p and the error constants C are computed by a least squares procedure such that $err_k \approx Ch_k^p$, where the error err_k is attained on the grid with cell width h_k. They are listed in Tab. 3.1.
We see that the discretizations converge to the analytic solution but immediately recognize that the first-order scheme cannot compete with the limited second-order discretizations. The error attained with the first-order scheme on the finest grid is of the size of the error of the second-order discretizations on the coarsest grid. Hence, the application of the first-order discretization for the taxis term in our more complex biomathematical models would require extremely fine meshes to achieve sufficient spatial resolution but this is not feasible in view of the implied computational

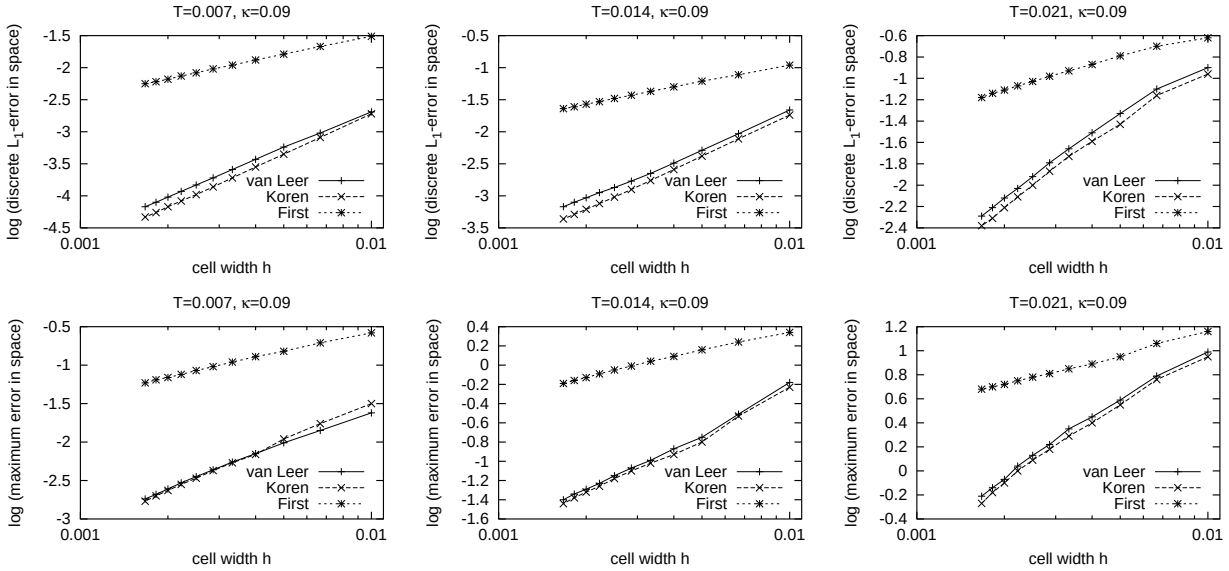

Figure 3.2: Plots of the logarithm of the measured $\|\cdot\|_1$-error (top row plots) and of the logarithm of the measured $\|\cdot\|_\infty$-error (bottom row plots) against the cell width h for different limiter functions and final times T (left to right) for Model 1 and smooth initial data ($\kappa = 0.09$).

	$T = 0.007$				$T = 0.014$				$T = 0.021$			
	$\|\cdot\|_1$-error		$\|\cdot\|_\infty$-error		$\|\cdot\|_1$-error		$\|\cdot\|_\infty$-error		$\|\cdot\|_1$-error		$\|\cdot\|_\infty$-error	
	C	p	C	p	C	p	C	p	C	p	C	p
Φ_{VL}	14.07	1.91	19.94	1.45	144.52	1.93	634.12	1.52	781.16	1.85	15168	1.57
Φ_K	26.55	2.07	62.13	1.64	276.04	2.09	541.79	1.51	799.18	1.89	14496	1.57
Φ_1	2.63	0.96	13.07	0.84	6.43	0.88	54.52	0.69	7.90	0.74	246.1	0.62

Table 3.1: Orders of convergence p and error constants C corresponding to the plots of Fig. 3.2.

effort. Therefore we will not consider the first-order spatial discretization of the flux terms for these models. Comparing the errors of the higher order discretizations we see that the discretizations using the Koren limiter are slightly more accurate than that using the van Leer limiter. The numbers in Tab. 3.1 show that these two discretizations almost attain the theoretical order two in the $\|\cdot\|_1$-norm and an approximate order of about 1.5 in the $\|\cdot\|_\infty$-norm. The latter is not surprising because the solution quickly develops a sharp (although smooth) peak and this peak is hard to approximate in the maximum norm. We note that the error constants C grow considerably large for increasing final time T. One reason is that the constant C depends on this final time, see Theorem 2, and another is the increased lack of spatial smoothness in the solution for increasing simulation time (see the scaling in the following solution plots to get an impression of the sharpness of the peak).
All three discretizations of Model 1 have nonnegative solutions at final time (at least for the high temporal accuracy requested when computing these solutions with DOPRI5—achieving the same for lower temporal accuracy requirements will be one of the topics of the next chapter). Further, the mass of the solution is conserved up to machine precision ($\approx 10^{-16}$) in all experiments.
Finally, we look at the symmetry of the solution of the MOL-ODE. We therefore plot, for a fixed value of t, all solution points $(\mathbf{x_i}, U_\mathbf{i}(t)), \mathbf{i} \in \mathcal{I}$, as points $(r(\mathbf{x_i}), U_\mathbf{i}(t))$ in a diagram. We plot the analytic PDE solution (2.8) of Model 1 at time t in the same manner in this diagram, and since

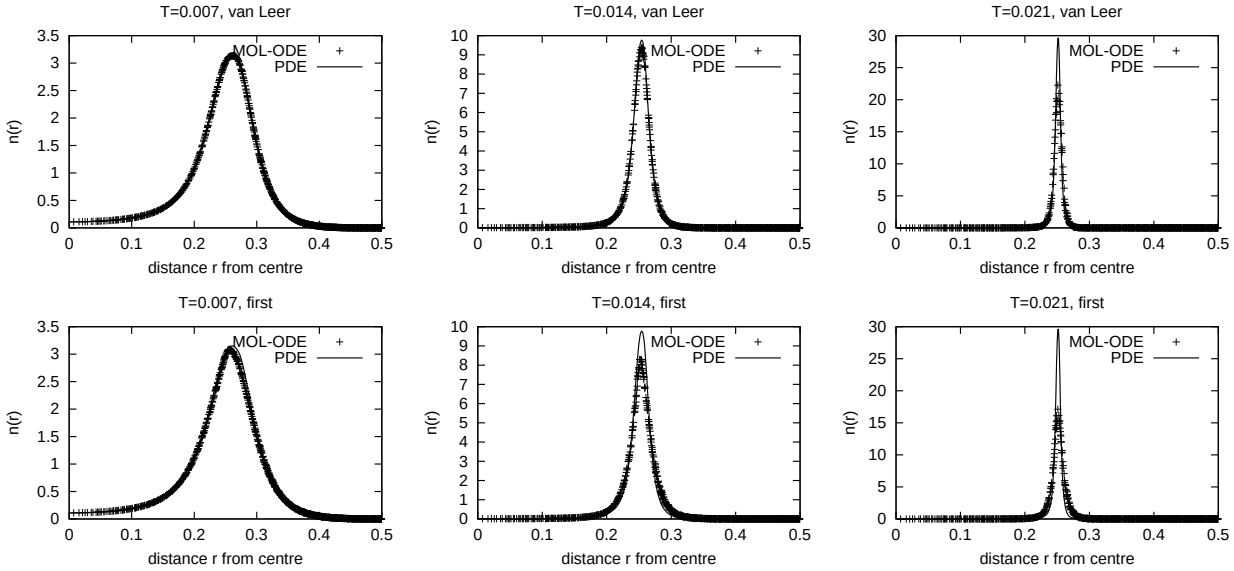

Figure 3.3: Values of the analytical solution of Model 1 with smooth initial data ($\kappa = 0.09$) and of the corresponding MOL-ODE (with van Leer limiter (top) and first-order limiter (bottom)) at three different final times T (left to right) plotted against the distance r between grid point and centre of the unit square. The spatial resolution is $h = 1/100$. (The results with Koren limiter (not given here) are almost indiscernible from the results with van Leer limiter.)

the PDE solution is radial symmetric, this corresponds to a single solution line in the diagram. In Fig. 3.3 we present some of these diagrams (for details see the caption there). We see that the solution points of the MOL-ODE are close to or even on the solution line of the PDE. Further, for a fixed value of r, there is no scattering of solution points of the MOL-ODE around the solution value of the PDE for r. This indicates that also the solution of the MOL-ODE is radial symmetric. It can also be seen that the first-order approximation results in a smeared peak whereas the higher order discretizations return a better resolved peak.

We now turn our attention to the discontinuous initial condition ($\kappa = 0$) in Model 1. When looking at discontinuous solutions then it makes no sense to measure the errors in the $\|\cdot\|_\infty$-norm (this error might be very large although the numerical approximation is very close to the true solution) and we only give plots of the spatial error in the $\|\cdot\|_1$-norm against the cell width, see Fig. 3.4, and the corresponding orders and error constants in Tab. 3.2.

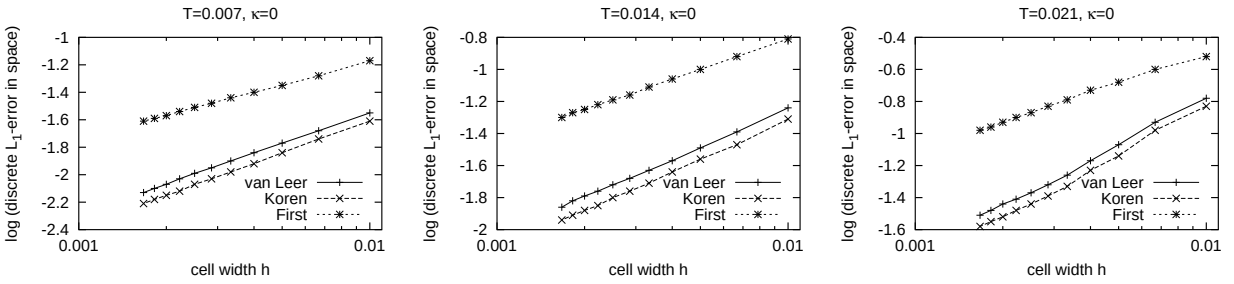

Figure 3.4: The same as in Fig. 3.2 but with nonsmooth initial data, $\kappa = 0$, and for the $\|\cdot\|_1$-norm only.

The same comments as given for smooth initial data apply regarding the first-order scheme. For

	$T = 0.007$		$T = 0.014$		$T = 0.021$	
	C	p	C	p	C	p
Φ_{VL}	0.87	0.74	2.05	0.78	13.52	0.96
Φ_K	0.88	0.78	1.94	0.80	13.86	0.99
Φ_1	0.88	0.56	2.79	0.63	5.20	0.61

Table 3.2: Orders p and constants C ($\|\cdot\|_1$-norm) corresponding to the plots of Fig. 3.4.

the higher order methods we observe that the differences are almost negligible. The order of convergence of the discretizations is clearly less than the theoretical order but this is expected and due to missing spatial smoothness of the solution, see [41, p. 121]. Also for nonsmooth initial data we have nonnegativity of the solution and conservation of initial mass up to machine precision. The plots in Fig. 3.5 show that the MOL-ODE solution is symmetric and they compare well with the results obtained for the same model and initial condition in the paper by Tyson et al. [58]. We note that the computation times for the approximate nonsmooth solutions with limiter Φ_K are considerably longer than with limiter Φ_{VL}.

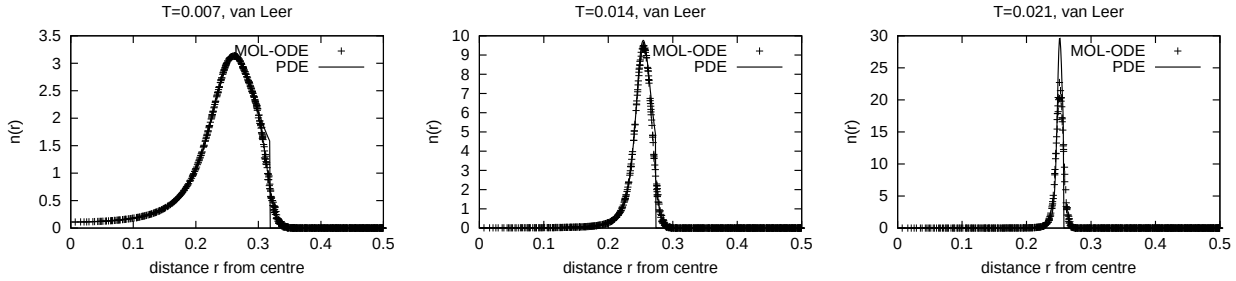

Figure 3.5: The same as in Fig. 3.3 but with nonsmooth initial data, $\kappa = 0$, and van Leer limiter only.

To summarize, we discourage the application of first-order approximations of taxis terms because an excessive amount of grid points is necessary to obtain a spatial accuracy which is comparable to the accuracy obtained by using higher order discretizations on very coarse meshes already. Further, the Koren limiter gives generally more accurate approximations than the van Leer limiter. However, the differences are not very big and both limiters can be recommended for application. In the numerical experiments in Chap. 5 we use the van Leer limiter function only.

Chapter 4

Time Stepping Methods

The result of the spatial discretization (described in Chap. 3) of the TDR system (2.5) is an IVP for a huge system of stiff, nonlinear ODEs which remains to be numerically integrated in time. We denote this system (repeating Eq. (3.1)) by

$$y'(t) = F(t, y(t)) \,, \quad t \geq t_0 \in \mathbb{R} \,, \quad y(t_0) = y_0 \in \mathbb{R}^m \,, \tag{4.1}$$

and henceforth assume that the right-hand side function F has the property (repeating (3.2))

$$F \text{ is continuous and (4.1) has a unique uncontinuable solution for all } t_0 \in \mathbb{R} \text{ and all } y_0 \in \mathbb{R}^m. \tag{4.2}$$

The aim of this chapter is to develop and discuss suitable numerical schemes to compute approximate solutions of (4.1) for times t up to a (moderately sized) final time T.
The components of the solution vector $y(t)$ of (4.1) are the semi-discrete approximations to the averages of population density n and chemical concentrations c_i in the elements of the chosen partition of the spatial domain (see Sec. 3.1). These semi-discrete approximations can be arranged in different orders in the vector $y(t)$ and this choice has a significant influence on the efficiency of the numerical schemes discussed in this chapter. For the problem class (2.5) and the partition described in Sec. 3.1, we obtain that $y(t)$ has a dimension $m = (l+1) \cdot M^d$. This number can be very large for $d > 1$ and standard integration methods for the solution of (4.1) are not always suitable. Therefore we develop and describe robust, efficient and sufficiently accurate methods for the numerical solution of systems (4.1) in this chapter. We restrict our attention to one-step methods. These generate approximations y_{k+1} to $y(t_{k+1})$ for $k = 0, 1, \ldots$ in a step by step fashion, starting with $y_0 = y(t_0)$, by using the approximate evolution operator Ψ acting on the last computed approximation y_k,

$$y_{k+1} := \Psi(t_k, \tau_k) y_k \,, \quad k = 0, 1, \ldots .$$

The operator Ψ depends, beside on the method coefficients and the right-hand side function F, on the time t_k and a selected time step size $\tau_k > 0$. The temporal grid points are defined by $t_{k+1} := t_k + \tau_k$. We discuss strategies to choose the time step size τ_k adaptively such that we can control the (local) error in the computation. Otherwise, we will neglect the dependence of the time step size on the step number k for ease of notation and write only τ in the following.
We are especially interested in the numerical integration of the large ODE systems (4.1) by means of splitting techniques. These are based on low-order explicit Runge-Kutta methods and linearly-implicit Rosenbrock-type methods. Both method classes and the necessary theory (local errors,

stability, time step size control) are introduced in Sec. 4.1. Following this introduction we present two splitting techniques, approximate matrix factorization (AMF) and operator splitting (OPS), Sec. 4.2.
The usefulness of splitting techniques becomes evident when we write the vector function F as

$$F(t,y) = F_0(t,y) + F_1(t,y)\,, \tag{4.3}$$

where we have collected all terms from the taxis discretization in F_0 (including corresponding boundary terms) and all diffusion and reaction terms in F_1. We separate these terms because the system $y'(t) = F_1(t,y(t))$ generally requires an implicit (or at least linearly-implicit) treatment because of stiffness, whereas the semi-discrete taxis system $y'(t) = F_0(t,y(t))$, which can be regarded as a discretized nonlinear advection equation, is better solved explicitely because this is often more efficient. The splitting techniques AMF and OPS make use of this separation and treat F_0 and F_1 differently. We can further split F_1 by separating terms of diffusion discretization in different spatial directions and reaction terms,

$$F_1(t,y) = \sum_{j=1}^{d} F_{D_j}(t,y) + F_R(t,y)\,. \tag{4.4}$$

This secondary splitting will be used to considerably reduce linear algebra costs within the schemes to be described and this is also where the order of the components in y becomes significant for efficiency.
In Sec. 3.2 we have introduced the notion of positive ODE systems and defined the class $\mathcal{P}$ of functions F such that (4.1) is a positive ODE system. Further, we have given conditions such that a right-hand side function F of (4.1) is an element of $\mathcal{P}$. In this chapter we always assume that

$$F, F_0, F_{D_j}, F_R \in \mathcal{P}\,.$$

In the previous chapter we have seen that this is possible for TDR systems with a suitable discretization in space and careful treatment of boundary conditions. In Sec. 4.3 we present three different theories from the literature which are concerned with the numerical solution of positive ODE systems. The aim is to obtain methods for the solution of such systems which guarantee nonnegative numerical approximations for reasonably large time steps. We mainly concentrate on the case of low-order ERK methods but also discuss the case of implicit schemes where appropriate. The reason for this special interest in explicit methods is that we want to apply such schemes for the numerical solution of the system $y'(t) = F_0(t,y(t))$ which often generates problems if nonnegative numerical approximations are requested. This is because F_0 corresponds to the taxis part in our models which is present in the cell density equation. The models represent pattern formation processes and hence this density will vary strongly in space and has steep moving fronts which cause positivity problems (lack of spatial smoothness). We derive an ERK method especially suited for the numerical solution of positive ODE systems in this section.
The theory presented and developed in Sec. 4.3 is then applied to the taxis ODE $y' = F_0(y)$ in Sec. 4.4. There we also discuss associated stability properties beside looking at positivity.
In Sec. 4.5 and Sec. 4.6 we detail, based on the theoretical investigations and on stability considerations, specific AMF and OPS schemes which we will use for the simulation of TDR systems. Finally, in Sec. 4.7 we discuss a few different approaches for the solution of system (4.1).

4.1 Runge-Kutta and Rosenbrock-type methods

There are excellent text books covering the theory and application of Runge-Kutta and Rosenbrock-type methods, e.g. [22, 23, 51, 55, 56, 10]. In this section we collect only definitions and results which are required for the understanding of this chapter. For further details (and proofs) we refer to the cited books.

An s-stage Runge-Kutta (RK) method for the solution of (4.1) can be characterized by a coefficient matrix $A = (a_{ij}) \in \mathbb{R}^{s,s}$ (with $a_{ij} = 0$ for all $j \geq i$ in case of explicit Runge-Kutta (ERK) methods), a weight vector $b = (b_i) \in \mathbb{R}^s$ and the knot vector $c = (c_i) := A\mathbb{1} \in \mathbb{R}^s$. In short, such a scheme is in general represented by the pair (A, b) or by its Butcher array

$$\begin{array}{c|c} c & A \\ \hline & b^{\mathsf{T}} \end{array} .$$

A given approximation y_k is advanced by a time step τ to yield y_{k+1} via

$$y_{k+1}^{(i)} = y_k + \tau \sum_{j=1}^{s} a_{ij} F(t_k + c_j\tau, y_{k+1}^{(j)}) \,,\ i = 1(1)s \,, \tag{4.5a}$$

$$y_{k+1} = y_k + \tau \sum_{i=1}^{s} b_i F(t_k + c_i\tau, y_{k+1}^{(i)}) \,. \tag{4.5b}$$

This defines the approximate evolution operator Ψ associated with the RK method (A, b).

We further consider s-stage W-methods (a class of Rosenbrock-type methods introduced by Steihaug and Wolfbrandt [53]) for the solution of (4.1) which are defined by coefficient matrices $A = (a_{ij})$ and $\Gamma = (\gamma_{ij}) \in \mathbb{R}^{s,s}$ (with $a_{ij} = 0$ for $j \geq i$, $\gamma_{ij} = 0$ for $j > i$, and $\gamma_{ii} = \gamma \in \mathbb{R}_+$ for all i), a weight vector $b = (b_i) \in \mathbb{R}^s$, and the knot vector $c = (c_i) := A\mathbb{1} \in \mathbb{R}^s$. A given approximation y_k is advanced by a time step τ to yield y_{k+1} via

$$(I - \tau\gamma T)k_i = \tau F\left(t_k + c_i\tau, y_k + \sum_{j=1}^{i-1} a_{ij}k_j\right) + \tau T \sum_{j=1}^{i-1} \gamma_{ij}k_j \,,\ i = 1(1)s \,, \tag{4.6a}$$

$$y_{k+1} = y_k + \sum_{i=1}^{s} b_i k_i \,. \tag{4.6b}$$

The matrix $T \in \mathbb{R}^{m,m}$ in the method is an arbitrary matrix. These methods require the solution of one linear system per stage for the unknown vector k_i. The system matrix $I - \tau\gamma T$ is the same in all stages so that only one LU-decomposition is required per time step. The methods are said to be *linearly-implicit*. For $T = 0$ the method reduces to the underlying ERK method (A, b). However, it is advantageous to choose T as an approximation to the Jacobian matrix $\frac{\partial F(t,y)}{\partial y}$ at (t_k, y_k) for accuracy and stability reasons. We will use the freedom in the choice of T to incorporate splitting in the scheme and to drastically decrease the linear algebra work per time step.

The stage equations (4.6a) have the matrix T on both sides. To avoid unnecessary matrix-vector multiplications (and to facilitate the approximate matrix factorization (AMF) to be discussed later)

we can replace (4.6) by the equivalent formula ($\widetilde{A} := A\Gamma^{-1}, \widetilde{b}^{\mathsf{T}} := b^{\mathsf{T}}\Gamma^{-1}$, and $\widetilde{\Gamma} := I - \gamma\Gamma^{-1}$)

$$(I - \tau\gamma T)\,\widetilde{k}_i = \tau\gamma F\left(t_k + c_i\tau, y_k + \sum_{j=1}^{i-1}\widetilde{a}_{ij}\widetilde{k}_j\right) + \sum_{j=1}^{i-1}\widetilde{\gamma}_{ij}\widetilde{k}_j\,,\ i = 1(1)s\,, \tag{4.7a}$$

$$y_{k+1} = y_k + \sum_{i=1}^{s}\widetilde{b}_i\widetilde{k}_i\,. \tag{4.7b}$$

This defines the approximate evolution operator Ψ associated with the W-method (A, Γ, b).

Order and order conditions:
A one-step method Ψ is of order $p \in \mathbb{N}$ if the local error $le(t,\tau)$, i.e. the error introduced by one time step of the method, satisfies (for sufficiently smooth problems (4.1))

$$\|le(t,\tau)\| := \|y(t+\tau) - \Psi(t,\tau)y(t)\| \le K\tau^{p+1}\,,\ K \in \mathbb{R}\,.$$

The global error $e(t_k) := y(t_k) - y_k$ is the error of the computed solution after several steps (with initial values $(t_0, y(t_0))$). Suppose that the method Ψ is of order p and we can write the method in the standard form for one-step methods, i.e. $\Psi(t,\tau)y = y + \tau\Phi(t,\tau,y)$, with so-called increment function Φ. The latter is the case for the ERK and W-methods considered here. Further suppose that in a neighbourhood of the exact solution of (4.1) the increment function $\Phi(t,\tau,y)$ is Lipschitz continuous in its last argument (with constant Λ). Then the global error is also of order p and satisfies

$$\|e(t_k)\| \le \widetilde{\tau}^p\frac{K}{\Lambda}\left(e^{\Lambda(t_k - t_0)} - 1\right)\,, \quad \widetilde{\tau} := \max\{\tau_0, \tau_1, \ldots, \tau_{k-1}\}\,. \tag{4.8}$$

RK or W-methods are of order p if their coefficients satisfy certain order conditions. In this work here we are mainly interested in low-order methods ($p = 2, 3$). The order conditions up to $p = 3$ (see [55, p. 39]) for an RK method (A, b) are

$$\sum_{i=1}^{s} b_i = 1\,, \quad \sum_{i=1}^{s} b_i c_i = \frac{1}{2}\,, \quad \sum_{i=1}^{s} b_i c_i^2 = \frac{1}{3}\,, \quad \sum_{i,j=1}^{s} b_i a_{ij} c_j = \frac{1}{6}\,. \tag{4.9}$$

If the first condition is satisfied then we have a method of order one, if also the second condition is satisfied then the order is two, and all four conditions must be satisfied for order three. The order conditions for a W-method (A, Γ, b) are listed for instance in [55, p. 136]). The method should have the same order independent of the choice of T (*W-method property*). Therefore, because of $T = 0$, for the W-method to be of order p, also its underlying ERK method (A, b) must be of order p and the respective order conditions in (4.9) must be satisfied. No additional condition is necessary for order one of the W-method. For order two there is the additional condition

$$\sum_{i=1}^{s} b_i\beta_i = \frac{1}{2} - \gamma\,, \quad \text{where} \quad \beta_{ij} := a_{ij} + \gamma_{ij}\,, \text{ and } \beta_i := \sum_{j=1}^{i-1}\beta_{ij}\,. \tag{4.10}$$

There are three additional conditions for order three (not listed here).

Linear stability properties

We consider (Dahlquist's) test equation $y' = \lambda y$, where $\lambda \in \mathbb{C}$ and apply RK and W-methods ($T := \lambda$ here) to this equation. Performing one time step with size τ of these methods applied to the test equation results in the recursion $y_{k+1} = R(z)y_k$, $z := \lambda\tau$, and $R(z)$ is called the stability function of the method at hand. The stability function is a rational function in z, which reduces to a polynomial for ERK methods. We have

$$R(z) = 1 + zb^{\mathsf{T}}(I - zA)^{-1}\mathbb{1} \quad \text{and} \quad R(z) = 1 + zb^{\mathsf{T}}(I - zB)^{-1}\mathbb{1},\ B := A + \Gamma\,,$$

for RK-methods (A, B) and W-methods (A, Γ, b), respectively. We denote with $\mathcal{S} := \{z \in \mathbb{C} \ : |R(z)| \leq 1\}$ the stability domain of a given method.

The solution of the test equation is stable for all $\lambda \in \mathbb{C}_{-,0}$ and we call a method A-stable if its stability domain $\mathcal{S} \supset \mathbb{C}_{-,0}$. This means that A-stable methods preserve this stability property, i.e. there is no stability restriction on the time step size $\tau > 0$ in the numerical solution of the test equation with $\lambda \in \mathbb{C}_{-,0}$. If the stability function of an A-stable method also satisfies

$$\lim_{\Re z \to -\infty} R(z) = 0$$

then this method is called L-stable. A consequence of this property is that the numerical solution of the test equation with $\Re\lambda \ll -1$ is damped to zero very quickly. This mimics the behaviour of the exact solution of the test equation with $\Re\lambda \ll -1$.

If $\mathcal{S} \supset W_\alpha$ for a value $\alpha \in [0, \pi/2]$, where $W_\alpha := \{z \in \mathbb{C} \ : \ |\arg(-z)| \leq \alpha\}$ is a closed wedge in the left complex half plane, then the corresponding method is called $A(\alpha)$-stable ($L(\alpha)$-stable methods are defined analogously). Obviously, A- and $A(\pi/2)$-stability as well as L- and $L(\pi/2)$-stability coincide. The stability functions of A(α)- or L(α)-stable methods are often called $A(\alpha)$- or $L(\alpha)$-acceptable. A discussion about the relevance and applicability of Dahlquist's test equation for the numerical solution of linear and general ODEs is given in [24].

Time step size control

For a very detailed discussion of step size control mechanisms we refer to [51, p. 334]. There are two major approaches to select the time step size τ_k in a one-step method Ψ adaptively: embedding and Richardson extrapolation. Suppose a time step size τ_k for the current time step is given. Then both approaches estimate the local error le_k in the current time step, that is the difference between $\widehat{y}_{k+1} := \Psi(t_k, \tau_k)y_k$ and the exact solution at $t_k + \tau_k$ of the ODE system with initial values (t_k, y_k). (Note that the local error here differs slightly from the local error in (4.8) because there $y_k = y(t_k)$ is assumed. However, this distinction does not affect the convergence of the method.) Both mechanisms accept the step if a norm of the estimate est_k of le_k is below a user-supplied tolerance; otherwise the step is rejected. In any case, a new step size τ_{new} is predicted, based on est_k, for the next time step of the method. The procedure should ensure that the time steps are sufficiently small in order to meet the tolerance requirement, and also large enough so that the problem is solved efficiently.

We first discuss embedding. Suppose we have a pair of methods Ψ and $\widetilde{\Psi}$ of orders p and $p + 1$, respectively. These methods generate approximations $\widehat{y}_{k+1} = \Psi(t_k, \tau_k)y_k$ and $\widetilde{y}_{k+1} = \widetilde{\Psi}(t_k, \tau_k)y_k$ to $y(t_{k+1})$. It is common that the methods Ψ and $\widetilde{\Psi}$ differ only in the weight vector b for RK and Rosenbrock-type methods; the coefficient matrices remain unchanged. This implies that $\widehat{y}_{k+1}$ and

$\widetilde{y}_{k+1}$ can be computed by approximately the same amount of work which is necessary to compute just a single approximation.
Then $le_k = \widehat{y}_{k+1} - \widetilde{y}_{k+1} + \mathcal{O}\left(\tau_k^{p+2}\right)$ and $est_k := \widehat{y}_{k+1} - \widetilde{y}_{k+1} = \mathcal{O}\left(\tau_k^{p+1}\right)$ provides an asymptotically correct, computable estimate of the local error in the lower order solution $\widehat{y}_{k+1}$. The user supplies relative ($RTOL_i \geq 0$) and absolute ($ATOL_i > 0$) tolerances for each component of the solution vector ($i = 1(1)m$). Then we compute the mixed (relative and absolute) error indicator

$$err := \left(\frac{1}{m} \sum_{i=1}^{m} \left(\frac{est_{k,i}}{RTOL_i |y_{k,i}| + ATOL_i} \right)^2 \right)^{\frac{1}{2}} . \tag{4.11}$$

Following the *error per step* criterion, the time step is accepted if $err \leq 1$ and rejected otherwise. If the step is accepted then we proceed with (in the methods to be presented here) $y_{k+1} := \widetilde{y}_{k+1}$, i.e. we use the higher order approximation $\widetilde{y}_{k+1}$ although the error is only estimated for the lower order approximation $\widehat{y}_{k+1}$. This is known as *local extrapolation*.
It remains to derive the new time step size τ_{new} such that, in case of a rejected step, the recomputed step with $\tau_k := \tau_{new}$, or, in case of an accepted step, the next step with $\tau_{k+1} := \tau_{new}$ is likely to pass the error test. A widely accepted definition is

$$\tau_{new} := \tau_k \min\{f_{max}, \max\{f_{min}, f_{saf}\ err^{-1/(p+1)}\}\} . \tag{4.12}$$

In this formula $0 < f_{min} \leq f_{max}$ are the minimum and maximum step size change ratios, and $0 \leq f_{saf} < 1$ is a safety factor. (The choice $f_{saf} = 0, f_{min} = f_{max} = 1$ (together with very relaxed tolerance requirements) leads to a constant step size scheme.) A closer examination of the error per step control with local extrapolation (as described) reveals that under suitable assumptions the global error in the solution is proportional to the tolerance requirements. Without local extrapolation the global error is only proportional to $TOL^{p/(p+1)}$, see [51, p. 350].
We now discuss the second step size control mechanism, Richardson extrapolation. In fact, the only difference to the algorithm described above is how we obtain the estimate est_k of the local error le_k. Suppose we have a method Ψ of order p. Then we compute two approximations

$$\widehat{y}_{k+1} := \Psi(t_k, \tau_k) y_k \quad \text{and} \quad \widetilde{y}_{k+1} := \Psi\left(t_k + \frac{\tau}{2}, \frac{\tau}{2}\right) \Psi\left(t_k, \frac{\tau}{2}\right) y_k .$$

The local error with respect to the solution $\widetilde{y}_{k+1}$ can then be estimated by

$$est_k := \frac{1}{2^p - 1} (\widetilde{y}_{k+1} - \widehat{y}_{k+1}) .$$

In the case of an accepted step, we set $y_{k+1} := \widetilde{y}_{k+1}$ (doubling [51, p. 364]). Note that this is not local extrapolation because the order is not increased; we use only the supposedly more accurate solution to advance the integration. Local extrapolation leads to setting $y_{k+1} := \widetilde{y}_{k+1} + est_k$. However, in this case properties like stability of the resulting scheme must be considered anew. We do not apply local extrapolation together with Richardson extrapolation in this work.

4.2 Introduction to approximate matrix factorization and operator splitting

Verwer et al. [62] successfully applied the 2-stage, second-order W-method ROS2 to advection–diffusion–reaction problems from atmospheric air pollution modelling. The AMF methodology applied in that paper also appears to be of interest for the time integration of the MOL-ODEs obtained as semi-discretizations of TDR systems (2.5).
Low-order W-methods are efficient for a wide range of stiff ODE problems, see e.g. [23, 55]. However, applying an s-stage W-method (in the form (4.6) or (4.7)) involves s linear solves with the matrix $I - \tau\gamma T$. Consequently, choosing T equal to the full Jacobian

$$\frac{\partial F(t_k, y_k)}{\partial y_k} = \frac{\partial F_0(t_k, y_k)}{\partial y_k} + \sum_{j=1}^{d} \frac{\partial F_{D_j}(t_k, y_k)}{\partial y_k} + \frac{\partial F_R(t_k, y_k)}{\partial y_k}$$

seems not practical because this matrix, although sparse, has a large bandwidth which grows with decreasing spatial grid width (for $d > 1$). This makes the direct solution of these systems prohibitively expensive. Moreover, due to the limiter functions used in the taxis discretization, F is only Lipschitz continuous so that the Jacobian might not even exist. The same situation occurs in the air pollution application [62] where ROS2 is applied with a matrix T approximating the true Jacobian. For the TDR models we use a similar approximation which yields the following choice for the matrix $I - \tau\gamma T$,

$$I - \tau\gamma T := \left(I - \tau\gamma \frac{\partial F_R(t_k, y_k)}{\partial y_k}\right) \prod_{j=1}^{d} \left(I - \tau\gamma \frac{\partial F_{D_j}(t_k, y_k)}{\partial y_k}\right). \tag{4.13}$$

By rearranging the right-hand side of (4.13), we see that this definition implies that we use a matrix T which depends on the time step size τ. The order of W-methods is independent of the matrix T. However, to take advantage of the factorization (4.13), it is now important that we use the W-methods in the transformed form (4.7). Using form (4.6) of the W-method would require to explicitly compute the matrix T from the factorization (4.13).
The approximation (4.13) is obtained in two steps. Firstly, we have neglected the taxis Jacobian F_0 which overcomes the possible difficulty of non-existence. This choice further underlies the assumption that explicit methods are in general more efficient than implicit ones when applied to the taxis ODE $y'(t) = F_0(t, y(t))$. Secondly, we have approximated the remainder matrix

$$I - \tau\gamma \frac{\partial F_1(t_k, y_k)}{\partial y_k} \tag{4.14}$$

by the factorized expression (4.13). With this factorization we avoid to solve linear systems which is still expensive (if $d > 1$) because the Jacobian of F_1 has a bandwidth $\mathcal{O}\left(h^{-(d-1)}\right)$, h denoting the spatial grid size. For efficiency it is important that the matrices involved are banded with a small bandwidth independent of h. This property is especially profitable for the fine spatial resolutions required in our models to resolve steep fronts for the cell density and it holds with (4.13). This can be seen easily. If the components in y are arranged such that all approximations (cell density n and chemical concentrations $\mathbf{c}$) corresponding to one grid cell form a block in y then the Jacobian

of F_R is a block diagonal matrix with block size $(l+1) \times (l+1)$ and we can solve for each block independently. If the components in y are arranged such that the approximations to the cell density n in all grid cells form a block and also all approximations to each of the l chemical concentrations in $\mathbf{c}$ then the Jacobians of F_{D_j}, $j = 1(1)d$, are block diagonal matrices. Further, for each $j = 1(1)d$, the components of each of the $l+1$ blocks in y can be arranged such that the corresponding submatrix in the Jacobian of F_{D_j} is tridiagonal, i.e. with this special arrangement of y we can solve the linear systems involving the Jacobian F_{D_j} efficiently. Altogether, solving linear systems with the factorized matrix (4.13) efficiently amounts to a sequential process where we have to rearrange the right-hand side vector of the system appropriately after each sub-step.
The factorization (4.13) is known as 'Approximate Matrix Factorization' (AMF) which has been used for a long time already for solving multi-space dimensional time-dependent PDE problems, see e.g. [5, 11, 28, 32, 48].
The application of AMF in W-methods does not affect the order of the method but it does of course affect the stability of the original W-method used with $T = F'(y_k)$. In [62] it is argued that with (4.13) the stability of the resulting ROS2 method with AMF is mainly governed by the stability of the modified Euler method (which is the underlying ERK method of ROS2) applied to the F_0 part only. We investigate this issue in Section 4.5 for the specific W-methods presented there.
If the split matrices in (4.13) do not commute then the order of the factors can be important for the performance of the method and the best choice can be problem specific. Our choice in (4.13) with the Jacobian of F_R as the first factor is guided by the assumption that the subsequent factors (corresponding to diffusion) will smooth the stage solutions and hence also result in a smoother step solution.

Remark 1 *The secondary splitting defined by (4.13) can be avoided and we can simply use (4.14) as matrix $I - \tau\gamma T$ in the W-methods. Direct methods are not suitable for the solution of the resulting systems because of the large bandwidth but we could employ iterative solvers. This immediately raises the issues of convergence and preconditioners. We do not consider iterative methods in this work because with the secondary splitting and AMF we obtain linear systems which can be solved very efficiently. However, the application of iterative linear system solvers can be a topic of further research.*

Whereas the W-methods applied with AMF perform a splitting at the level of linear algebra, it is also possible to directly split Eq. (4.1) at the problem level, that is, to apply operator splitting (OPS). Like AMF, operator splitting is a popular approach for solving multi-space dimensional time-dependent PDE problems. Operator splitting has been considered in [14] for the tumour-induced angiogenesis Model 2 (different initial TAF concentration). The method proceeds as follows. Given an approximation y_k at time t_k and a step size τ, we compute

$$y_{k+1} = \Psi_0\left(t_k + \frac{\tau}{2}, \frac{\tau}{2}\right) \Psi_1\left(t_k, \tau\right) \Psi_0\left(t_k, \frac{\tau}{2}\right) y_k \,. \tag{4.15}$$

The operators Ψ_0 and Ψ_1 are approximate evolution operators for the split functions F_0 and F_1 (see Eq. (4.3)), respectively. Specifically, $\Psi_i(\widetilde{t}, \tau)u$ approximates the solution of the IVP

$$y'(t) = F_i(t, y(t)) \,, \quad t \geq \widetilde{t} \,, \qquad y(\widetilde{t}) = u \,,$$

at $t = \widetilde{t} + \tau$. The formula (4.15) is known as Strang-splitting [54]. If the right-hand side function is linear and autonomous, $F_i(t, y) = A_i y\,, A_i \in \mathbb{R}^{m,m}$, and the matrices A_i commute then the solution of a splitting step (4.15) coincides with the exact solution provided that the subproblems are solved exactly; otherwise a so-called splitting error is introduced [39]. If the operators Ψ_i are at least second-order accurate approximations of the exact evolution operators, i.e. we use second order methods for the solution of the subproblems, then the order of the approximation (4.15) equals two. The stability of (4.15) is determined by the stability properties of Ψ_0 and Ψ_1.
It is effective to select an explicit method Ψ_0 and an implicit method Ψ_1. We will employ ERK methods as explicit schemes Ψ_0. W-methods applied with AMF appear to be of interest for the application as implicit schemes Ψ_1.
Operator splitting is applied in the order given in (4.15) because then we use only half the step size of the splitting step for the explicit method. This doubles the stability (and positivity domain, see next section) of the explicit method and hence is expected to lead overall to less time steps and subsequently less computational effort. Other splitting orders are possible, e.g.

$$y_{k+1} = \Psi_1\left(t_k + \frac{\tau}{2}, \frac{\tau}{2}\right) \Psi_0\left(t_k, \tau\right) \Psi_1\left(t_k, \frac{\tau}{2}\right) y_k\,,$$

$$\text{or} \qquad y_{k+1} = \Psi_1\left(t_k + \frac{\tau}{2}, \frac{\tau}{2}\right) \Psi_0\left(t_k + \frac{\tau}{2}, \frac{\tau}{2}\right) \Psi_0\left(t_k, \frac{\tau}{2}\right) \Psi_1\left(t_k, \frac{\tau}{2}\right) y_k\,,$$

but a numerical assessment revealed that they are not advantageous for our problem class (more time steps, greater amount of work per time step).

4.3 Positive methods for positive ODE systems

Positive ODE systems (4.1) arise in a great variety of applications, e.g. when modelling chemical reactions, in the semi-discretization of air pollution [31] and, as we have seen, biomathematical models. The quantity $y(t)$ usually describes the concentration or density of some species. In such a situation we are naturally interested in obtaining nonnegative numerical approximations y_k of the solution $y(t_k)$ at discrete time points t_k by an appropriate numerical method. This requirement is not met in general. We consider one-step methods for the solution of (4.1) here; for multi-step methods see for instance [29, 6].
In order to characterize positivity properties of numerical schemes we give the definition of *positive one-step methods* from [27].

Definition 2 *Let there be given a one-step method for the solution of (4.1), a subclass $\mathcal{F} \subset \mathcal{P}$ and a threshold $0 < \tau^+ \leq \infty$. The method is called* positive *on $\mathcal{F}$ with threshold τ^+ if the numerical approximations obtained by the method are uniquely defined and are nonnegative whenever the method is applied to the IVP (4.1) with any $F \in \mathcal{F},\ t_0 \in \mathbb{R},\ y_0 \geq 0$ and with step size τ satisfying $0 < \tau \leq \tau^+$. If this holds with $\tau^+ = \infty$ then the method is called* unconditionally positive, *otherwise* conditionally positive, *on $\mathcal{F}$.*

Obviously, the approximations are always uniquely defined for explicit methods.
We say that a method taken from a class of methods has *optimal positivity* on a certain problem class $\mathcal{F}$ if it is a positive method on $\mathcal{F}$ with a step size restriction τ^+ and all other methods from the given class have, for positivity on $\mathcal{F}$, a step size restriction $\widetilde{\tau}^+ \leq \tau^+$.

It is useful to define subclasses of the class $\mathcal{P}$ of positive problems for the following investigations. Let $g(t) \geq 0$ be a given continuous, vector-valued function, $\alpha \in \mathbb{R}_{+,0}$ and define classes of linear functions

$$\mathcal{L}_g^+(\alpha) := \{F | F(t,y) = Py + g(t) \text{ where } P \in \mathbb{R}^{m,m},\ P + \alpha I \geq 0\}. \tag{4.16}$$

IVPs with right-hand side functions taken from these classes are positive (apply Theorem 1).
Further, following [27], we consider classes of nonlinear, dissipative functions and define subclasses by using the so-called circle condition [37]. The right-hand side F of an IVP (4.1) fulfils the circle condition with constant $\rho \in \mathbb{R}_+$ in some vector norm $\|\cdot\|$ (e.g. the p-norms) if

$$\|\rho(\widetilde{y} - y) + (F(t,\widetilde{y}) - F(t,y))\| \leq \rho\|\widetilde{y} - y\| \quad \text{for all } t \in \mathbb{R}, y, \widetilde{y} \in \mathbb{R}^m. \tag{4.17}$$

Now, for any $\rho \in \mathbb{R}_+$, define

$$\mathcal{D}^+(\rho) := \{F | F \in \mathcal{P} \text{ and } F \text{ satisfies (4.17) with constant } \rho \text{ in some } p\text{-norm }, p \in [1,\infty]\}. \tag{4.18}$$

The following lemma characterizes the hierarchy in the parametrized classes $\mathcal{L}_g^+(\alpha)$ and $\mathcal{D}^+(\rho)$.

Lemma 7 *Let $\alpha_1 \geq \alpha_2 \geq 0$ and $\rho_1 \geq \rho_2 > 0$. Then we have*

$$\mathcal{L}_g^+(\alpha_2) \subset \mathcal{L}_g^+(\alpha_1) \subset \mathcal{P} \quad \textit{and} \quad \mathcal{D}^+(\rho_2) \subset \mathcal{D}^+(\rho_1) \subset \mathcal{P}.$$

Proof The statements follow directly from the definitions of $\mathcal{L}_g^+(\alpha)$ and $\mathcal{D}^+(\rho)$. □
In the following subsections we consider the positivity of RK methods applied to problems from $\mathcal{L}_g^+(\alpha)$ and $\mathcal{D}^+(\rho)$ (Sec. 4.3.1 and Sec. 4.3.2, respectively), and the positivity of ERK schemes applied to general problems (4.1) with $F \in \mathcal{P}$ in Sec. 4.3.3. One of the main goals is to identify a low-order ERK method with appropriate positivity properties for a broad range of positive ODE problems. We summarize the results on ERK methods in Sec. 4.3.4.

4.3.1 Positivity of RK and W-methods on $\mathcal{L}_g^+(\alpha)$

The results of this section are based on work by Bolley and Crouzeix [6], and Kraaijevanger and van de Griend [36, 19].
RK methods or W-methods (using $T := P$) applied with fixed step size τ to a problem taken from the class $\mathcal{L}_g^+(\alpha)$ yield the recursion

$$y_{k+1} = R(\tau P)y_k + \tau \sum_{i=1}^{s} R_i(\tau P) g(t_k + c_i\tau)\,. \tag{4.19}$$

Here $R(z)$ and $R_i(z)$ are rational functions with real coefficients; $R(z)$ is the stability function of the method. These functions are polynomials in the case of ERK methods. For a definition and some properties of matrix functions (e.g. $R(\tau P), R_i(\tau P)$ above) see Sec. A.2.
We need the concept of absolute monotonicity of rational functions and the so-called threshold factor, see e.g. [19, 36, 37], for the study of positivity of the recursion (4.19).

Definition 3 *A rational function R is called* absolutely monotonic *at a point $z \in \mathbb{R}$ if $R(z)$ is defined and R as well as all its derivatives are nonnegative in z. The function R is called absolutely monotonic on the interval $I \subset \mathbb{R}$ if R is absolutely monotonic at every $z \in I$.*

Definition 4 *The* threshold factor of a rational function R, *denoted by $T(R)$, is defined as*

$$T(R) := \sup\{r \mid r \in \mathbb{R}_{+,0} \text{ and } R \text{ is absolutely monotonic on } [-r, 0]\}\,.$$

This definition differs slightly from the definition given in [19]. Here we have $T(R) = 0$ if $R(z)$ is absolutely monotonic in $z = 0$ but not in a left neighbourhood of $z = 0$. $T(R)$ is not defined if $R(z)$ is not even absolutely monotonic in $z = 0$. In contrast, the definition in [19] would result in $T(R) = 0$ in both cases.

Lemma 8 *If R is a polynomial then the threshold factor $T(R)$ is given by*

$$T(R) = \sup\{r \mid r \in \mathbb{R}_{+,0} \text{ and } R \text{ is absolutely monotonic in } z = -r)\}\,.$$

Proof The statement follows with Lemma 3.1 from [37]. □
Except for the case $T(R) = 0$ (see the comments after Definition 4), the statement of Lemma 8 coincides with the definition of the threshold factor of polynomials in [36].
We have $T(R) = \infty$ for the stability function of the implicit Euler method, $R(z) = (1 - z)^{-1}$. However, Bolley and Crouzeix [6] show that $R(z) = \exp(z) + \mathcal{O}\left(z^{p+1}\right)$ with $p \geq 2$ for $z \to 0$ implies $T(R) < \infty$.
M- and $\bar{M}$-matrices are important in the theory of Bolley and Crouzeix [6] and are introduced now.

Definition 5 ([37, p. 497], [6, p. 241]) *A matrix $B \in \mathbb{R}^{m,m}$ is said to be an M-matrix if $b_{ij} \leq 0$ for all $i \neq j$, B is nonsingular, and $B^{-1} \geq 0$.*
A matrix $A \in \mathbb{R}^{m,m}$ is said to be an $\bar{M}$-matrix if for all $\alpha \in \mathbb{R}_+$ the matrix $\alpha I + A$ is an M-matrix.

The next two lemmas give useful characterizations of M-matrices.

Lemma 9 ([21, p. 151]) *Let $B \in \mathbb{R}^{m,m}$ such that $b_{ij} \leq 0$ for all $i \neq j$ and define $D := \operatorname{diag}(B)$. Then the statements*

$$B \text{ is nonsingular and } B^{-1} \geq 0\,,$$

and

$$(1) \quad b_{ii} > 0 \text{ for all } i = 1(1)m\,, \quad (2) \quad M := I - D^{-1}B \geq 0\,, \quad \text{and } (3) \quad \rho(M) < 1\,,$$

where $\rho(M)$ denotes the spectral radius of the matrix M, are equivalent.

Lemma 10 ([49, p. 30]) *Let $B, C \in \mathbb{R}^{m,m}$ be two matrices which satisfy $B \leq C$ and $c_{ij} \leq 0$ for all $i \neq j$. Then if B is an M-matrix, so is the matrix C.*

Let $C := \alpha I + B$. Then the following corollary follows from Lemma 10.

Corollary 3 *If $B \in \mathbb{R}^{m,m}$ is an M-matrix then B is an $\bar{M}$-matrix.*

We want to characterize $\bar{M}$-matrices and therefore need the Perron–Frobenius theorem for nonnegative matrices, see e.g. [21, p. 150] or [43].

Theorem 3 (Perron–Frobenius) *Let $A \in \mathbb{R}^{m,m}$ be a nonnegative matrix, $A \geq 0$. Then the spectral radius $\rho(A) \geq 0$ is an eigenvalue of A and for this eigenvalue exists a nonnegative eigenvector.*

Theorem 4 *Let $A \in \mathbb{R}^{m,m}$ be a real square matrix. Then A is an $\bar{M}$-matrix if and only if*

1. *$a_{ij} \leq 0$ for all $i \neq j$,*
2. *$a_{ii} \geq 0$ for all $i = 1(1)m$, and*
3. *A has no eigenvalue $\lambda \in (-\infty, 0)$.*

Proof
Necessity: Suppose A is an $\bar{M}$-matrix. This implies $B := \alpha I + A$ is an M-matrix for all $\alpha > 0$ and hence, $b_{ij} \leq 0$ for all $i \neq j$, B is nonsingular, and, by Lemma 9, $b_{ii} > 0$ for all i. Therefore conditions 1. and 2. are satisfied for A. The regularity of B implies that $Bx = 0$ has the trivial solution only and therefore, for arbitrary $\alpha > 0$, $Ax = -\alpha x$ is satisfied for $x = 0$ only, i.e. A has no eigenvalue $\lambda \in (-\infty, 0)$ and condition 3. is satisfied.
Sufficiency: We prove that conditions 1., 2. and 3. imply that A is an $\bar{M}$-matrix, i.e. that $B := \alpha I + A$ is an M-matrix for all $\alpha > 0$. First note that $b_{ij} \leq 0$ for all $i \neq j$ and all $\alpha > 0$ by condition 1., and further $b_{ii} > 0$ for all i and all $\alpha > 0$ by condition 2.
We use Lemma 9 to show that B is nonsingular and $B^{-1} \geq 0$ for all $\alpha > 0$. Let $D := \mathrm{diag}(B)$ and define $M := I - D^{-1}B$. Then we obtain $M = -(\mathrm{diag}(\alpha I + A))^{-1}\widetilde{A}$, where $\widetilde{A} := A - \mathrm{diag}(A)$. Therefore, $M \geq 0$ by conditions 1. and 2. Denote the spectral radius of M for a given value of $\alpha > 0$ by $\rho(M, \alpha)$. By Theorem 3 we have that $\rho(M, \alpha) \geq 0$ is an eigenvalue of M. We will show that $\rho(M, \alpha) < 1$ for all $\alpha > 0$. $\rho(M, \alpha)$ depends continuously on α and $\rho(M, \infty) = 0$. Hence, if there exists $\alpha > 0$ such that $\rho(M, \alpha) > 1$ then there exists $\widetilde{\alpha} > 0$ such that $\rho(M, \widetilde{\alpha}) = 1$. We demonstrate that $\rho(M, \widetilde{\alpha}) = 1$ leads to a contradiction. $\rho(M, \widetilde{\alpha})$ is an eigenvalue of M and hence there exists a non-trivial vector x such that $Mx = x$. This leads to $\widetilde{A}x = -\mathrm{diag}(\widetilde{\alpha}I + A)x$ and subsequently to $Ax = -\widetilde{\alpha}x$. Hence A has a negative eigenvalue and this contradicts condition 3. Therefore, $\rho(M, \alpha) \neq 1$ for all $\alpha > 0$ and, by continuity, $\rho(M, \alpha) < 1$ for all $\alpha > 0$. Now, with Lemma 9, follows that B is nonsingular and $B^{-1} \geq 0$ for all $\alpha > 0$. Hence, B is an M-matrix for all $\alpha > 0$ and this completes the proof. □
We note that condition 3. of Theorem 4 ensures that the stage equations in a W-method applied with $T = -A$ have unique solutions independent of the time step size $\tau > 0$.
We can now formulate results of Bolley and Crouzeix [6].

Lemma 11 ([6, Lemma 3]) *Let $R(z)$ be a rational function with threshold factor $T(R) \geq 0$. $R(A) \geq 0$ for any matrix A satisfying $-A$ is an $\bar{M}$-matrix and $\max_i\{-a_{ii}\} \leq \mu$ if and only if $T(R) \geq \mu$.*

Theorem 5 ([6, Theorem 2]) *If the rational functions $R(z)$ and $R_i(z)$ of scheme (4.19) have threshold factors $T(R), T(R_i) \geq \mu$ and if $\tau^+ \in \mathbb{R}_+$ satisfies the condition $\tau^+\alpha \leq \mu$ for a value $\alpha \in \mathbb{R}_{+,0}$ then the scheme (4.19) is positive on the class $\{F \in \mathcal{L}_g^+(\alpha) \mid F(t,y) = Py + g(t),\ -P$ is an $\bar{M}$-matrix$\} \subset \mathcal{L}_g^+(\alpha)$ with threshold τ^+.*

We observe that the time step restriction for positivity of a scheme (4.19) is proportional to a method dependent constant (the threshold factor here) and at the same time proportional to the inverse of a problem dependent constant (the class parameter α here). This will also be the case for the other positivity concepts which we are going to discuss.
If we only consider polynomials $R(z)$ in Lemma 11 then the restriction to matrices A such that $-A$ is an $\bar{M}$-matrix can avoided.

Lemma 12 *Let $R(z)$ be a polynomial with threshold factor $T(R) \geq 0$. $R(A) \geq 0$ for any matrix A satisfying $a_{ij} \geq 0$ for all $i \neq j$ and $\max\{0, \max_i\{-a_{ii}\}\} \leq \mu$ if and only if $T(R) \geq \mu$.*

Proof
Necessity: The matrices A allowed in Lemma 11 are also allowed in this lemma. Further, the polynomial R here can be regarded as a rational function R in Lemma 11. Hence, necessity follows from Lemma 11.
Sufficiency: Because $\max\{0, \max_i\{-a_{ii}\}\} \leq \mu$ we obtain $B := \mu I + A \geq 0$. The series

$$g(z) := R(-\mu) + \ldots + \frac{R^{(k)}(-\mu)}{k!} z^k + \ldots$$

converges for every $z \in \mathbb{C}$ because for k large enough the kth derivative of R vanishes. The coefficients $\frac{R^{(k)}(-\mu)}{k!}$ are nonnegative (absolute monotonicity of R in $-\mu$). Hence we have, see Lemma 21 in Sec. A.2,

$$R(A) = r(-\mu)I + \ldots + \frac{R^{(k)}(-\mu)}{k!}(\mu I + A)^k + \ldots$$

and because of $\mu I + A \geq 0$ we have $R(A) \geq 0$. □
With the help of this lemma we can now also reformulate Theorem 5 for the case of explicit methods (4.19).

Theorem 6 *If the function $R(z)$ and $R_i(z)$ of scheme (4.19) are polynomials having threshold factors $T(R), T(R_i) \geq \mu$ and if $\tau^+ \in \mathbb{R}_+$ satisfies the condition $\tau^+\alpha \leq \mu$ for a value $\alpha \in \mathbb{R}_{+,0}$ then the scheme (4.19) is positive on the class $\mathcal{L}_g^+(\alpha)$ with threshold τ^+.*

In the next two subsections we will study the threshold factors of low-order polynomials (i.e. the stability functions of low-order ERK methods) and of restricted Padé approximations (i.e. stability functions of some W-methods).

4.3.1.1 Threshold factors of polynomials

The absolute monotonicity of the stability polynomial of an ERK method is crucial with respect to the allowable time step size in order to guarantee positivity of the method when applied to the problem class $\mathcal{L}_0^+(\alpha)$. This can be seen from Theorem 6. The absolute monotonicity of polynomials is studied in [36] and it is stated that s-stage ERK methods of order $p = s$ have a threshold factor $T(R) = 1$, whereas s-stage ERK methods of order $p = s - 1$ can have a threshold factor $T(R) = 2$. This means that, at the cost of just one matrix-vector product, the allowable time step size with respect to positivity of the method is doubled. Further, [36] gives the optimal stability polynomials $R_{s,p}^+$ in these two cases:

$$R_{s,s}^+(z) = T_s(z) \text{ for } s \geq 1 \quad \text{and} \quad R_{s,s-1}^+(z) = T_{s-1}(z) + \frac{1}{2}\frac{z^s}{s!} \text{ for } s \geq 2\,,$$

where $T_s(z) := \sum_{i=0}^{s} \frac{z^i}{i!}$ is the Taylor polynomial of degree s of $\exp(z)$.
We are interested in second- or third-order methods here. Numerical experiments in [36] demonstrate that (on a linear test problem) the 3-stage method of order two performs more efficient with respect to positivity compared to the s-stage methods of order $p = s$ for $s = 2, 3$ (optimal stability polynomial $R_{s,p}^+$ for positivity on $\mathcal{L}_0^+(\alpha)$ in each case). Therefore we consider 3-stage explicit Runge-Kutta methods of order two with optimal stability polynomial $R_{3,2}^+$ for positivity on $\mathcal{L}_0^+(\alpha)$ in this section. We will use the free parameters in this class of methods to satisfy positivity conditions for nonlinear problem sets and further order conditions.
Consider a 3-stage ERK method (A, b). The conditions for order two are the first two conditions in (4.9). The stability polynomial of a 3-stage ERK method of order two is $R_{3,2}(z) = 1 + z + \frac{1}{2}z^2 + b_3a_{32}a_{21}z^3$. On the other hand, the optimal stability polynomial for 3-stage ERK methods of order two with respect to positivity on the problem class $\mathcal{L}_0^+(\alpha)$ is $R_{3,2}^+(z) = 1 + z + \frac{1}{2}z^2 + \frac{1}{12}z^3$, see above. Hence, beside the two order conditions, the parameters of the method have to satisfy $b_3a_{32}a_{21} = \frac{1}{12}$. Solving for these three conditions results in the class of 3-stage, second-order ERK methods with optimal positivity on the problem class $\mathcal{L}_0^+(\alpha)$. We refer to this class of methods as (Class A) and their Butcher array is given below; denote $\gamma := b_3a_{32}$.

$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{12\gamma} & 0 & & \\ \frac{1}{b_3}\left(\frac{1}{2} - \frac{b_2}{12\gamma} - \gamma\right) & \frac{\gamma}{b_3} & 0 & \\ \hline & 1 - b_2 - b_3 & b_2 & b_3 \end{array} \qquad \begin{aligned} b_2, b_3, \gamma &\in \mathbb{R}\,, \\ b_3, \gamma &\neq 0\,. \end{aligned} \qquad \text{(Class A)}$$

This class of methods forms the basis for all our further investigations regarding positivity of ERK schemes. In Sec. 4.3.2 and Sec. 4.3.3 we discuss two approaches for positivity on nonlinear problem sets and we identify a unique method from (Class A) which has optimal positivity for both approaches simultaneously. This method can also be shown to have optimal positivity on the problem class $\mathcal{L}_g^+(\alpha)$, see Sec. 4.3.4.
We can also use the free parameters b_2, b_3 and γ in (Class A) to satisfy one order three condition (with the aim of improving the accuracy of the scheme). The third-order condition $\sum_{i,j} b_i a_{ij} c_j = b_3a_{32}a_{21} = \frac{1}{6}$ cannot be satisfied because of the condition on the stability polynomial. However, the other third-order condition $\sum_i b_i c_i^2 = \frac{1}{3}$ can be satisfied (resulting in a third-order scheme for

quadrature problems). We have $b_3c_3 = \frac{1}{2} - b_2c_2$. Substituting this in the third-order condition yields $\frac{1}{3} = b_2c_2^2 + \frac{1}{b_3}\left(\frac{1}{2} - b_2c_2\right)^2$. Employing $c_2 = \frac{1}{12\gamma}$, we arrive after some calculations at the methods of (Class B) whose Butcher array is given below. In Sec. 4.3.4, we state some results concerning positivity properties of this class of methods applied to nonlinear problems.

$$\begin{array}{c|ccc} 0 & & & \\ \frac{1}{12\gamma} & 0 & & \\ \frac{1}{b_3}\left(\frac{1}{2} - \frac{b_2}{12\gamma} - \gamma\right) & \frac{\gamma}{b_3} & 0 & \\ \hline & 1 - b_2 - b_3 & b_2 & \frac{(6\gamma - b_2)^2}{48\gamma^2 - b_2} \end{array} \qquad \begin{array}{rcl} b_2, \gamma & \in & \mathbb{R}\,, \\ \gamma & \neq & 0\,, \\ b_2 & \neq & 6\gamma\,, \\ b_2 & \neq & 48\gamma^2\,. \end{array} \qquad \text{(Class B)}$$

4.3.1.2 Threshold factors of restricted Padé approximations

Absolute monotonicity of rational functions R and their threshold factors $T(R)$ are studied in [19]. In this work we consider s-stage W-methods and if these methods have an order $p \geq s$ then their stability functions $R(z)$ are so-called restricted Padé approximations to $\exp(z)$ for $z \to 0$.

Definition 6 ([55, p. 142]) *A rational function $R(z) = (1 - \gamma z)^{-s} P(z)$, where $s \in \mathbb{N}, \gamma \in \mathbb{R}_+$, and $P(z)$ is a polynomial of degree r, satisfying*

$$|R(z) - \exp(z)| = \mathcal{O}\left(z^{r+1}\right) \quad \textit{for} \quad z \to 0\,,$$

i.e. the approximation order is at least r, is called a restricted Padé approximation to $\exp(z)$.

Lemma 13 ([55, pp. 142]) *The stability function $R(z)$ of an s-stage W-method of order $p \geq s$ is uniquely defined by the method parameter $\gamma \in \mathbb{R}_+$. $R(z)$ is a restricted Padé approximation to $\exp(z)$ with approximation order $r \geq s$ and is given by*

$$R(z) = (1 - \gamma z)^{-s} \sum_{j=0}^{s} z^j \sum_{i=0}^{j} \binom{s}{i} \frac{(-\gamma)^i}{(j-i)!}\,.$$

In the remainder of this section we discuss the absolute monotonicity of the stability function of s-stage W-methods with order $p \geq s$ for $s = 1$ and $s = 2$.

Theorem 7 *The stability function*

$$R(z) = \frac{1 + (1-\gamma)z}{1 - \gamma z}, \quad \gamma > 0\,,$$

of a 1-stage W-method of order $p \geq 1$ has a threshold factor $T(R) = \infty$ for $\gamma \geq 1$ and a threshold factor $T(R) = (1-\gamma)^{-1}$ for $\gamma \in (0, 1)$.

Proof We have $1 - \gamma z > 0$ if $\gamma > 0$ and $z \leq 0$. If $\gamma \geq 1$ then $1 + (1-\gamma)z \geq 0$ and hence also $R(z) \geq 0$ for all $z \leq 0$. If $\gamma \in (0, 1)$ then $1 + (1-\gamma)z \geq 0$ and hence also $R(z) \geq 0$ for all $z \in [-(1-\gamma)^{-1}, 0]$.

It remains to show that derivatives of $R(z)$ do not require further restrictions on z in order to have nonnegative values. We obtain $R'(z) = (1-\gamma z)^{-2}$. This is the square of the stability function of the implicit Euler method which is absolutely monotonic for all $z \leq 0$. Thus $R'(z)$ is absolutely monotonic for all $z \leq 0$. This completes the proof of the theorem.
Note: The kth derivative ($k \geq 1$) of $R(z)$ is given by $R^{(k)}(z) = k!\gamma^{k-1}(1-\gamma z)^{-(k+1)}$. This shows absolute monotonicity of $R'(z)$ for all $z \leq 0$, $\gamma > 0$ without using results about the implicit Euler scheme. □
We now turn our attention to 2-stage W-methods of order $p \geq 2$. We obtain their stability function from Lemma 13,

$$R(z) = \frac{1 + (1-2\gamma)z + \left(\frac{1}{2} - 2\gamma + \gamma^2\right) z^2}{(1-\gamma z)^2}, \quad \gamma > 0. \tag{4.20}$$

This function is A-acceptable for $\gamma \geq \frac{1}{4}$, L-acceptable for $\gamma_1 = 1 - \frac{1}{2}\sqrt{2}$ and $\gamma_2 = 1 + \frac{1}{2}\sqrt{2}$, and of approximation order three to $\exp(z)$ for $\gamma = \frac{1}{6}(3+\sqrt{3})$, see [55, p. 144]. We give an expression for the derivatives of the stability function.

Lemma 14 *The kth derivative $R^{(k)}(z)$ of (4.20) for $k \geq 1$ is given by*

$$R^{(k)}(z) = \frac{\gamma^{k-2}\left[\frac{k!}{2}((1-2\gamma)k + 4\gamma - 1) + (1-3\gamma)k!\gamma z\right]}{(1-\gamma z)^{k+2}}. \tag{4.21}$$

Proof Substituting $k = 1$ in (4.21) or computing the derivative of (4.20) gives in both cases $R^{(1)}(z) = (1 + (1-3\gamma)z)(1-\gamma z)^{-3}$. For $k > 1$ let us consider the expression

$$R^{(k)}(z) = \frac{\gamma^{k-2}\left[a_k + b_k\gamma z\right]}{(1-\gamma z)^{k+2}}, \quad a_2 = 1,\ b_2 = 2 - 6\gamma. \tag{4.22}$$

This proves the lemma for $k = 2$ by simple calculation. The derivative of $R^{(k)}(z)$ is

$$R^{(k+1)}(z) = \frac{\gamma^{k-1}\left[(k+2)a_k + b_k + (k+1)b_k\gamma z\right]}{(1-\gamma z)^{k+3}},$$

and hence a_k, b_k satisfy the system of difference equations:

$$a_{k+1} = (k+2)a_k + b_k,\ b_{k+1} = (k+1)b \quad \text{with initial data} \quad a_2 = 1,\ b_2 = 2 - 6\gamma.$$

This system is decoupled and has the solution $b_k = (1-3\gamma)k!$ and

$$a_k = \frac{1}{6}(k+1)! + \sum_{i=2}^{k-1} \frac{b_i(k+1)!}{(i+2)!} = \frac{k!}{2}\left((1-2\gamma)k + 4\gamma - 1\right).$$

Substituting this in the expression (4.22) completes the proof. □
If $\gamma > \frac{1}{2}$ then the stability function (4.20) is not absolutely monotonic in $z = 0$ (and hence the threshold factor $T(R)$ is not defined for $\gamma > \frac{1}{2}$). This can be seen as follows. It is necessary that $R^{(k)}(0) \geq 0$ for all $k \geq 1$ for absolute monotonicity of R in $z = 0$. Considering (4.21), this is

satisfied if and only if $(1-2\gamma)k + 4\gamma - 1 \geq 0$ for all $k \geq 1$. For $k = 1, 2$ this holds for all $\gamma > 0$ and for $k > 2$ we obtain that

$$\gamma \leq \frac{1-k}{2(2-k)}$$

must hold. The term on the right-hand side is always greater than $\frac{1}{2}$ for $k > 2$ and tends to $\frac{1}{2}$ for $k \to \infty$. Therefore, absolute monotonicity of R in $z = 0$ is only given if $\gamma \leq \frac{1}{2}$ ($R(0)$ is nonnegative for all $\gamma > 0$). As a consequence of this result we consider $\gamma \in \left(0, \frac{1}{2}\right]$ in the following only.

Now we will derive for each $\gamma \in \left(0, \frac{1}{2}\right]$ the maximum value μ_γ such that the derivatives (4.21), for all $k \geq 1$, of the stability function (4.20) are nonnegative for all $z \in [-\mu_\gamma, 0]$. Following this, in Lemma 15, we give the maximum value $\mu_{\gamma,0}$ for each $\gamma \in (0, \frac{1}{2}]$ such that the stability function (4.20) itself is nonnegative for all $z \in [-\mu_{\gamma,0}, 0]$.

Consider the derivatives (4.21) for $k \geq 1$. For $\gamma \in \left[\frac{1}{3}, \frac{1}{2}\right]$ we have that $(1-3\gamma)\gamma z \geq 0$ for all $z \leq 0$. Furthermore, we have shown above that for $\gamma \leq \frac{1}{2}$ we have $(1-2\gamma)k + 4\gamma - 1 \geq 0$ for all $k \geq 1$. Hence $\mu_\gamma = \infty$ for $\gamma \in \left[\frac{1}{3}, \frac{1}{2}\right]$. Now consider $\gamma \in \left(0, \frac{1}{3}\right)$. Then we require $(1-2\gamma)k + 4\gamma - 1 + 2(1-3\gamma)\gamma z \geq 0$ for all $k \geq 1$. This is satisfied if and only if for all $k \geq 1$

$$z \geq -\frac{(1-2\gamma)k + 4\gamma - 1}{2(1-3\gamma)\gamma} =: \alpha_k \,.$$

We observe that $\alpha_k \geq \alpha_{k+1}$. Hence, the most restrictive condition is $z \geq \alpha_1$ and this yields $\mu_\gamma = (1-3\gamma)^{-1}$ for $\gamma \in \left(0, \frac{1}{3}\right)$.

Lemma 15 *The maximum value $\mu_{\gamma,0}$ such that the stability function $R(z)$ given in Eq. (4.20) is nonnegative for all $z \in [-\mu_{\gamma,0}, 0]$ is given by*

$$\mu_{\gamma,0} = \infty \;\; \textit{if} \;\; \gamma \in \left(0, \frac{1}{4}\right], \quad \textit{and} \quad \mu_{\gamma,0} = -\frac{2}{2\gamma - 1 - \sqrt{4\gamma - 1}} \;\; \textit{if} \;\; \gamma \in \left(\frac{1}{4}, \frac{1}{2}\right] .$$

Proof The denominator of $R(z)$ is always positive for $z \leq 0$ and $\gamma > 0$. Hence we have to investigate nonnegativity of the numerator of $R(z)$ which is given by

$$P(z) = 1 + (1-2\gamma)z + \left(\frac{1}{2} - 2\gamma + \gamma^2\right) z^2 \,.$$

The discriminant of $P(z)$ is $D = (4\gamma - 1)\left(\frac{1}{2} - 2\gamma + \gamma^2\right)^{-2}$.

The discriminant D reveals that $P(z)$ has complex zeros for $\gamma \in \left(0, \frac{1}{4}\right)$ and a double zero for $\gamma = \frac{1}{4}$. Now follows that $R(z) \geq 0$ for all $z \leq 0$ because $P(0) > 0$, i.e. $\mu_{\gamma,0} = \infty$ for $\gamma \in \left(0, \frac{1}{4}\right]$. If $\gamma \in \left(\frac{1}{4}, 1 - \frac{1}{2}\sqrt{2}\right)$ then $P(z)$ has two real zeros. We have $P(0) > 0$ and $P'(0) > 0$ for the γ-values considered and therefore the zeros of P are to the left of $z = 0$ and the greater of both, namely

$$z_0 = -\frac{1}{2}\frac{1-2\gamma}{\frac{1}{2} - 2\gamma + \gamma^2} + \frac{1}{2}\frac{\sqrt{4\gamma - 1}}{\left|\frac{1}{2} - 2\gamma + \gamma^2\right|} = \frac{2}{2\gamma - 1 - \sqrt{4\gamma - 1}} \,,$$

defines the value of $-\mu_{\gamma,0}$.

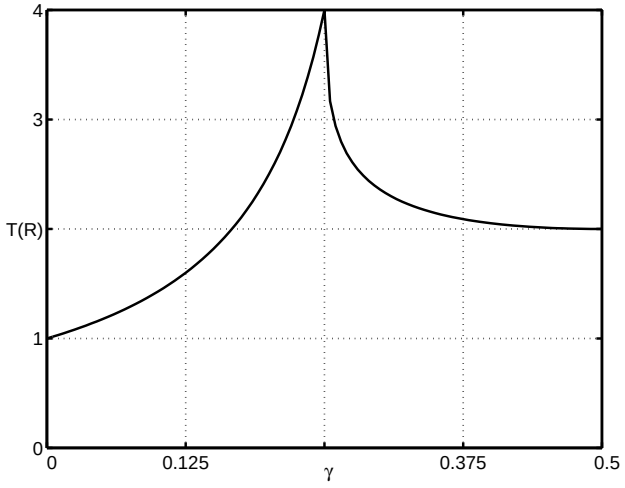

Figure 4.1: The threshold factor $T(R)$ of the stability function $R(z)$, see Eq. (4.20), of two-stage W-method of order $p \geq 2$ vs. the free parameter $\gamma \in (0, \frac{1}{2}]$ of $R(z)$. The threshold factor is not defined for $\gamma > \frac{1}{2}$.

If $\gamma = 1 - \frac{1}{2}\sqrt{2}$ then $P(z)$ reduces to $P(z) = 1 + (\sqrt{2} - 1)z$ and we obtain $\mu_{\gamma,0} = (\sqrt{2} - 1)^{-1}$. If $\gamma \in \left(1 - \frac{1}{2}\sqrt{2}, \frac{1}{2}\right]$ then $P(z)$ has two real zeros – one to the left and the other to the right of $z = 0$ because $P(0) > 0$ and $\frac{1}{2} - 2\gamma + \gamma^2 < 0$. The smaller of both, namely

$$z_0 = -\frac{1}{2}\frac{1 - 2\gamma}{\frac{1}{2} - 2\gamma + \gamma^2} - \frac{1}{2}\frac{\sqrt{4\gamma - 1}}{\left|\frac{1}{2} - 2\gamma + \gamma^2\right|} = \frac{2}{2\gamma - 1 - \sqrt{4\gamma - 1}}\,,$$

defines the value of $-\mu_{\gamma,0}$. □

Obviously we have that the threshold factor $T(R)$ of $R(z)$ for each value of $\gamma \in (0, \frac{1}{2}]$ is given by $T(R) = \min\{\mu_{\gamma,0}, \mu_\gamma\}$. The following theorem makes this precise, see also Fig. 4.1.

Theorem 8 *The stability function $R(z)$, see Eq. (4.20), of a 2-stage W-method of order $p \geq 2$ has a threshold factor $T(R) = (1 - 3\gamma)^{-1}$ for $\gamma \in (0, \frac{1}{4}]$ and a threshold factor $T(R) = -\frac{2}{2\gamma - 1 - \sqrt{4\gamma - 1}}$ for $\gamma \in (\frac{1}{4}, \frac{1}{2}]$. The threshold factor of R is not defined for $\gamma > \frac{1}{2}$.*

We see that the largest value $T(R) = 4$ is attained for $\gamma = \frac{1}{4}$. According to [19], $T(R) = 4$ is the optimal threshold factor in the class of rational approximations of order $p \geq 2$ of $\exp(z)$ for $z \to 0$ with polynomial denominator and numerator of degree two. Unfortunately, $\gamma = \frac{1}{4}$ is just the borderline for A-acceptability of $R(z)$ and therefore slightly larger values of γ should be preferred. A good choice, leading to an L-acceptable stability function $R(z)$ is $\gamma = 1 - \frac{1}{2}\sqrt{2} \approx 0.29289$. Verwer et al. [62] recommend, based on numerical experiments and some theoretical support, the choice $\gamma = 1 + \frac{1}{2}\sqrt{2} > \frac{1}{2}$ in their air pollution application. Then the stability function is also L-acceptable (however with a larger error constant) and is nonnegative for all $z \leq 0$. The latter is not true for $\gamma = 1 - \frac{1}{2}\sqrt{2}$. We return to the issue of selecting an appropriate value for γ in Sec. 4.5.

4.3.2 Positivity of RK methods on $\mathcal{D}^+(\rho)$

Absolute monotonicity of the rational functions $R(z)$ and $R_i(z)$ of an RK method is not sufficient to guarantee positivity of the method when applied to nonlinear problem classes like $\mathcal{D}^+(\rho)$. This

case is studied by Horvath [27] and we apply his theory in this section to the methods of (Class A). An important property of RK methods with respect to the problem class $\mathcal{D}^+(\rho)$ is the *radius of absolute monotonicity of an RK method* [37], which we denote by $T(A,b)$, where (A,b) is the RK method at hand. The radius $T(A,b)$ is used by Kraaijevanger [37] in the study of contractivity of RK methods and also used in the nonlinear positivity theory for RK methods by Horvath [27].
We first define the term *absolute monotonicity of an RK method* (A,b) and then its radius of absolute monotonicity $T(A,b)$, see [37].

Definition 7 *An RK method (A,b) is said to be absolutely monotonic at a given point $z \in \mathbb{R}$ if*

$$\begin{aligned} & I - zA \text{ is nonsingular}\,, \\ R(z) &= 1 + zb^T(I-zA)^{-1}\mathbb{1} \geq 0\,, \\ A(z) &= A(I-zA)^{-1} \geq 0\,, \\ b(z) &= b^T(I-zA)^{-1} \geq 0\,, \\ e(z) &= (I-zA)^{-1}\mathbb{1} \geq 0\,. \end{aligned}$$

Further, (A,b) is said to be absolutely monotonic on an interval $I \subset \mathbb{R}$ if it is absolutely monotonic for any $z \in I$.

Definition 8 *Let an RK method (A,b) be given. We define the radius of absolute monotonicity of (A,b), denoted by $T(A,b)$, by*

$$T(A,b) := \sup\left\{z | z \in \mathbb{R}_{+,0} \text{ and } (A,b) \text{ is absolutely monotonic on } [-z,0]\right\}.$$

The first condition in Definition 7 is obviously satisfied for ERK schemes. Further, we have that the threshold factor of the stability function of an RK method (A,b) is greater than or equal to the radius of absolute monotonicity of this method because the function $R(z)$ in Definition 7 is just the (rational) stability function of the RK method (A,b). Hence we have

$$T(R) \geq T(A,b).$$

One of the main results of Horvath [27, Theorem 6] is the following theorem.

Theorem 9 *Let (A,b) be an irreducible RK method and $\rho > 0$. Then we have that (A,b) is positive on $\mathcal{D}^+(\rho)$ with threshold $\tau^+ = T(A,b)/(2\rho)$ whenever $T(A,b) > 0$.*

Remark 2 *By irreducibility of RK methods we mean both, irreducibility in the DJ and in the HS sense. For concepts of reducibility we refer to [9, 10, 23]. Note that the ERK schemes of (Class A) and (Class B) are always irreducible.*

We need $T(A,b) > 0$ in Theorem 9 for positivity of the RK method (A,b) on the problem class $\mathcal{D}^+(\rho)$. This is also necessary for contractivity of the RK method applied to such problems [37]. Further, larger values of $T(A,b)$ lead to more relaxed time step restrictions for positivity (and contractivity) on these classes. We see that $T(A,b)$ is the method dependent constant and ρ is the problem dependent constant in the time step restriction for positivity of RK methods on the class $\mathcal{D}^+(\rho)$.
The following lemma with statements from [37] characterizes all irreducible RK schemes with $T(A,b) > 0$ (part 1.) and simplifies the computation (of lower bounds) of $T(A,b)$ (part 2.).

Lemma 16 ([37]) *For irreducible RK methods (A, b) holds:*

1. $T(A,b) > 0 \quad \Leftrightarrow \quad A \geq 0\,,\ b > 0\,,$ *and for all* $i, j\ (\,(A^2)_{ij} \neq 0 \Rightarrow A_{ij} \neq 0\,)\,.$

2. *Let* $r > 0\,.$ *Then* $T(A,b) \geq r \quad \Leftrightarrow \quad (A,b)$ *is absolutely monotonic in* $-r$ *and* $A \geq 0\,.$

We turn our attention to methods from (Class A). We already know that these methods have a threshold factor $T(R) = 2$ and therefore we have $T(A,b) \leq 2$. We will show that exactly one method (A,b) of (Class A) satisfies $T(A,b) = 2$, see Theorem 10.

Lemma 17 *Let (A,b) be a scheme from (Class A). We have $T(A,b) > 0$ if and only if*

$$b_3 \in (0,1)\,, \quad b_2 \in \left(0, \frac{3}{4}\right), \quad b_2 + b_3 < 1\,, \text{ and } \gamma \in \left(\frac{1}{4} - \sqrt{\frac{1}{16} - \frac{b_2}{12}}, \frac{1}{4} + \sqrt{\frac{1}{16} - \frac{b_2}{12}}\right).$$

Proof If the conditions given in the lemma are equivalent to $\{\, A \geq 0,\, b > 0$ and for all i, j holds $(\,(A^2)_{ij} \neq 0 \Rightarrow A_{ij} \neq 0\,)\,\}$ then the statement of the lemma follows with Lemma 16, part 1.
Sufficiency: $b > 0 \Rightarrow b_1 = 1 - b_2 - b_3, b_2, b_3 > 0 \Rightarrow b_2, b_3 < 1$ and $b_2 + b_3 < 1$.
$\gamma \neq 0$ and $a_{21} \geq 0 \Rightarrow \gamma > 0 \Rightarrow a_{21} > 0$ and with $b_3 > 0$ also $a_{32} > 0$. Further $(A^2)_{ij} \neq 0$ only for $i = 3, j = 1$, that is $(A^2)_{31} = a_{21}a_{32} \neq 0$, and this implies that also $a_{31} > 0$ holds.
$a_{31} > 0 \Rightarrow \gamma^2 - \frac{1}{2}\gamma + \frac{b_2}{12} < 0 \Rightarrow b_2 < \frac{3}{4}$ and $\gamma \in \left(\frac{1}{4} - \sqrt{\frac{1}{16} - \frac{b_2}{12}}, \frac{1}{4} + \sqrt{\frac{1}{16} - \frac{b_2}{12}}\right)$.
Necessity: The conditions on the b_i imply $b > 0$. $b_2 < \frac{3}{4} \Rightarrow \gamma > 0 \Rightarrow a_{32} > 0$ and $a_{21} > 0$. Further, $a_{31} > 0$ follows with the conditions on γ. □
For 3-stage ERK methods (A,b) we obtain

$$(I - zA)^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ -za_{21} & 1 & 0 \\ -za_{31} & -za_{32} & 1 \end{pmatrix}^{-1} = \begin{pmatrix} 1 & 0 & 0 \\ za_{21} & 1 & 0 \\ za_{31} + z^2 a_{32}a_{21} & za_{32} & 1 \end{pmatrix}.$$

Let (A,b) be an ERK method from (Class A) with $T(A,b) > 0$. Then (A,b) is absolutely monotonic in z ($z \in [-2,0]$) if and only if the conditions $A(z), b(z)$, and $e(z) \geq 0$ are satisfied (see Definition 7). This is true if and only if

$$\begin{aligned} C_1(z) &:= a_{31} + za_{32}a_{21} \geq 0\,, \\ C_2(z) &:= b_1 + z(b_2a_{21} + b_3a_{31}) + z^2 b_3 a_{32}a_{21} \geq 0\,, \\ C_3(z) &:= b_2 + zb_3a_{32} \geq 0\,, \\ C_4(z) &:= 1 + za_{21} \geq 0\,, \\ C_5(z) &:= 1 + z(a_{31} + a_{32}) + z^2 a_{32}a_{21} \geq 0\,. \end{aligned} \tag{4.23}$$

Lemma 18 *Let (A,b) be a scheme from (Class A) satisfying the conditions of Lemma 17. Then (A,b) is absolutely monotonic in z ($z \in [-2,0]$) if and only if $C_i(z) \geq 0$, $i = 1(1)5$, and this is the*

case if and only if

$$\begin{aligned} z &\geq 12\gamma - 6 + \frac{b_2}{\gamma}\,, \\ 0 &\leq 1 - b_2 - b_3 + \left(\frac{1}{2} - \gamma\right) z + \frac{1}{12} z^2\,, \\ z &\geq -\frac{b_2}{\gamma}\,, \\ z &\geq -12\gamma\,, \\ 0 &\leq z^2 + \left(6 - \frac{b_2}{\gamma}\right) z + 12 b_3\,. \end{aligned}$$

Proof The statement follows by simplifying and rearranging the conditions $C_i(z) \geq 0$, $i = 1(1)5$, by using the method coefficients of (Class A). □

Theorem 10 *Let (A, b) be a scheme from (Class A). Then $T(A, b) = 2$ if and only if $b_2 = b_3 = \frac{1}{3}$ and $\gamma = \frac{1}{6}$.*

Proof Let $T(A, b) = 2$. Then the conditions of Lemma 17 are satisfied. Further, (A, b) is absolutely monotonic in $z = -2$ and therefore the conditions of Lemma 18 are satisfied for $z = -2$. The third condition implies $\gamma \leq \frac{b_2}{2}$ and the fourth condition $\gamma \geq \frac{1}{6}$. These two bounds on γ make $b_2 \geq \frac{1}{3}$ necessary. The first condition of Lemma 18 for $z = -2$ implies

$$0 \geq \gamma^2 - \frac{1}{3}\gamma + \frac{b_2}{12} =: p(\gamma)\,.$$

We have $p(0) > 0$ and the discriminant is given by $D = \frac{1}{36} - \frac{b_2}{12}$. In order to have the condition satisfied for some γ we need $D \geq 0$ and this is the case only for $b_2 \leq \frac{1}{3}$. Hence $b_2 = \frac{1}{3}$ is necessary and this immediately implies $\gamma = \frac{1}{6}$. The fifth condition of Lemma 18 for $z = -2$ now implies $b_3 \geq \frac{1}{3}$ whereas the second condition requires $b_3 \leq \frac{1}{3}$. Hence we need $b_3 = \frac{1}{3}$.
On the other hand, if (A, b) is the scheme from (Class A) with $b_2 = b_3 = \frac{1}{3}$ and $\gamma = \frac{1}{6}$ then it satisfies the conditions of Lemma 17. We already know that $T(A, b) \leq 2$. With Lemma 16 holds $T(A, b) \geq 2$ if (A, b) is absolutely monotonic in $z = -2$ and this is the case if the conditions of Lemma 18 are satisfied for $z = -2$. By inspection we see that this is the case. □

4.3.3 Positivity of ERK methods for general nonlinear problems

The positivity of ERK methods applied to general positive ODEs (4.1) is considered in [31]. The approach is based on the reformulation of ERK methods as convex combination of forward Euler steps—an idea used by Shu and Osher [52] in the derivation of RK total variation diminishing time discretizations. Let $\alpha_{ij} \geq 0$ be given for $i = 2(1)s + 1$ and $j = 1(1)i - 1$ such that $\sum_{j=1}^{i-1} \alpha_{ij} = 1$. Consider an s-stage ERK method (A, b) and denote $a_{s+1,j} := b_j$ for $j = 1(1)s$ and define

$$\beta_{ij} := a_{ij} - \sum_{l=j+1}^{i-1} \alpha_{il} a_{lj}\,.$$

Then a time step τ from y_k to y_{k+1} with the ERK method is equivalent to

$$\begin{aligned} y^{(1)} &:= y_k\,, \\ y^{(i)} &:= \sum_{j=1}^{i-1} \left(\alpha_{ij} y^{(j)} + \tau\beta_{ij} f\left(t_k + c_j\tau, y^{(j)}\right)\right), \quad i = 2(1)s+1\,, \\ y_{k+1} &:= y^{(s+1)}\,. \end{aligned} \tag{4.24}$$

The freedom in the choice of the α_{ij} is used to yield nonnegative coefficients β_{ij} (which is not always possible) and to obtain, for a given ERK scheme, the optimal result from the following lemma.

Lemma 19 (see [52, 31]) *Let (A, b) be a given ERK scheme and assume that the coefficients β_{ij} in (4.24) are nonnegative. Consider an ODE $y'(t) = F(t, y(t))$ with $F \in \mathcal{P}$. If $u + \tau F(t, u) \geq 0$ for all $u \geq 0$, all t and all step sizes $0 < \tau \leq \tau^0$ then the ERK method (A, b) is positive for the given ODE under the step size restriction*

$$\tau^+ := \min_{1\leq j<i\leq s+1} \frac{\alpha_{ij}}{\beta_{ij}}\tau^0, \qquad \textit{where } \frac{\alpha_{ij}}{\beta_{ij}} := +\infty \textit{ for } \beta_{ij} = 0\,. \tag{4.25}$$

Proof We show that $\widetilde{u} := \alpha_{ij}u + \tau\beta_{ij}f(t_k + c_j\tau, u) \geq 0$ for all step sizes $0 < \tau \leq \tau^+$. If $\beta_{ij} = 0$ then this is obviously true, so assume henceforth $\beta_{ij} > 0$. If $\alpha_{ij} = 0$ then $\tau^+ = 0$ and there exists no $0 < \tau \leq \tau^+$. If $\alpha_{ij} > 0$ then

$$\widetilde{u} = \alpha_{ij}\left(u + \tau\frac{\beta_{ij}}{\alpha_{ij}} f(t_n + c_j\tau, u)\right) \geq 0$$

if $\tau\frac{\beta_{ij}}{\alpha_{ij}} \leq \tau^0$, i.e. $\tau \leq \frac{\alpha_{ij}}{\beta_{ij}}\tau^0$. □

We refer to $\min\limits_{1\leq j<i\leq s+1} \frac{\alpha_{ij}}{\beta_{ij}}$ as the *positivity factor* of a given ERK method (A, b) in this work. The positivity factor constitutes the method dependent constant in the step size bound for positivity in this approach. The problem dependent constant is the inverse of τ^0.
In Sec. 4.4.1 we demonstrate that the taxis ODE defined in Sec. 3.3.1 allows $\tau^0 > 0$ and hence fits into this framework.

Remark 3 *The positivity theory for RK methods presented in Sec. 4.3.2 is applicable to certain classes of positive (nonlinear) ODEs. On the other hand, Lemma 19 gives a statement about the positivity of ERK methods applied to general positive ODEs. It requires knowledge of the positivity of the forward Euler method applied to the given ODE.*

Now we investigate the positivity factor of ERK methods (A, b) from (Class A). We restrict attention to such methods which satisfy $T(A, b) > 0$ (see Lemma 17). Our aim is to construct methods with factor 2. The following theorem shows that exactly one method with this property exists and that this method is the same method as derived in Theorem 10.

Theorem 11 *Let (A, b) be a scheme from (Class A) with $T(A, b) > 0$ which can be rewritten in the form (4.24) with coefficients $\beta_{ij} \geq 0$. This method has a positivity factor ≥ 2 if and only if $b_2 = b_3 = \frac{1}{3}$, and $\gamma = \frac{1}{6}$. The largest possible positivity factor of this method is 2 and is attained for the choice $\alpha_{21} = 1, \alpha_{31} = 0, \alpha_{32} = 1, \alpha_{41} = \frac{1}{3}, \alpha_{42} = 0$, and $\alpha_{43} = \frac{2}{3}$.*

Proof A method (A, b) considered in the theorem has a positivity factor ≥ 2 if and only if we can select α_{ij} such that $\frac{\alpha_{ij}}{\beta_{ij}} \geq 2$ for all $1 \leq j < i \leq 4$. We show (following this paragraph) that this implies $b_2 = b_3 = \frac{1}{3}$, and $\gamma = \frac{1}{6}$, and further that this results in the unique choice of the values of α_{ij} as given in the theorem. On the other hand, simple calculation shows that this specific method has a positivity factor two for the choice of α_{ij} given in the theorem.
We have $\alpha_{21} = 1$ and $\beta_{21} = \frac{1}{12\gamma} > 0$. Hence $\frac{\alpha_{21}}{\beta_{21}} \geq 2 \Leftrightarrow \gamma \geq \frac{1}{6}$. Further, $\beta_{32} = a_{32} = \frac{\gamma}{b_3} > 0$ and hence $\frac{\alpha_{32}}{\beta_{32}} \geq 2 \Leftrightarrow \alpha_{32} \geq \frac{2\gamma}{b_3}$. Using $\alpha_{32} \leq 1$ this implies $b_3 \geq 2\gamma \geq \frac{1}{3}$. $\beta_{43} = b_3 > 0$ implies $\frac{\alpha_{43}}{\beta_{43}} \geq 2 \Leftrightarrow \alpha_{43} \geq 2b_3$. With $\alpha_{43} \leq 1$ follows $b_3 \leq \frac{1}{2}$ and hence $\gamma \leq \frac{1}{4}$.
We have $\beta_{42} = b_2 - \alpha_{43}\frac{\gamma}{b_3} \geq 0 \Leftrightarrow \alpha_{43} \leq \frac{b_2 b_3}{\gamma}$. This implies $2\gamma \leq b_2$ and hence $b_2 \geq \frac{1}{3}$ and $b_1 \leq \frac{1}{3}$. Next, $\beta_{31} = a_{31} - a_{21}\alpha_{32} \geq 0 \Leftrightarrow \alpha_{32} \leq \frac{a_{31}}{a_{21}} = \frac{1}{b_3}(6\gamma - 12\gamma^2 - b_2)$. Using $b_2 \geq 2\gamma$ implies $\alpha_{32} \leq \frac{2\gamma}{b_3}(2 - 6\gamma)$ and $\gamma \geq \frac{1}{6}$ leads to $\alpha_{32} \leq \frac{2\gamma}{b_3}$. Hence we need $\alpha_{32} = \frac{2\gamma}{b_3}$. This simplifies the expression of β_{31} yielding $\beta_{31} = \frac{1}{b_3}\left(\frac{1}{3} - \frac{b_2}{12\gamma} - \gamma\right)$. Now consider two cases.
Case 1 ($\beta_{31} = 0$): This is only possible for $\gamma = \frac{1}{6}$ and $b_2 = \frac{1}{3}$ because we already know $b_2 \geq \frac{1}{3}$. Now we have $\alpha_{43} \leq \frac{b_2 b_3}{\gamma} = 2b_3$ and therefore $\alpha_{43} = 2b_3$. This also gives $\beta_{42} = 0$. Finally, $\beta_{41} = b_1 - a_{21}\alpha_{42} - a_{31}\alpha_{43} = \frac{1}{3} - b_3 - \frac{\alpha_{42}}{2} \geq 0 \Leftrightarrow b_3 \leq \frac{1}{3} - \frac{\alpha_{42}}{2}$. Using $b_3 \geq \frac{1}{3}$ leads to $b_3 = \frac{1}{3}$ and $\alpha_{42} = 0$. Hence we obtain the method given in the theorem. Also, all parameters α_{ij} have fixed values now (as given in the theorem).
Case 2 ($\beta_{31} > 0$): This is only possible if $b_2 < \frac{1}{3}$ and this is in contradiction with $b_2 \geq \frac{1}{3}$.
This completes the proof. □

4.3.4 Further results on positivity of ERK methods and the method RK32

The investigations in Sec. 4.3.2 and Sec. 4.3.3 have singled out a unique method from (Class A) (the method with $b_2 = b_3 = \frac{1}{3}$, and $\gamma = \frac{1}{6}$) with favourable nonlinear positivity properties. We refer to this method as RK32 (because it has three stages and order two) in the following. The Butcher array of this method is given in Fig. 4.2. RK32 is optimal with respect to positivity on the class $\mathcal{L}_0^+(\alpha)$ by construction. We have not yet looked at the positivity of RK32 on $\mathcal{L}_g^+(\alpha)$ for $g \neq 0$. This is our next task.
We apply a method (A, b) from (Class A) to a problem from class $\mathcal{L}_g^+(\alpha)$. The polynomials $R_i(z)$, $i = 1, 2, 3$ in Eq. (4.19) of a 3-stage ERK method are given by

$$R_1(z) = b_1 + (b_2 a_{21} + b_3 a_{31})z + b_3 a_{32} a_{21} z^2\,, \quad R_2(z) = b_2 + b_3 a_{32} z\,, \quad R_3(z) = b_3\,,$$

and simplify for methods from (Class A) to

$$R_1(z) = 1 - b_2 - b_3 + \left(\frac{1}{2} - \gamma\right) z + \frac{1}{12} z^2\,, \quad R_2(z) = b_2 + \gamma z\,, \quad R_3(z) = b_3\,.$$

For optimal positivity of the methods applied to problems from $\mathcal{L}_g^+(\alpha)$ we need, according to Theorem 6, that the threshold factors $T(R_i)$ of the polynomials R_i and the threshold factor $T(R)$ of the stability polynomial are as large as possible. It is $T(R) = 2$ by construction of (Class A) and therefore we are interested in methods with $T(R_i) \geq 2$. The method RK32 is easily shown to satisfy this requirement and is hence optimal with respect to positivity on the class $\mathcal{L}_g^+(\alpha)$.

Beside the methods of (Class A), we have also defined the methods of (Class B) which satisfy additionally one of the third-order conditions. We identified the admissible range of the parameters (b_2, γ) such that $T(A, b) > 0$ holds for methods (A, b) of (Class B). Within this parameter range we determined numerically the method which yields the largest radius of absolute monotonicity $T(A, b)$ and hence is optimal with respect to positivity on the problem class $\mathcal{D}^+(\rho)$, see Theorem 9. This is the method with $b_2 = 0.3572, \gamma = 0.3039$ leading to $T(A, b) = 1.1754$.
Rewriting the ERK method of (Class B) with $b_2 = 0.3572$ and $\gamma = 0.3039$ as a convex combination of forward Euler steps (see formula (4.24)) with $\alpha_{3,1} = 0.3213, \alpha_{4,1} = 0.38$ and $\alpha_{4,2} = 0.0000764$ results in a positivity factor of ≈ 1.1754 (see Lemma 19), and this factor is optimal (numerical search).
These numerically obtained values (optimal $T(A, b)$, positivity factor) for the methods of (Class B) are slightly better than those which hold for some s-stage methods of order s (all values equal one) but they are worse than the values for the optimal method RK32 of (Class A) where all values equal two. Further, the methods of (Class B) are still of second-order only and the advantage of having one of the third-order conditions satisfied is expected to be marginal compared with methods of (Class A). Therefore we will omit the results of this method in our numerical tests.
In the literature we find other ERK schemes which are recommended for the solution of positive ODEs. In [17] we compare RK32 with the following schemes: modified Euler (ME, two stages, second-order) and Runge-Kutta-Fehlberg method 2(3) (RKF2(3), three stages, third-order, see [22, 52, 31]). Both methods have $T(R) = T(A, b) = 1$, positivity factor 1, and they are recommended in [52]. Fig. 4.2 gives the Butcher arrays of these methods together with embedded methods which have one order less than the primary methods. These embedded methods can be used to estimate the local error in the computation and to adaptively change the time step size. The time step selection strategy by embedding is based on accuracy only (positivity is not taken into account) and we advance a time step always with the higher order solution (local extrapolation). Therefore the positivity properties of the embedded methods are not essential.

$$\begin{array}{c|cc} 0 & 0 & \\ 1 & 1 & 0 \\ \hline & \frac{1}{2} & \frac{1}{2} \\ \hline & 1 & 0 \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & & \\ 1 & 1 & 0 & \\ \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & 0 \\ \hline & \frac{1}{6} & \frac{1}{6} & \frac{2}{3} \\ \hline & \frac{1}{2} & \frac{1}{2} & 0 \end{array} \qquad \begin{array}{c|ccc} 0 & 0 & & \\ \frac{1}{2} & \frac{1}{2} & 0 & \\ 1 & \frac{1}{2} & \frac{1}{2} & 0 \\ \hline & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ \hline & \frac{1}{2} & \frac{1}{2} & 0 \end{array}$$

Figure 4.2: Butcher arrays for ME, RKF2(3), and RK32 (from left to right). The last row of each array defines an embedded method.

In [17] we present numerical experiments with these three ERK methods. Test examples are semi-discretizations (with flux limiter) of a linear advection equation and of a TDR system (here the ERK methods are used in an OPS scheme). The experiments demonstrated that RK32 allows for the largest time steps in order to obtain nonnegative solutions of comparable accuracy. From the point of efficiency, ME and RK32 are comparable (ME requires only two function evaluations per time step), and RKF2(3) is more expensive. The larger time steps allowed by RK32 pay off in splitting schemes especially for lower accuracy requirements. This advantage should be more

pronounced if the implicit part in the splitting scheme becomes more expensive. The higher order of RKF2(3) pays off for higher accuracy demands. We will employ ME, RK32, and RKF2(3) in our numerical methods to be described in Sec. 4.5 and Sec. 4.6.

4.4 Positivity and stability of ERK methods for the taxis ODE

In this section we consider the positive taxis ODE derived in Sec. 3.3.1,

$$\frac{\mathrm{d}}{\mathrm{d}\,t}U_{\mathbf{i}}(t) = \mathcal{H}_T(\mathbf{U}(t);\mathbf{i})\,, \quad \mathbf{i}\in\mathcal{I}\,. \tag{4.26}$$

At first we demonstrate in Sec. 4.4.1 that Eq. (4.26) fits into the framework of Sec. 4.3.3 by deriving an expression for the threshold step size τ^0 required in Lemma 19. Then, in Sec. 4.4.2, we discuss linear stability properties of some ERK schemes (including RK32) when applied to a linearized version of Eq. (4.26).

4.4.1 Positivity of the forward Euler method for the taxis ODE

We denote with F_j the taxis discretization in spatial direction j,

$$F_j(\mathbf{U}(t),t) := -\frac{1}{h}\left(\mathcal{T}_j(\mathbf{U};\mathbf{i}) - \mathcal{T}_j(\mathbf{U};\mathbf{i}-\mathbf{e}_j)\right)\,,$$

which can be written in the form $F_j(\mathbf{U}(t),t) = \beta_1 U_{\mathbf{i}-\mathbf{e}_j} - \beta_0 U_{\mathbf{i}} + \beta_2 U_{\mathbf{i}+\mathbf{e}_j}$, with values β_0,β_1,β_2 depending on the signs of the local velocities (and also on components of $\mathbf{U}$ through the smoothness monitor r), see below.
If $v_{\mathbf{i},j}, v_{\mathbf{i}-\mathbf{e}_j,j} \geq 0$ then we obtain, in the non-exceptional case of the state interpolants (3.12a), the values

$$\beta_0 = \frac{1}{h}\left(v_{\mathbf{i},j} + \frac{v_{\mathbf{i},j}}{2}\Phi(r_{\mathbf{i},j}) - \frac{\Phi(r_{\mathbf{i}-\mathbf{e}_j,j})}{2r_{\mathbf{i}-\mathbf{e}_j,j}}v_{\mathbf{i}-\mathbf{e}_j,j}\right) \leq \frac{v_{\mathbf{i},j}}{h}\left(1+\frac{\delta}{2}\right),$$

$$\beta_1 = \frac{1}{h}\left(v_{\mathbf{i}-\mathbf{e}_j,j} + \frac{v_{\mathbf{i},j}}{2}\Phi(r_{\mathbf{i},j}) - \frac{\Phi(r_{\mathbf{i}-\mathbf{e}_j,j})}{2r_{\mathbf{i}-\mathbf{e}_j,j}}v_{\mathbf{i}-\mathbf{e}_j,j}\right) \geq 0\,, \quad \text{and} \quad \beta_2 = 0\,.$$

The bounds on β_0 and β_1 follow from the assumptions (3.14b) on the limiter function Φ (also the value of δ). The same bounds on β_0 and β_1 (and also $\beta_2 = 0$) are obtained if exceptional cases occur in the definition of the state interpolants. We assume $\mathbf{U} \geq 0$. One step with the forward Euler method yields the approximation $\widetilde{\mathbf{U}}$ at the new time level

$$\widetilde{U}_{\mathbf{i}} = U_{\mathbf{i}} + \tau(\beta_1 U_{\mathbf{i}-\mathbf{e}_j} - \beta_0 U_{\mathbf{i}} + \beta_2 U_{\mathbf{i}+\mathbf{e}_j}) = (1-\tau\beta_0)U_{\mathbf{i}} + \tau\beta_1 U_{\mathbf{i}-\mathbf{e}_j} + \tau\beta_2 U_{\mathbf{i}+\mathbf{e}_j}\,.$$

Hence $\widetilde{U}_{\mathbf{i}} \geq 0$ provided $\tau\beta_0 \leq 1$ and $\tau\beta_1, \tau\beta_2 \geq 0$, and this is the case if $\tau \leq h(v_{\mathbf{i},j}(1+\delta/2))^{-1}$.
Now consider the case that $v_{\mathbf{i},j}, v_{\mathbf{i}-\mathbf{e}_j,j} \leq 0$. This leads, in the non-exceptional case of the state interpolants (3.12b), to values

$$\beta_0 = \frac{1}{h}\left(-v_{\mathbf{i}-\mathbf{e}_j,j} + \frac{-v_{\mathbf{i}-\mathbf{e}_j,j}}{2}\Phi(r_{\mathbf{i},j}^{-1}) - \frac{\Phi(r_{\mathbf{i}-\mathbf{e}_j,j}^{-1})}{2r_{\mathbf{i}-\mathbf{e}_j,j}^{-1}}(-v_{\mathbf{i},j})\right) \leq \frac{-v_{\mathbf{i}-\mathbf{e}_j,j}}{h}\left(1+\frac{\delta}{2}\right),$$

$$\beta_1 = 0\,, \quad \text{and} \quad \beta_2 = \frac{1}{h}\left(-v_{\mathbf{i},j} + \frac{-v_{\mathbf{i}-\mathbf{e}_j,j}}{2}\Phi(r_{\mathbf{i},j}^{-1}) - \frac{\Phi(r_{\mathbf{i}-\mathbf{e}_j,j}^{-1})}{2r_{\mathbf{i}-\mathbf{e}_j,j}^{-1}}(-v_{\mathbf{i},j})\right) \geq 0\,.$$

The bounds on β_0, β_2 follow from the assumptions (3.14b) on the limiter function. (The same bounds hold in the exceptional cases of (3.12b).) Hence, similar as in the previous case, an Euler step yields $\widetilde{U}_{\mathbf{i}} \geq 0$ if $\tau \leq h(-v_{\mathbf{i}-\mathbf{e}_j,j}(1+\delta/2))^{-1}$.
Along the same lines we can treat the cases where $(v_{\mathbf{i},j} \leq 0, v_{\mathbf{i}-\mathbf{e}_j,j} \geq 0)$ or $(v_{\mathbf{i},j} \geq 0, v_{\mathbf{i}-\mathbf{e}_j,j} \leq 0)$. Then we obtain the bounds $(\beta_0 \leq 0, \beta_1 \geq 0, \beta_2 \geq 0)$ and $(\beta_0 \leq 1/h \cdot (v_{\mathbf{i},j} - v_{\mathbf{i}-\mathbf{e}_j,j})(1+\delta/2), \beta_1 \geq 0, \beta_2 \geq 0)$, respectively. Altogether we obtain that the Euler step yields a nonnegative result $\widetilde{U}_{\mathbf{i}}$ for all $\mathbf{i} \in \mathcal{I}$ if

$$\tau \leq \frac{h}{v^{(j)}(1+\delta/2)}\,, \quad \text{where} \quad v^{(j)} := 2\max_{\mathbf{i}\in\mathcal{I}} |v_{\mathbf{i},j}|\,.$$

Note that the factor 2 in the expression for $v^{(j)}$ can be replaced by 1 if the local velocities $v_{\mathbf{i},j}$ in spatial direction j have a uniform sign.
Summation of F_j over all spatial directions j now yields a step size bound τ^0 such that the forward Euler method applied to the taxis ODE (4.26) is positive for all $0 < \tau \leq \tau^0$. We obtain

$$\tau^0 = \frac{h}{v(1+\delta/2)}\,, \quad \text{where} \quad v := \sum_{j=1}^{d} v^{(j)}\,.$$

This generalizes the result given by Hundsdorfer et al. [31] to the case of non-constant velocities.

4.4.2 Discussion of linear stability

We start with considering the stability of some ERK methods applied to Dahlquist's (linear) test equation $y' = \lambda y$, $\lambda \in \mathbb{C}$, i.e. we study the stability polynomial of these ERK methods. Methods taken from (Class A) and (Class B) have the same stability polynomial, namely $R_{3,2}^+(z)$ (see Sec. 4.3.1.1). The stability domain of $R_{3,2}^+(z)$ is given in the plots of Fig. 4.3. For comparison, we also print the linear stability domains of the s-stage ERK methods of order s for $s = 2, 3$ in these plots. The domain of the 3-stage methods of (Class A) and (Class B) is stretched by a factor of about two in the real direction compared with the domains of the s-stage methods of order s. With respect to the imaginary direction, there is only little stretching compared with the 2-stage, second-order methods and a slight disadvantage near the imaginary axis compared with the 3-stage, third-order methods. Altogether, the 3-stage, second-order methods appear to have favourable linear stability properties.
Later we will apply the derived ERK method RK32 or other ERK methods to the (nonlinear) taxis ODE (4.26). In order to get some information about the behaviour of the ERK methods applied to (4.26), we will now consider a simple test problem, the scalar advection equation in one space dimension,

$$\partial_t u + v\partial_x u = 0\,, \quad v \in \mathbb{R}\,,\ t \geq 0\,,\ u(0,x) = u_0(x)\,. \tag{4.27}$$

This is a linear, constant coefficient problem and we assume periodic boundary conditions. Hence we can apply Fourier analysis as in [61, p. 17], see also e.g. [20].
We use the state interpolation approach, see Sec. 3.3.1, to discretize the spatial derivatives on a grid with grid width h. The limiter function makes the resulting MOL-ODE nonlinear and this

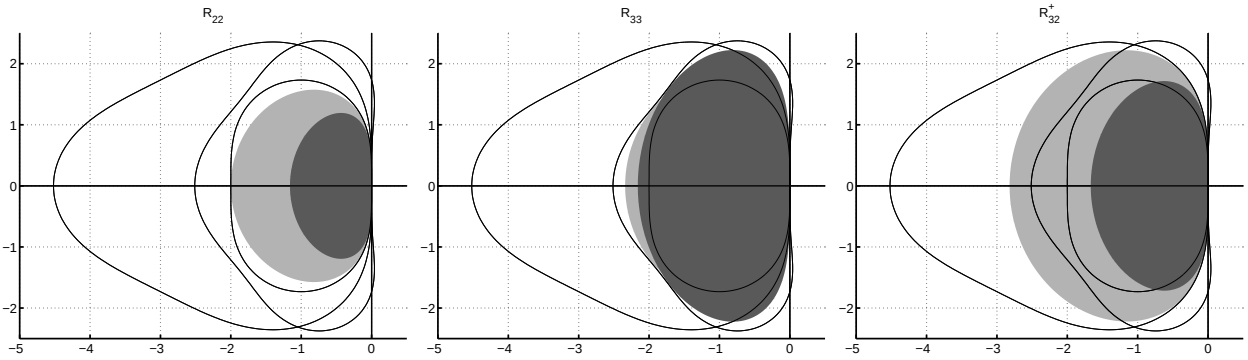

Figure 4.3: The lines in each of the plots are the boundaries of the stability domains of methods with stability polynomials $R^+_{3,2}$, $R_{3,3}$, and $R_{2,2}$ (from largest to smallest; the stability domain is the respective interior area). The dark gray shaded area in each plot is the stability domain of methods with stability polynomial as given in the plot title if the eigenvalues are confined to $\lambda_{1/3}(\xi)$, i.e. the limiter function Φ^L_K (the light gray shaded area corresponds to eigenvalues $\lambda_0(\xi)$, i.e. the limiter function Φ^L_{VL}).

prevents the application of Fourier analysis. However, if we assume spatially smooth profiles u then the smoothness monitor r, Eq. (3.13), is approximately equal to one. Therefore we linearize the limiter function Φ around $r = 1$ and this in turn leads to a linear MOL-ODE. The linearized version of the Koren and van Leer limiter are

$$\Phi^L_K(r) = K_{1/3}(r)\,, \quad \text{and} \quad \Phi^L_{VL}(r) = K_0(r)\,, \quad \text{where} \quad K_\kappa(r) = \frac{1-\kappa}{2} + \frac{1+\kappa}{2} r\,.$$

Remark 4 *The semi-discretization of problem (4.27) with the state interpolation approach using the linearized Koren limiter function $\Phi^L_K(r)$ is equivalent to the semi-discretization with the third-order upwind biased discretization ($\kappa = \frac{1}{3}$-method [31]) of the spatial derivative in (4.27).*

We obtain eigenvalues $\lambda_\kappa(\xi)$ of the linearized discretization operators (along the lines described in [61, p. 17]),

$$\lambda_\kappa(\xi) = -\frac{|v|}{h}\frac{1-\kappa}{2}\left((\cos\xi - 1)^2 + i\,\mathrm{sign}(v)\sin\xi\left(\frac{3-\kappa}{1-\kappa} - \cos\xi\right)\right), \quad \xi \in [0, 2\pi]\,.$$

We see that the real part of $\lambda_\kappa(\xi)$ for $\kappa = 0$ is slightly more negative than the real part of $\lambda_\kappa(\xi)$ for $\kappa = \frac{1}{3}$. This implies that the discretization with (linearized) van Leer limiter has slightly more damping than the discretization with the (linearized) Koren limiter.
When applying an ERK method with stability polynomial $R(z)$ and step size τ to the linear MOL-ODEs obtained with the linearized limiters Φ^L_K or Φ^L_{VL} then for stability we require that

$$|R(\tau\lambda_\kappa(\xi))| \le 1 \quad \text{for all} \quad \xi \in [0, 2\pi]\,.$$

We define $\nu := \frac{\tau|v|}{h}$ and want to maximize ν under the restriction that the scheme is stable. Then we obtain (numerically) the values in Tab. 4.1 (first two lines) and the stability domains as illustrated in the plots of Fig. 4.3. We see that methods with stability polynomial $R_{3,3}$ allow for the largest values of ν and hence largest time steps τ with respect to stability for the MOL-ODE obtained with limiter Φ^L_K, followed by methods with $R^+_{3,2}$ and then methods with $R_{2,2}$. If the limiter is Φ^L_{VL} then we obtain the order $R^+_{3,2}$, $R_{3,3}$, and $R_{2,2}$.

maximum values of ν for	RK32 ($R^+_{3,2}$)	ME ($R_{2,2}$)	RKF2(3) ($R_{3,3}$)
stability with limiter Φ^L_K	1.25	0.87	1.62
stability with limiter Φ^L_{VL}	1.41	1	1.17
positivity (nonlinear MOL-ODE)	1	0.5	0.5

Table 4.1: Maximum values of ν for stability and positivity.

With the help of Sec. 4.4.1 and Lemma 19 we can derive a value of ν such that the schemes RK32, ME, and RKF2(3) are positive when applied to the nonlinear MOL-ODE obtained by the state interpolation approach with limiter function Φ_K or Φ_{VL}. These values are given in the third line of Tab. 4.1. They are the same for ME and RKF2(3) because both have a positivity factor equal to one, whereas RK32 has a twice as large value caused by the positivity factor equal to two for this method.
The values of ν in Tab. 4.1 are best balanced for the RK32 method.

4.5 Rosenbrock-type methods with AMF

4.5.1 Two-stage methods ROS2(γ)-AMF and ROS3-AMF

A family of 2-stage, second-order W-methods (4.6) is given by the coefficients [10],

$$A = \begin{pmatrix} 0 & 0 \\ \dfrac{1}{2b_2} & 0 \end{pmatrix}, \quad b = \begin{pmatrix} 1-b_2 \\ b_2 \end{pmatrix}, \quad \text{and} \quad \Gamma = \begin{pmatrix} \gamma & 0 \\ -\dfrac{\gamma}{b_2} & \gamma \end{pmatrix}, \tag{4.28}$$

where $b_2 \neq 0$ and $\gamma > 0$ are still free parameters. The order of the method is independent of the matrix T (W-method property).
Verwer et al. [62] successfully applied the method with parameters (4.28) where $b_2 = \frac{1}{2}$ to advection–diffusion–reaction problems from atmospheric air pollution modelling. The parameter $\gamma > 0$ is still free in this scheme which we refer to as ROS2(γ). The underlying ERK method of ROS2(γ) is the modified Euler (ME) method and a first-order embedded method is provided by the choice $\widetilde{b} = \begin{pmatrix} 1 & 0 \end{pmatrix}^{\mathsf{T}}$.
The stability function $R(z)$ for all schemes (4.6) with parameters (4.28) is given by Eq. (4.20). This function is A-acceptable for $\gamma \geq \frac{1}{4}$ and L-acceptable for $\gamma_\pm = 1 \pm \frac{1}{2}\sqrt{2}$. Theorem 8 gives the radius of absolute monotonicity of the stability function $R(z)$. This radius $T(R)$ is important for positivity of the scheme when applied to linear problems with right-hand side functions from the class $\mathcal{L}^+_0(\alpha)$. The theorem states that the radius $T(R)$ is not defined for $\gamma > \frac{1}{2}$ (i.e. the method is not positive on $\mathcal{L}^+_0(\alpha)$), and the radius is largest for $\gamma = \frac{1}{4}$, i.e. exactly for the limit value of γ which yields A-stability.
We now turn our attention to the ROS2(γ) method applied, in the form (4.7), with AMF as defined by the factorization (4.13). We refer to these methods as ROS2(γ)-AMF. The order of the method is not affected by AMF but we now consider a more realistic scalar stability test equation than

$y' = \lambda y$. This test equation has the form

$$\frac{\mathrm{d}\, y(t)}{\mathrm{d}\, t} = \left(\lambda_T + \lambda_0 + \sum_{j=1}^{d} \lambda_j \right) y(t)\,, \tag{4.29}$$

where we assume that λ_T corresponds to an eigenvalue of the taxis discretization (see Sec. 4.4.2), λ_0 to a reaction eigenvalue, and $\lambda_j, j = 1(1)d$ to the eigenvalues of the diffusion discretization in the d spatial directions. The matrix $I - \tau\gamma T$ in the method is now defined by the factorization (4.13)

$$(I - \tau\gamma T) = (1 - \tau\gamma\lambda_0) \prod_{j=1}^{d} (1 - \tau\gamma\lambda_j) =: p\,. \tag{4.30}$$

The factorized stability function of a method (4.7, 4.28) with respect to Eq. (4.29) is

$$R(z_T, z_0, z_1, \ldots, z_d) = 1 + 2zp^{-1} + \left(\frac{1}{2} z^2 - z \right) p^{-2}\,, \tag{4.31}$$

where $z_j := \tau\lambda_j$ for $j = 0(1)d$, $z_T := \tau\lambda_T$, and $z := z_T + \sum_{j=0}^{d} z_j$. The strong impact of AMF on the stability of a Rosenbrock-type method (4.7) is nicely illustrated by considering the damping of the scheme at infinity. Without AMF there exist methods (4.7) with a stability function $R(z)$ which satisfies $|R(\infty)| = 0$. On the other hand, the stability function $R(z_T, z_0, z_1, \ldots, z_d)$ of a Rosenbrock-type method applied with AMF yields $|R(z_T, \infty, \ldots, \infty)| = 1$ for $d \geq 1$ (even if $z_T = 0$), i.e. there is no damping at all at infinity. Further, $R(z_T, 0, 0 \ldots, 0)$ is the stability polynomial of the underlying ERK scheme and therefore it is sensible to assume that z_T is within the stability domain of this ERK method.

Let us assume that z_T is within the stability domain of the ME method (the underlying ERK scheme of ROS2(γ)), and that $z_j \leq 0$, $j = 1(1)d$, because the λ_j, $j = 1(1)d$, correspond to the diffusion discretization. Then, for a given γ-value, we are looking for the maximum value θ such that ROS2(γ)-AMF is stable, i.e. $|R(z_T, z_0, \ldots, z_d)| \leq 1$, for all z_0 in the closed wedge W_θ, see p. 39. This case is considered by Hundsdorfer [32] for the values $d = 1$ and $d = 2$ and he states that the largest values of θ are obtained for values of $\gamma \in [0.5, 0.8]$—the best choice being $\gamma \approx 0.59$ leading to $\theta \approx 77°$. Further, for $\gamma \in [\frac{1}{4}, \frac{1}{2}]$ no angle $\theta > 0°$ is obtained, and the choice $\gamma = \gamma_+$, results in $\theta \approx 11°$.

With these results in mind we will consider the following ROS2(γ)-AMF methods in our numerical experiments:

- ROS2(γ_-)-AMF: ROS2(γ_-) is L-stable and is positive on $\mathcal{L}_0(\alpha)$ under a step size restriction. ROS2(γ_-)-AMF has a value $\theta = 0°$.

- ROS2(γ_+)-AMF: ROS2(γ_+) is L-stable and is not positive on $\mathcal{L}_0(\alpha)$. ROS2(γ_+)-AMF has a value $\theta = 11°$. This γ-value is also used in [62].

- ROS2(0.59)-AMF: ROS2(0.59) is A-stable and is not positive on $\mathcal{L}_0(\alpha)$. ROS2(0.59)-AMF has a value $\theta \approx 77°$.

Another Rosenbrock-type method (4.6) with parameters (4.28) where $b_2 = \frac{3}{4}$ and $\gamma = \gamma_3 := \frac{1}{2}+\frac{1}{6}\sqrt{3}$ is discussed in [38]. This scheme is also second-order if applied as a W-method. However, the method is third-order if applied to autonomous ODEs ($F(t,y) \equiv F(y)$) and with a matrix $T = \frac{\partial F(y)}{\partial y} + \mathcal{O}(\tau)$, i.e. a first-order approximation of the exact Jacobian of F. We refer to this scheme as ROS3. The AMF (4.13) does not lead to such a matrix T if $F_0 \neq 0$ and therefore ROS3 applied with AMF, i.e. ROS3-AMF, to the full MOL-ODE is second-order. With $\widetilde{b} = \begin{pmatrix} 1 & 0 \end{pmatrix}^{\mathsf{T}}$ we obtain a first-order embedded method. We will also use this method in our numerical experiments and apply it to the full MOL-ODE (in the sense of a second-order W-method)

- ROS3-AMF: ROS3 is A-stable, not positive on $\mathcal{L}_0(\alpha)$, and of second-order (third-order if applied to autonomous ODEs and with a matrix T which is a $\mathcal{O}(\tau)$-approximation of the true Jacobian). ROS3-AMF method has a value $\theta \approx 56°$.

We are looking for appropriate schemes Ψ_1 to solve ODEs with right-hand side F_1 within the OPS framework, see Sec. 4.2. In this case, the factorization (4.13) leads to a $\mathcal{O}(\tau)$-approximation T of the Jacobian of F_1. Further, all models which we look at here (see Sec. 2.3) give rise to autonomous MOL-ODEs. Hence, if Ψ_1 is chosen as the ROS3-AMF method then we obtain a third-order accurate scheme. We will also consider the second-order methods ROS2(γ)-AMF with $\gamma = \gamma_-$ and $\gamma = 0.59$ as methods Ψ_1 in the OPS schemes (to be detailed in Sec. 4.6).
With the OPS methods in mind, we are now looking at the stability of ROS2(γ)-AMF and ROS3-AMF applied to the test equation (4.29) with $\lambda_T = 0$. Then the factorized stability function R of these schemes is given by Eq. (4.31) with $z_T = 0$. Lanser et al. [38, Theorem 1] prove that $|R(0, z_0, z_1)| \leq 1$ holds for all $z_0, z_1 \in \mathbb{C}_{-,0}$ if and only if $\gamma \geq \gamma_3 = \frac{1}{2} + \frac{1}{6}\sqrt{3}$. This implies that we have the A-stability property of the factorized schemes if $d = 1$. If $d = 2$ then we must restrict $z_0, z_1, z_2 \in W_{45°}$, i.e. $A(45°)$-stability [32]. This result can be improved if we assume that $z_1, \ldots, z_d \leq 0$. Then the cases $d = 1$ and $d = 2$ coincide again and Hundsdorfer [32] obtains values θ_0 (the subscript indicating that $z_T = 0$) such that $|R(0, z_0, z_1, \ldots, z_d)| \leq 1$ if $z_0 \in W_{\theta_0}$. The values are: $\theta_0 = 90°$ for $\gamma \geq \gamma_3$ (hence for the methods ROS2(γ_+)-AMF and ROS3-AMF), $\theta_0 \approx 81°$ for ROS2(0.59)-AMF, and finally $\theta_0 = 0°$ for the ROS2(γ_-)-AMF.

4.5.2 Three-stage methods ROS32(γ)-AMF

In Sec. 4.3 we derive the 3-stage, second-order ERK method RK32 with favourable positivity properties. Our aim here is the construction of a 3-stage, second-order W-method with underlying ERK scheme RK32. We hope that this method will combine the good positivity properties of the underlying ERK method with the good stability properties of linearly-implicit methods.
Construction of the method class ROS32(γ): The parameters A and b of the Rosenbrock-type method (4.6) are determined by the underlying ERK scheme RK32 already, see Fig. 4.2, and we are left to find parameters Γ. The additional condition for order two of a W-method, Eq. (4.10), now yields $\gamma_{21} = -(3\gamma + \gamma_{31} + \gamma_{32})$, and we are left with the free parameters $\gamma, \gamma_{31}, \gamma_{32}$. We can compute the stability function $R(z)$ of the methods (with respect to Dahlquist's test equation),

$$R(z) = \frac{1}{(1-\gamma z)^3}\left[1 + (1-3\gamma)z + \left(\frac{1}{2} - 3\gamma + 3\gamma^2\right)z^2 + r_3 z^3\right], \tag{4.32}$$

where

$$r_3 := \frac{1}{12} - \frac{1}{6}\gamma_{31} - \frac{1}{3}(\gamma_{31}\gamma_{32} + \gamma_{32}^2) - (1+\gamma_{32})\gamma + 2\gamma^2 - \gamma^3 \,. \tag{4.33}$$

We obtain $|R(\infty)| = 0$ if and only if $r_3 = 0$ and this is equivalent to

$$\gamma_{31} = \frac{-1}{1+2\gamma_{32}} \left(6\gamma^3 - 12\gamma^2 + 6(1+\gamma_{32})\gamma + 2\gamma_{32}^2 - \frac{1}{2} \right), \ \gamma_{32} \in \mathbb{R} \setminus \left\{ -\tfrac{1}{2} \right\}, \ \gamma > 0 \,, \tag{4.34a}$$

or

$$\gamma_{31} \in \mathbb{R} \,, \ \gamma_{32} = -\frac{1}{2} \,, \ \gamma = \gamma_\pm := 1 \pm \frac{1}{2}\sqrt{2} \,. \tag{4.34b}$$

With this necessary condition for L-stability satisfied, the stability function of our methods is now given by Eq. (4.32) with $r_3 = 0$. This function is A-acceptable if and only

$$\gamma \in [g_-, g_+] \approx [0.180, 2.186] \,, \ \text{ where } \ g_\pm = \frac{3}{4} + \frac{1}{4}\sqrt{3} \pm \frac{1}{12}\sqrt{72 + 42\sqrt{3}} \,. \tag{4.35}$$

The interval includes the values $\gamma_\pm$ given in Eq. (4.34b). Hence, if the parameters Γ satisfy $\gamma_{21} = -(3\gamma + \gamma_{31} + \gamma_{32})$ and also the conditions (4.34) and (4.35) then we obtain the class of L-stable, 3-stage, second-order W-methods with underlying ERK scheme RK32. We restrict our attention to such methods in the following.

We will use the third-stage solution of the methods as embedded solution in the time step size control. This solution has order one. With respect to Dahlquist's test equation we obtain the internal stability function $R_3(z)$ of the third-stage solution,

$$R_3(z) = \frac{1}{(1-\gamma z)^2} \left[1 + (1-2\gamma)z + \left(\gamma^2 - \gamma + \frac{1}{4} - \frac{1}{2}(3\gamma + \gamma_{31} + \gamma_{32}) \right) z^2 \right] . \tag{4.36}$$

The relevance of internal stability, i.e. the stability of the stage solutions, for the solution of stiff ODE problems is discussed in [60]. We here search for methods with L-acceptable internal stability function $R_3(z)$ and hence require that $|R_3(\infty)| = 0$. This is satisfied if and only if

$$\gamma_{31} = 2\gamma^2 - 5\gamma + \frac{1}{2} - \gamma_{32} \,. \tag{4.37}$$

Under this condition, it easily follows that the internal stability function $R_3(z)$ is A-acceptable, and hence L-acceptable, for all values of $\gamma \in [\gamma_-, \gamma_+] = [1 - \frac{1}{2}\sqrt{2}, 1 + \frac{1}{2}\sqrt{2}]$. Further, the internal stability function of the second-stage solution is A-acceptable if and only if $\gamma \geq \frac{1}{4}$, and even L-acceptable if and only if $\gamma = \frac{1}{2}$.

Let us see whether we can combine the requirement (4.37) for internal L-stability of the third-stage solution with the conditions (4.34a) or (4.34b). Obviously, the case (4.34b) results in two L-stable, 3-stage, second-order W-methods with underlying ERK scheme RK32 and L-acceptable third-stage stability function $R_3(z)$. These methods are given by the parameters Γ:

$$\gamma = \gamma_\pm := 1 \pm \frac{1}{2}\sqrt{2} \,, \ \gamma_{21} = -2\gamma^2 + 2\gamma - \frac{1}{2} \,, \ \gamma_{31} = 2\gamma^2 - 5\gamma + 1 \,, \ \gamma_{32} = -\frac{1}{2} \,. \tag{4.38a}$$

Let us turn to discuss the case (4.34a). Notice that we now have two conditions on the parameter γ_{31}, namely (4.37) and the one in (4.34a). They are both satisfied simultaneously if and only if

$$\gamma_{32} = \frac{6\,\gamma^2 - 10\,\gamma + 1}{4(1-\gamma)} \quad \text{for} \quad \gamma \neq 1\,.$$

We exclude $\gamma = \gamma_\pm$ because these values lead to $\gamma_{32} = -\frac{1}{2}$ and this case is not considered here. Both conditions on γ_{31} cannot be satisfied simultaneously if $\gamma = 1$ and hence we also exclude this value of γ. The interval $[\gamma_-, \gamma_+]$ for γ (leading to an L-acceptable third-stage stability function $R_3(z)$) is a subinterval of the given admissible range $[g_-, g_+]$ of γ-values for an L-acceptable stability function $R(z)$, see (4.35). Therefore, we obtain that the class of L-stable, 3-stage, second-order W-methods with underlying ERK scheme RK32 and L-acceptable third-stage stability function $R_3(z)$ is given by the parameters Γ satisfying either the conditions

$$\begin{gathered} \gamma \in (\gamma_-, \gamma_+)\,,\ \gamma \neq 1\,, \\ \gamma_{21} = -2\gamma^2 + 2\gamma - \frac{1}{2}\,,\ \gamma_{31} = \frac{8\gamma^3 - 22\gamma^2 + 12\gamma - 1}{4(\gamma-1)}\,,\ \gamma_{32} = \frac{6\gamma^2 - 10\gamma + 1}{4(1-\gamma)}\,, \end{gathered} \tag{4.38b}$$

or the conditions stated in (4.38a). This class of W-methods with class parameter γ forms the basis of the following investigations and we refer to it as ROS32(γ).

A third-order method for linear, autonomous problems: We can choose the free parameter γ in the class ROS32(γ) such that the stability function $R(z)$ of the method is a third-order approximation to the exponential function. This implies that the resulting W-method, if applied with exact Jacobian, is third-order accurate for linear, autonomous ODE systems. This method, ROS32(γ_3), is given by the choice

$$\gamma = \gamma_3 := -\frac{1}{2}\sqrt{2}\cos\left(\frac{1}{3}\arctan\left(\frac{1}{4}\sqrt{2}\right)\right) + 1 + \frac{1}{2}\sqrt{3}\sqrt{2}\sin\left(\frac{1}{3}\arctan\left(\frac{1}{4}\sqrt{2}\right)\right) \approx 0.436\,.$$

AMF and stability for ROS32(γ)-AMF : We can apply the methods ROS32(γ) in the form (4.7) with AMF and obtain schemes referred to as ROS32(γ)-AMF. We apply these schemes to the test equation (4.29) and obtain the factorized stability function $R(z_T, z_0, z_1, \ldots, z_d)$. The method parameters Γ satisfy the relation $r_3 = 0$, see (4.33). This simplifies the factorized stability function and yields (p is defined as in (4.30)):

$$\begin{aligned} R(z_T, z_0, z_1, \ldots, z_d) = 1 &+ \frac{(-1 + 12\gamma^3 + 12\gamma + 2\gamma_{31})\,z}{12\gamma^2 p} + \frac{(6\gamma^2 + \gamma\gamma_{31})\,z^2}{6\gamma^2 p^2} \\ &+ \frac{(-12\gamma^3 - 12\gamma + 18\gamma^2 + 1 - 2\gamma_{31})\,z}{6\gamma^2 p^2} + \frac{z^3}{12p^3} + \frac{(-6\gamma^2 - 2\gamma\gamma_{31})\,z^2}{12\gamma^2 p^3} \\ &+ \frac{(-1 + 12\gamma^3 + 12\gamma - 24\gamma^2 + 2\gamma_{31})\,z}{12\gamma^2 p^3}\,. \end{aligned} \tag{4.39}$$

For a given value of $\gamma \in [\gamma_-, \gamma_+], \gamma \neq 1$, the corresponding value of γ_{31} is defined by the conditions (4.38a) or (4.38b). As in the previous section, let us now assume that z_T is within the stability domain of the method RK32 (the underlying ERK scheme of ROS32(γ)) and that $z_j \leq 0$, j=1(1)d. Then, for a given γ-value, we are looking (numerically) for the maximum value θ such that the

method ROS32(γ)-AMF is stable, i.e. $|R(z_T, z_0, \ldots, z_d)| \leq 1$, for all $z_0 \in W_\theta$. The values θ (for the cases $d = 1$ and $d = 2$) are obtained by a similar algorithm as described in [32] and are shown in Fig. 4.4. The results are independent of the value of d ($d = 1$ or $d = 2$). For $\gamma = \gamma_3$ we obtain $\theta \approx 50°$. The optimal value for γ with respect to the above stability property appears to be $\gamma \approx 0.335$ resulting in $\theta \approx 64°$.

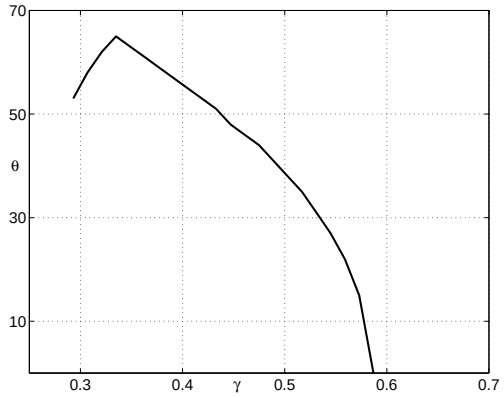

Figure 4.4: Maximum values of $\theta = \theta(\gamma)$ such that the factorized stability function R of ROS32(γ)-AMF satisfies $|R(z_T, z_0, \ldots, z_d)| \leq 1$ for all z_T within the stability domain of RK32, all $z_0 \in W_\theta$, and all $z_1, \ldots, z_d \leq 0$. We have considered $\gamma \in [\gamma_-, \gamma_+], \gamma \neq 1$, and obtained numerically the same maximum values of θ in the cases $d = 1$ and $d = 2$. For values of $\gamma > 0.58$ we could not find a value $\theta > 0$ such that the stability condition is satisfied.

Based on these theoretical investigations, we will apply two schemes from the class ROS32(γ)-AMF in our numerical experiments:

- ROS32(γ_3)-AMF: ROS32(γ_3) is an L-stable, 3-stage, second-order W-method with underlying ERK scheme RK32 and L-acceptable third-stage stability function $R_3(z)$. The method is third-order accurate for linear, autonomous problems (exact Jacobian). ROS32(γ_3)-AMF has a value $\theta \approx 50°$.
- ROS32(0.335)-AMF: ROS32(0.335) is an L-stable, 3-stage, second-order W-method with underlying ERK scheme RK32 and L-acceptable third-stage stability function $R_3(z)$. ROS32(0.335)-AMF has the optimal value $\theta \approx 64°$.

4.6 Selection of schemes Ψ_0 and Ψ_1 for the OPS methods

In this section we detail the integration schemes which we use as approximate evolution operators Ψ_0 and Ψ_1 in the OPS approach (Strang splitting), Eq. (4.15). Sec. 4.3 and Sec. 4.5 provide a variety of different explicit and linearly-implicit time stepping schemes, backed by theoretical investigations, which we can now choose from. A couple of other implicit schemes from the literature which could be a good choice for Ψ_1 are discussed in Sec. 4.7.
As discussed before, the taxis part F_0 of the problem should be integrated explicitly in time. Following the discussion in Sec. 4.3 and Sec. 4.4, we consider the ERK schemes ME, RK32, and RKF2(3) as candidates for the explicit method Ψ_0 in the OPS schemes.

Also, as discussed before, the diffusion–reaction part F_1 should be integrated by an implicit scheme for stability reasons. Therefore we select for Ψ_1 some of the Rosenbrock-type methods considered in the previous section. We apply them with AMF to reduce the amount of linear algebra work. However, in the situation of OPS, the factorization (4.13) leads to a matrix T in the Rosenbrock-type scheme which is a $\mathcal{O}(\tau)$-approximation of the true Jacobian of the right-hand side function F_1. (Remember that in Sec. 4.5, the right-hand side function is always $F = F_0 + F_1$, and we have neglected the Jacobian of F_0 when defining the matrix T.) This approximation property of T ensures that the ROS3-AMF method applied as approximate evolution operator Ψ_1 will be third-order accurate (if F_1 is autonomous). We also consider the methods ROS2(γ_-)-AMF and ROS2(0.59)-AMF as implicit methods Ψ_1 in OPS schemes. We do not consider the scheme ROS2(γ_+)-AMF as method Ψ_1 in the OPS approach because numerical tests (not presented here) with the biomathematical models described in Sec. 2.3.2, have shown the inefficiency of the resulting OPS methods. Further, notice the poor numerical performance of ROS2(γ_+)-AMF in comparison with ROS2(γ_-)-AMF when applied to the (unsplit) MOL-ODE system, see Chap. 5. We also do not consider the 3-stage, second-order W-methods ROS32(γ)-AMF as methods Ψ_1 because they are constructed with the unsplit MOL-ODE system in mind, as considered in the previous section. They would lead to computational overhead (three stages but only order two) in an OPS setting.
We refer to a specific OPS method by the name OPS–Ψ_1–Ψ_0 and, to summarize, we will test the explicit methods $\Psi_0 \in \{$ ME, RK32, RKF2(3) $\}$, and the implicit methods $\Psi_1 \in \{$ ROS2(γ_-), ROS2(0.59), ROS3 $\}$ applied with AMF.

4.7 Alternative methods for the MOL-ODE and different splitting approaches

The numerical solution of ODEs has been and is a highly active field of numerical analysis and scientific computing. A variety of special problem as well as all-purpose ODE solvers have been developed. In this section we present two schemes, VODPK and ROWMAP, which aim at solving large stiff ODE systems. We will use both schemes as reference methods in our numerical tests. Further, we mention some methods which can be employed as implicit approximate evolution operators Ψ_1 in OPS methods. We also discuss a different splitting approach for the MOL-ODE, source splitting [33], and a splitting approach on the PDE level of the TDR system [58]. The latter splitting method has been tested successfully for a TDR system describing a bacterial pattern formation process.

General purpose ODE solver for stiff ODEs.
Implicit or linearly-implicit methods have to be used for the solution of the MOL-ODE due to stability requirements (stiffness). These methods require some information about the Jacobian matrix of the problem. The Jacobian matrix of the MOL-ODE resulting from the semi-discretization of a TDR system is of large dimension and, although sparse, has a very large bandwidth. This rules out the use of band solvers for the solution of linear systems involving the Jacobian. To circumvent this problem we restrict attention to ODE solvers which do not require the Jacobian explicitly but only products of the Jacobian and vectors where the vector is arbitrary. A finite difference approximation of these products can be computed from the right-hand side function of the ODE system

with just two function evaluations. Guided by these requirements we select the codes VODPK and ROWMAP.
VODPK [7] is a variable-coefficient ODE solver with the preconditioned Krylov method GMRES for the solution of linear systems. It is based on the VODE and LSODPK packages. We use VODPK with default parameters and set the flag MF=21 (method based on BDF formulas up to order 5 with restarted GMRES). Experiments with preconditioned VODPK (using AMF) resulted in increased computational costs and no performance gain. Therefore we do not consider this case here.
ROWMAP [64] is based on the 4-stage ROW-methods of order 4 of the code ROS4 [23] and implements a special multiple Arnoldi process (MAP [50, 63]) for the solution of the stage equations. We use the code with default parameters. The order 4 of ROWMAP is obtained provided that the products of the Jacobian and vectors are exact. We already mentioned earlier that the discretization of the taxis part of the TDR systems is only Lipschitz continuous due to the application of limiter functions. Further, as in the case of Model 2, nonlinear reactions can lead to more non-differentiable terms in the right-hand side of the MOL-ODE. Hence, we cannot guarantee that the products of the Jacobian and vectors used in ROWMAP (computed by a finite difference approximation) are exact and therefore we must expect a less than fourth order behaviour of ROWMAP. This, in turn, has an influence on the reliability of the time step size control.
Inaccurate products of the Jacobian and vectors are not a problem for the order of the code VODPK because there these products are only used in the Newton process for the solution of the nonlinear stage equations. Inexact products may merely slow down the convergence of the Newton process.

Alternative methods for Ψ_1 in OPS schemes.
In [14] we have successfully used a linearly-implicit variant of the trapezoidal splitting method [30, 12] as method Ψ_1 in the OPS approach. This method has a consistency order two and, applied to the test equation $y'(t) = A_1 y(t) + A_2 y(t)$ with real matrices A_1 and A_2, we obtain an amplification matrix (the equivalence of the stability function in the scalar case) which is A-acceptable (i.e. its norm is bounded by one), if the matrices A_1 and A_2 commute and if they have a nonpositive logarithmic matrix norm, see [12]. Other linearly-implicit splitting methods are derived in [12] which are A- and also L-stable for the test equation above even if the matrices do not commute. The L-stable methods are especially interesting for very stiff problems. Numerical experiments with these methods have shown that their application in OPS schemes does not lead to an improved performance compared to the application of Rosenbrock-type AMF methods in OPS for the models discussed in this paper. Therefore we do not consider them in the numerical experiments section here. However, for different models or other applications they might be the methods of choice.

Alternative splitting approaches.
Another splitting approach for the solution of the MOL-ODE (4.1), source splitting, is considered in [33] and applied to a variation of Model 2 in [15]. The formulas

$$y_{k+1} = z(t_k + \tau) + \frac{\tau}{2}(F_0(z(t_k + \tau)) - F_0(y_k)) \,, \tag{4.40a}$$

$$z'(t) = F_0(y_k) + F_1(z(t)) \,,\ t \in [t_k, t_k + \tau], z(t_k) = y_k \,, \tag{4.40b}$$

define a time step of a source splitting method for problem (4.1) with a right-hand side function $F(t, y) = F_0(t, y) + F_1(t, y)$. This method reduces to the modified Euler (ME) ERK method if

applied with $F_1 \equiv 0$ and is of second-order (assuming exact integration of (4.40b)). Note that F_0 is treated as an additional, constant source in the ODE (4.40b). We have solved the ODE (4.40b) in each time step of the source splitting method with one step of ROS2(γ_-)-AMF in [15]. Numerical experiments in this paper also demonstrate that the performance of the source splitting method approximately equals the performance of the ROS2(γ_-)-AMF method applied to the (not splitted) MOL-ODE. No numerical experiments with the source splitting method are presented in this thesis.

Recently, Tyson et al. [58] have described a splitting algorithm and applied it to a specific TDR model. They perform a splitting of the TDR model at the PDE level already and their approach is closely related to our proposed OPS schemes. They use the software package CLAWPACK [40] to deal with the taxis part of the problem and also an L-stable implicit method for the diffusion–reaction part. The AMF methods presented in this thesis differ from the approach of Tyson et al. and also from the OPS approach as the AMF methods do not split the equations and hence avoid the associated splitting error. We did not compare our proposed algorithms with the methods proposed by Tyson et al. [58] yet.

Chapter 5

Numerical Experiments and Discussion

In this chapter we present numerical results which demonstrate the performance of the time stepping algorithms of Chap. 4 applied to the MOL-ODEs obtained by the discretization in space (Chap. 3) of the biomathematical models described in Sec. 2.3.2.
The discretization in space is done on grids with grid width $h = \frac{1}{100}$ or $h = \frac{1}{200}$, see Sec. 3.1. These resolutions are sufficiently fine to resolve the phenomena exhibited by the different models. Following standard practice, we have implemented the AMF methods (Sec. 4.5) and the OPS methods (Sec. 4.6) with variable time step sizes (Sec. 4.1) in FORTRAN77. The embedded first-order solution is used to obtain an estimate of the local error of the current time step in the AMF schemes. The time step is selected on the basis of an error per step (EPS) control which aims to keep this estimate below a mixed (relative and absolute) threshold depending on the user supplied tolerance TOL ($= ATOL = RTOL$). The higher order solution is used to advance an accepted step (local extrapolation). The OPS methods use Richardson extrapolation to obtain a local error estimate of the current step and then the same EPS control to select the step size. They step forward with the solution obtained from two half-steps (doubling). The Jacobians of the diffusion and the reaction parts of the right-hand side of the MOL-ODE are evaluated at the beginning of a time step (AMF) or at the beginning of a Richardson step (OPS). We compute finite difference approximations to the true Jacobians of the split functions.
We compare the computed (with our or other methods) solutions, y_{comp}, of a MOL-ODE at final time (corresponding to the examples considered here) against a reference solutions y_{ref}. We obtain these reference solutions of the ODE systems with the standard integrator VODPK [7] requiring the very stringent tolerance $TOL = 10^{-12}$. The error estimate $err := \|y_{comp} - y_{ref}\|$ between computed solutions and reference solutions is measured in the scaled l_2-norm,

$$\|v\| = \left(\frac{1}{m} v^{\mathsf{T}} v\right)^{\frac{1}{2}}, \qquad v \in \mathbb{R}^m .$$

This norm is used throughout this chapter, and in tables and figures we print the logarithm to the base 10 of the error norm. This procedure implies that we consider the temporal error (including splitting errors) of the solution y_{comp} in the numerical experiments. We do not consider the spatial accuracy of the solution here (this has been done for Model 1 in Sec 3.4).
For each model (Model 2, Model 3, Model 4) we consider two different scenarios: $\varepsilon > 0$ (as defined in the corresponding model description) and $\varepsilon = 0$, and for each of these a coarse ($h = \frac{1}{100}$)

and a fine ($h = \frac{1}{200}$) spatial mesh. The integration schemes are run for seven tolerance values $TOL = 10^{-3}, 10^{-3.5}, \ldots, 10^{-6}$, except indicated otherwise, in all four test cases for each model. We do not provide details of all numerical experiments but instead give a short summary of the observations. The numerical experiments with Model 2 are discussed in more detail. All test runs are performed on one processor of a HP Convex X-Class server under HP-UX.

5.1 Tumour–induced angiogenesis — Model 2

Description of solution: The solution n of the equation for the EC density of this problem has initially peaks near the right boundary of the domain. The cells there are migrating to the left—forming a stream which moves up the present TAF (c_1) gradient as time proceeds. No cell proliferation takes place in the beginning of the simulation because the c_1 concentration at the cells is below the threshold $c_1^\star$. Later proliferation leads to a strong, local increase of the cell density. The cells also take up TAF. This results in changes in the TAF gradients and causes lateral cell movement and hence a widening of the cell streams. The cell streams turn towards the centre of the TAF source (the tumour) once they are close enough to the left boundary. Fig. 5.1 gives cell density plots at three different output times for the model with ($\varepsilon = 0.001$) and without ($\varepsilon = 0$) cell random motility (notice the different final times). We see that the process proceeds faster if cell random motility is present and that in this case also the lateral cell movement is more pronounced (leading to a closed wave front towards final time).

Numerical order of convergence: We start with assessing the numerical order of convergence of the OPS and AMF methods by using the ODE solver with fixed time step size. We select the method OPS-ROS2(γ_-)-RK32 for this test because it will turn out that this method will be one of the most efficient for this example. We consider only this method because the variable step size experiments later in this section suggest that the other methods behave similarly (almost parallel lines in the accuracy vs. CPU time plots). We discretize Model 2 on a spatial grid with $h = \frac{1}{100}$. In the first test we use a random motility coefficient $\varepsilon = 0.001$ of the ECs and the corresponding final time $T = 1.0$. Then we obtain the following table:

steps	30	35	40	50	75	100	150	200	300	400	600	800
err	-1.51	-1.65	-1.75	-1.92	-2.24	-2.48	-2.81	-3.06	-3.41	-3.65	-4.01	-4.26
order	—	2.09	1.72	1.75	1.82	1.92	1.87	2.00	1.99	1.92	2.04	2.00

The first row of the table gives the number of times steps taken to reach T, the attained norm of the error is given in the second row, and the numerical order of convergence in the third row. The numerical solution blows up if the number of time steps is reduced to 25. We can clearly observe the expected second order convergence behaviour. We also note that the numerical solution is positive for steps ≥ 40 and that there is some slight undershoot (negative solution values) for steps $= 30$ and steps $= 35$.

We now turn to the situation without cell random motility, $\varepsilon = 0$, and hence $T = 1.2$ (everything else unchanged). Now we obtain the table:

steps	35	40	50	75	100	150	200	300	400	600	800
err	-0.96	-1.43	-1.67	-2.05	-2.28	-2.63	-2.88	-3.26	-3.54	-3.92	-4.19
order	—	8.11	2.48	2.16	1.84	1.99	2.00	2.16	2.24	2.16	2.16

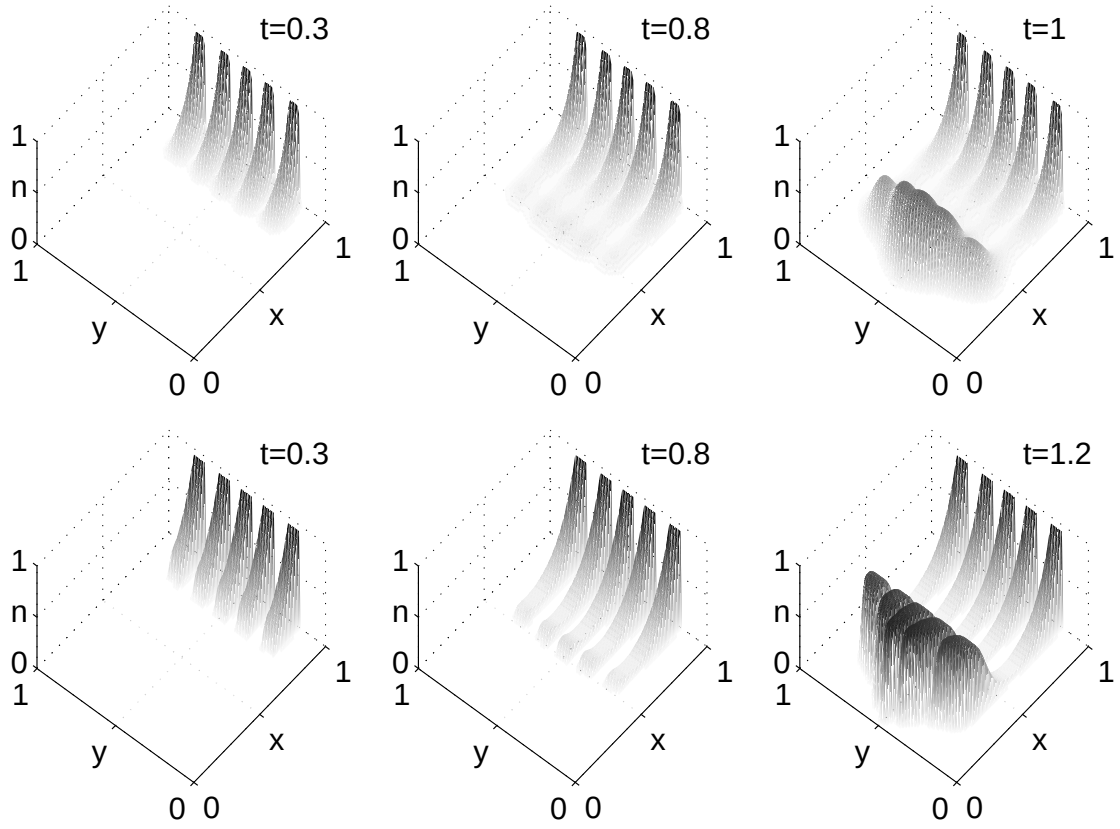

Figure 5.1: EC density n solutions of Model 2 for different simulations times and with or without random motility of the ECs. The random motility coefficient of the ECs is $\varepsilon = 0.001$ in the top row plots, and there is no EC random motility in the bottom row plots ($\varepsilon = 0$); all other parameters are as in Sec. 2.3.2.1.

The order of the method is again two. The numerical solution blows up if we only use 25 or 30 time steps. The smallest component in the numerical solution is exactly zero for steps ≥ 75; there are negative solution components (small in magnitude) if the number of time steps is smaller and for too few time steps this causes the solution to blow up.

Comparison of AMF methods: We now turn to investigate the performance of the variable step size implementations and first concentrate on the AMF schemes. Fig. 5.2 gives the error vs. CPU time plots for the cases with and without EC random motility on a spatial grid with $h = \frac{1}{100}$. Clearly, all methods converge to the reference solution with approximately the same order of convergence.

In the case of $\varepsilon = 0.001$, Fig. 5.2 (left), only ROS2(γ_-)-AMF and ROS32(0.335)-AMF return a solution for all requested tolerances TOL; the solutions of the other schemes blow up for tolerance requirements which are too weak. In the case of $\varepsilon = 0$ no such problems are observed and the methods return solutions even for $TOL = 10^{-3}$. Below we claim that the blow-up of the (numerical) solutions in Model 2 is caused by negative solution values in combination with the reaction term in the taxis equation. Negative solution values are introduced if the selected time step sizes are too large. The number of time steps (scaled by the final time T) which a given method takes to reach the final time T for a given value of TOL is considerably larger if $\varepsilon = 0$ than if $\varepsilon = 0.001$.

(The error in the numerical solution is smaller if $\varepsilon = 0$ compared to the error if $\varepsilon = 0.001$ for fixed TOL; the (scaled) number of time steps taken by a method to reach a certain error err is similar for both values of ε.) Hence, in the case of $\varepsilon = 0$ the step size control selects smaller time steps (caused by the steep fronts (non-smoothness) in the EC density solution in this case). On the other hand, if $\varepsilon > 0$ then the steep fronts are smoothed and the step size control selects larger time steps which in turn causes the negative solution components.

The successfully computed solutions of all methods are nonnegative except for some (small in magnitude) negative values if weaker (successful) tolerances are requested. In the case $\varepsilon = 0$, the smallest solution component is exactly zero for all requested tolerances for the methods ROS2(γ_-)-AMF, ROS2(0.59)-AMF, and ROS32(0.335)-AMF.

The method ROS3-AMF is slightly more accurate than the other AMF methods for this model and higher accuracy demands but not very robust for less strict accuracy requirements. The method ROS2(γ_+)-AMF shows the worst performance. ROS2(γ_-)-AMF appears to be very robust; ROS2(0.59)-AMF is slightly more efficient. There are hardly any differences between ROS32(γ_3)-AMF and ROS32(0.335)-AMF. The latter method is slightly more stable (as supported by the theory). Similar conclusions can be drawn from numerical experiments if the spatial resolution is refined to a grid with $h = \frac{1}{200}$.

Comparison of OPS methods: Let us next consider the performance of the variable step size implementations of OPS schemes applied to Model 2. Fig. 5.2 gives the error vs. CPU time plots for the cases with and without EC random motility on a spatial grid with $h = \frac{1}{100}$.

Again, all methods converge to the reference solution with approximately the same order of convergence. The OPS methods almost avoid a blow-up of the numerical solutions. Blow-up is only observed for $TOL = 10^{-3}$ and the methods OPS-ROS2(0.59)-RK32 and OPS-ROS2(0.59)-RKF2(3) if $\varepsilon = 0.001$, and OPS-ROS3-RK32 if $\varepsilon = 0.001$ or $\varepsilon = 0$.

It is noteworthy that the OPS schemes with implicit method ROS2(γ_-)-AMF demonstrate the most stable behaviour of all schemes tested here. This can probably be attributed to the L-stability of the underlying Rosenbrock method but is in contrast to the stability property derived for the factorized scheme applied to the test equation (4.29). An explanation here could be that ROS2(γ_-)-AMF is stable with respect to (4.29) on a very large domain for the values $\lambda_0, \ldots, \lambda_d$ but this domain does

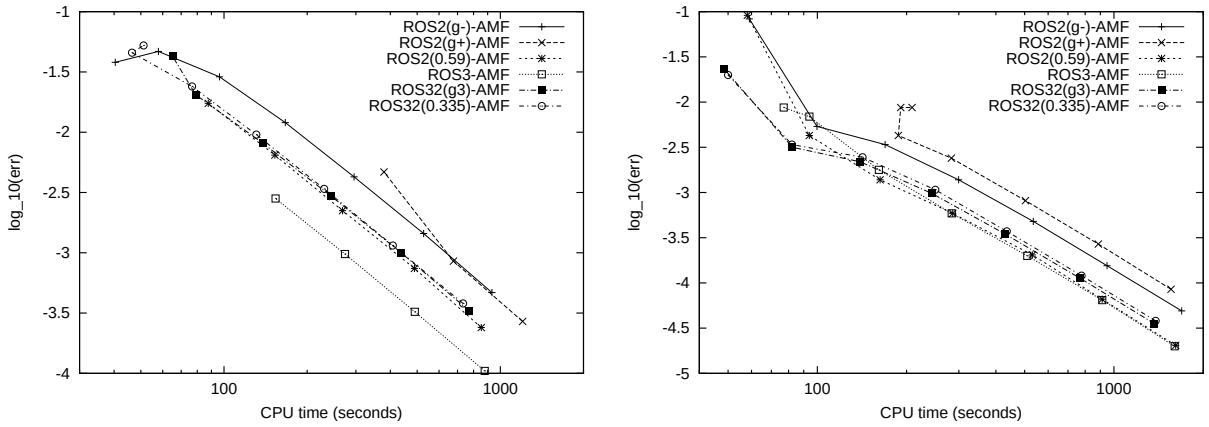

Figure 5.2: Error vs. CPU time plots of various AMF methods applied to Model 2 with $\varepsilon = 0.001$, $T = 1$ (left) and $\varepsilon = 0$, $T = 1.2$ (right). We use $h = \frac{1}{100}$; all other parameters are as given in Sec. 2.3.2.1.

not extend to infinity. However, this issue has not been investigated. The schemes with implicit method ROS2(0.59)-AMF or ROS3-AMF have problems with stability if the required tolerance is not sufficiently stringent.
If we take a look at the behaviour of the OPS schemes with respect to the choice of the explicit method then we see that RKF2(3) gives the most accurate results, followed by RK32 and eventually ME. Especially in Fig. 5.3 (bottom) we see that OPS-ROS2(γ_-)-RKF2(3) suffers from stability problems in the lower accuracy range whereas the other two OPS methods with implicit scheme ROS2(γ_-)-AMF are still unaffected in this range of accuracy.
Altogether we recommend the application of OPS-ROS2(γ_-)-RK32 and OPS-ROS2(γ_-)-RKF2(3)

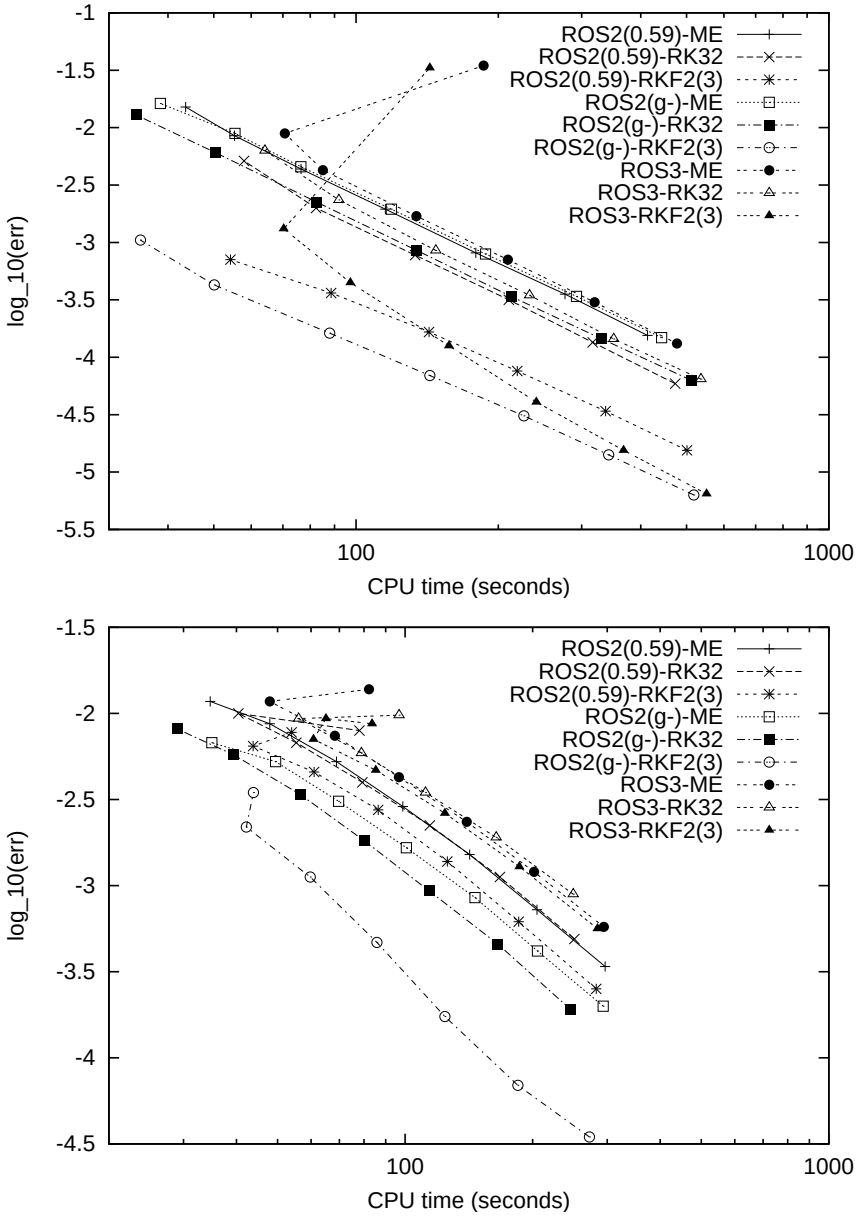

Figure 5.3: Error vs. CPU time plots of various OPS methods applied to Model 2 with $\varepsilon = 0.001$, $T = 1$ (top) and $\varepsilon = 0$, $T = 1.2$ (bottom). We use $h = \frac{1}{100}$; all other parameters are as given in Sec. 2.3.2.1.

for the solution of Model 2. Similar conclusions can be drawn from numerical experiments on the fine spatial grid with $h = \frac{1}{200}$.

Comparison with reference methods: Fig. 5.4 gives the error vs. CPU time plots of a selection of AMF and OPS methods, and of the reference methods (general purpose, stiff ODE solver) VODPK and ROWMAP for the solution of Model 2 on the coarse and the fine spatial grid.

We see that the OPS schemes demonstrate the best performance, followed by the AMF schemes. The reference methods are not suitable for the solution of Model 2 if only low to moderate accuracies are required. They have many rejected steps and the numerical solutions obtained by ROWMAP blow up except for the two most strict tolerance requirements in the case $\varepsilon = 0$. Both reference methods return numerical solutions with negative components in the accuracy range considered. For higher accuracy demands they perform well and will, due to their higher order, eventually outperform the splitting methods.

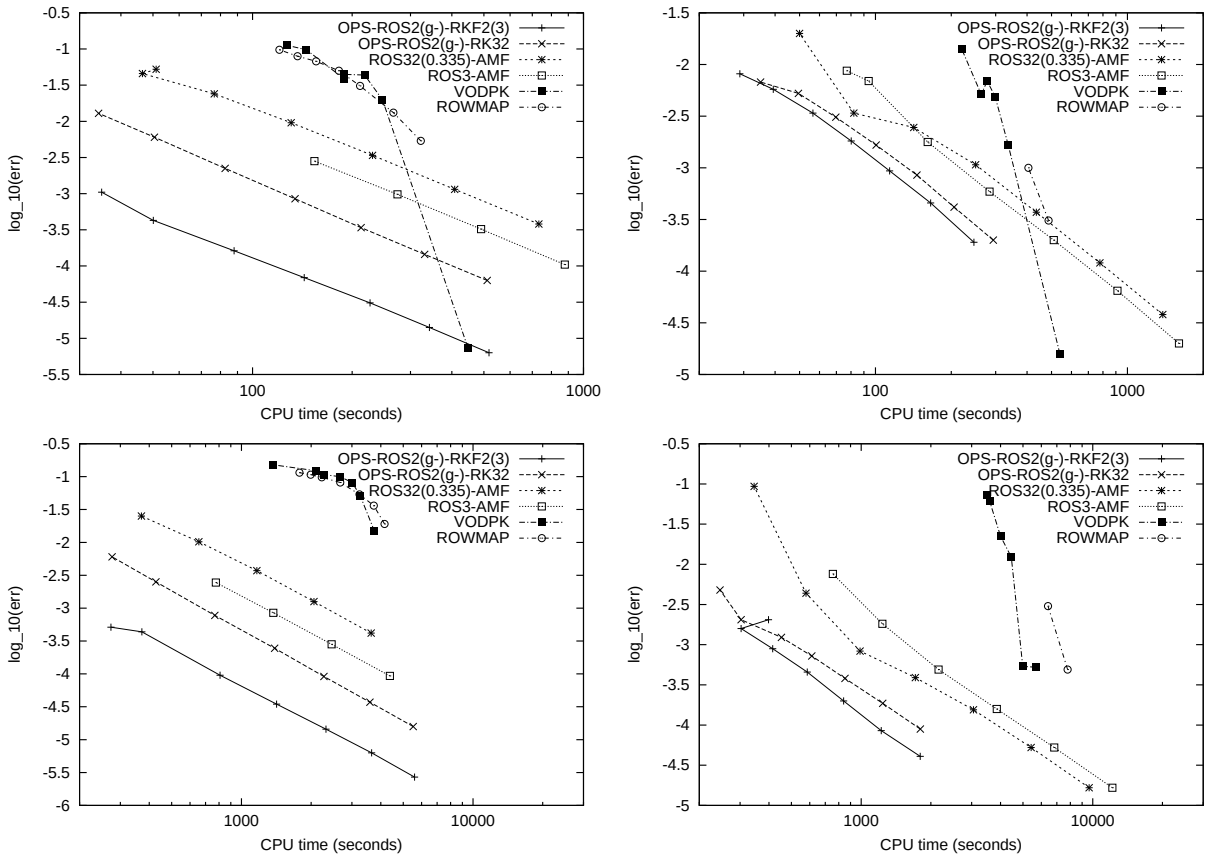

Figure 5.4: Error vs. CPU time plots of various methods applied to Model 2 with $\varepsilon = 0.001$, $T = 1$ (left column) and $\varepsilon = 0$, $T = 1.2$ (right column), and spatial grid width $h = \frac{1}{100}$ (top row) and $h = \frac{1}{200}$ (bottom row). All other parameters are as given in Sec. 2.3.2.1.

Positivity and blow-up: Especially with the AMF schemes and the reference methods we have seen that the numerical solutions may blow up if the requested tolerance is too weak. Looking at the course of integration, we can see that in such a case negative solution values appear in elements of the numerical solution which approximate the solution of the taxis equation. These negative values grow larger in magnitude – at first slowly and then more and more rapidly until

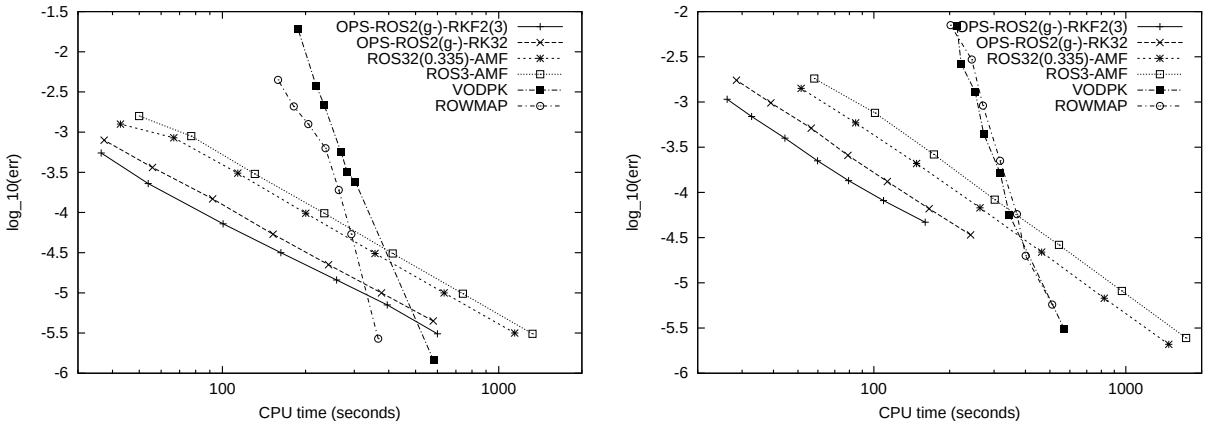

Figure 5.5: Error vs. CPU time plots of various methods applied to Model 2 with $\varepsilon = 0.001$, $T = 1$ (left) and $\varepsilon = 0$, $T = 1.2$ (right). We use $h = \frac{1}{100}$ and $\beta = \mu = 0$; all other parameters are as given in Sec. 2.3.2.1.

they blow up.

The behaviour described is caused by the reaction term of the taxis equation. To this end, consider the simplified reaction term $f_R(n) := \widetilde{c_1}\mu n(1-n) - \beta n$ with nonnegative parameters, where $\widetilde{c_1}$ represents the term $\max\{0, c_1 - c_1^*\}$ of the original reaction function. The scalar Jacobian $\frac{\partial f_R(n)}{\partial n} = \widetilde{c_1}\mu - \beta - 2\widetilde{c_1}\mu n$ is positive for $n = 0$ if $\widetilde{c_1} > \frac{\beta}{\mu}$. Hence the fixed point $n = 0$ of $f_R(n)$ is unstable in this case. The choice of parameters of Model 2 leads to $\frac{\beta}{\mu} = 0.04$ and hence the conditions for instability are easily satisfied.

We have studied numerically the effect of switching off the reaction term in the taxis equation and indeed, all methods can successfully compute the solution up to the final time T even for $TOL = 10^{-3}$. The results on a spatial grid with $h = \frac{1}{100}$ are given in Fig. 5.5. There we can also see that the performance of the methods ROWMAP and VODPK also improves considerably but the splitting methods still have a clear advantage.

Another technique to enforce a nonnegative solution is to apply *clipping*, see e.g. [62]. This means that after each time step of a method all negative components of the solution are set to zero. Clipping interferes with mass conservation and should therefore only applied with care and if really necessary. In our case, clipping prevents the blow-up of numerical solutions but does not improve the performance of the methods. Therefore we rather recommend to apply the methods with a more stringent tolerance requirement TOL.

5.2 Tumour–induced angiogenesis — Model 3

Description of solution: The most interesting part of this model is again the evolution of the EC density n depicted in Fig. 5.6 for the cases with ($\varepsilon = 0.00035$) and without ($\varepsilon = 0$) cell random motility. In contrast to Model 2, the model considered here does not take account of EC proliferation but concentrates on the development of the ECs near the tips of the new blood vessels. The total mass of ECs is conserved in the model.

The solution n of the equation for the EC density of this problem has initially peaks near the left boundary of the domain. The cells there are migrating to the right and move up the present TAF (c_1)

gradient as time proceeds. The gradient in the fibronectin concentration c_2 acts counterproductive by slowing down the migration of the ECs towards the right domain boundary. On the other hand, it enhances the lateral movement of ECs. This lateral movement is clearly visible in the plots: the outer EC clusters move laterally as the time increases and subsequently, around time $t = 5$, join to form one central cluster. Later at time $t = 10$ we see that the cells have spread out even more in the domain; there is also some movement backward to the left boundary. From now on, the ECs advance only slowly towards the tumour (right-hand boundary) due to the chemotactic function chosen in Model 3.

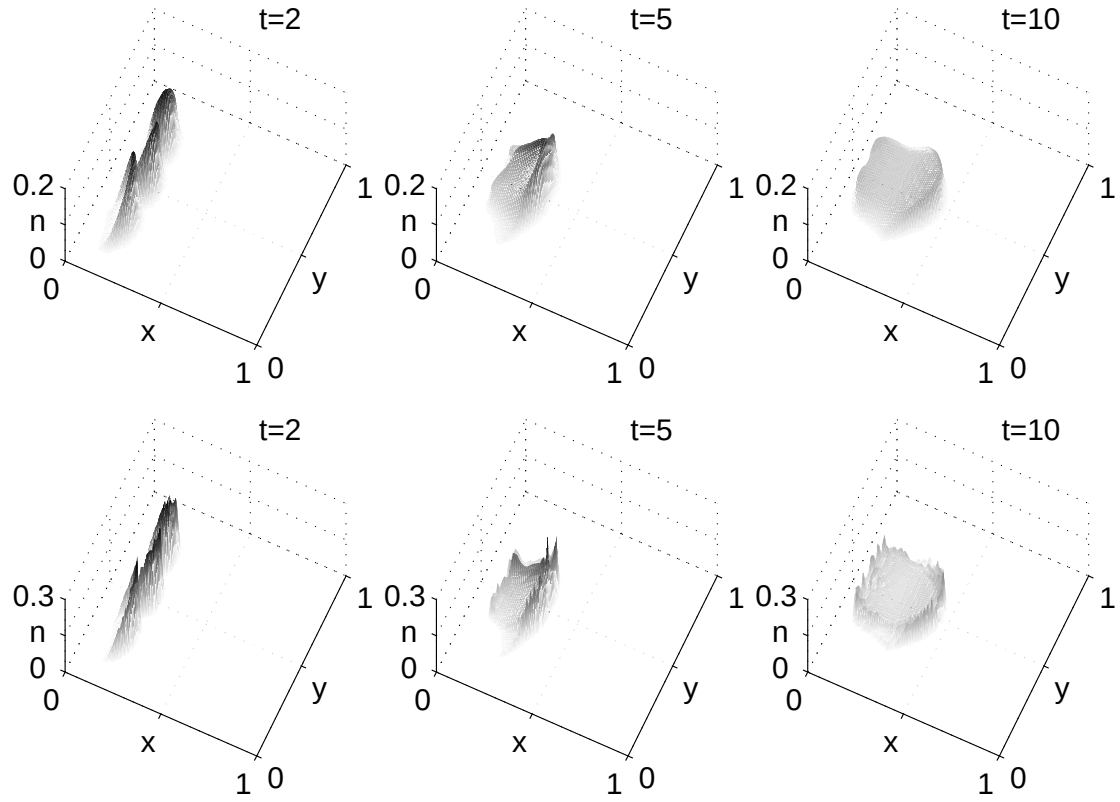

Figure 5.6: EC density n solutions of Model 3 for different simulations times and with or without random motility of the ECs. The random motility coefficient of the ECs is $\varepsilon = 0.00035$ in the top row plots, and there is no EC random motility in the bottom row plots ($\varepsilon = 0$); all other parameters are as in Sec. 2.3.2.1.

Comparison of selected AMF and OPS schemes with reference methods:
The most promising AMF methods for this model are ROS32(0.335)-AMF (ROS32(γ_3)-AMF behaves similarly) and ROS2(γ_-)-AMF (ROS2(0.59)-AMF behaves similarly). The enhanced accuracy of ROS3-AMF, as observed in the experiments with Model 2, does not show up for this model. Again, ROS2(γ_+)-AMF cannot compete with the other AMF schemes. There are almost no rejected time steps for all methods. The total mass of the solution components corresponding to the cell density n in the model at final simulation time ($T = 5$) is almost the same as the mass at initial time. The difference is of the order of machine precision ($\approx 10^{-16}$) except for the weakest tolerances used where the difference can be as large as 10^{-6}. Therefore it is justified to say that the

integration schemes are mass conservative. Small negative components in the numerical solution are no difficulty in the numerical solution of Model 3. This is apparently in contrast to what is observed in the numerical solution of Model 2. However, there we have argued that the difficulties arise because of the reaction term in the cell density equation of Model 2. Here we have no such reaction term and hence this supports the arguments given in the previous section. We find (small in magnitude) negative components in the numerical solution for low tolerance requirements only, especially in the case with random cell motility ($\varepsilon = 0.00035$). This can be explained, as in Model 2, by observing that for the same value of TOL the methods take considerably more time steps to reach the final time if $\varepsilon = 0$ as if $\varepsilon = 0.00035$. In the following we consider the schemes ROS32(0.335)-AMF and ROS2(γ_-)-AMF only and their error vs. CPU time plots are given, in comparison with other methods, in Fig. 5.7.

We turn to discuss the numerical results obtained with the OPS methods applied to Model 3. They all behave almost identical in the moderate and higher accuracy range. The schemes with implicit method ROS3-AMF demonstrate small stability problems for low accuracy demands; the schemes with implicit method ROS2(γ_-)-AMF appear to be the most robust with this respect—especially for the case $\varepsilon = 0.00035$ on the finer spatial grid. Therefore we consider the schemes OPS-ROS2(γ_-)-RK32 and OPS-ROS2(γ_-)-RKF2(3) in the comparison in Fig. 5.7. The OPS schemes

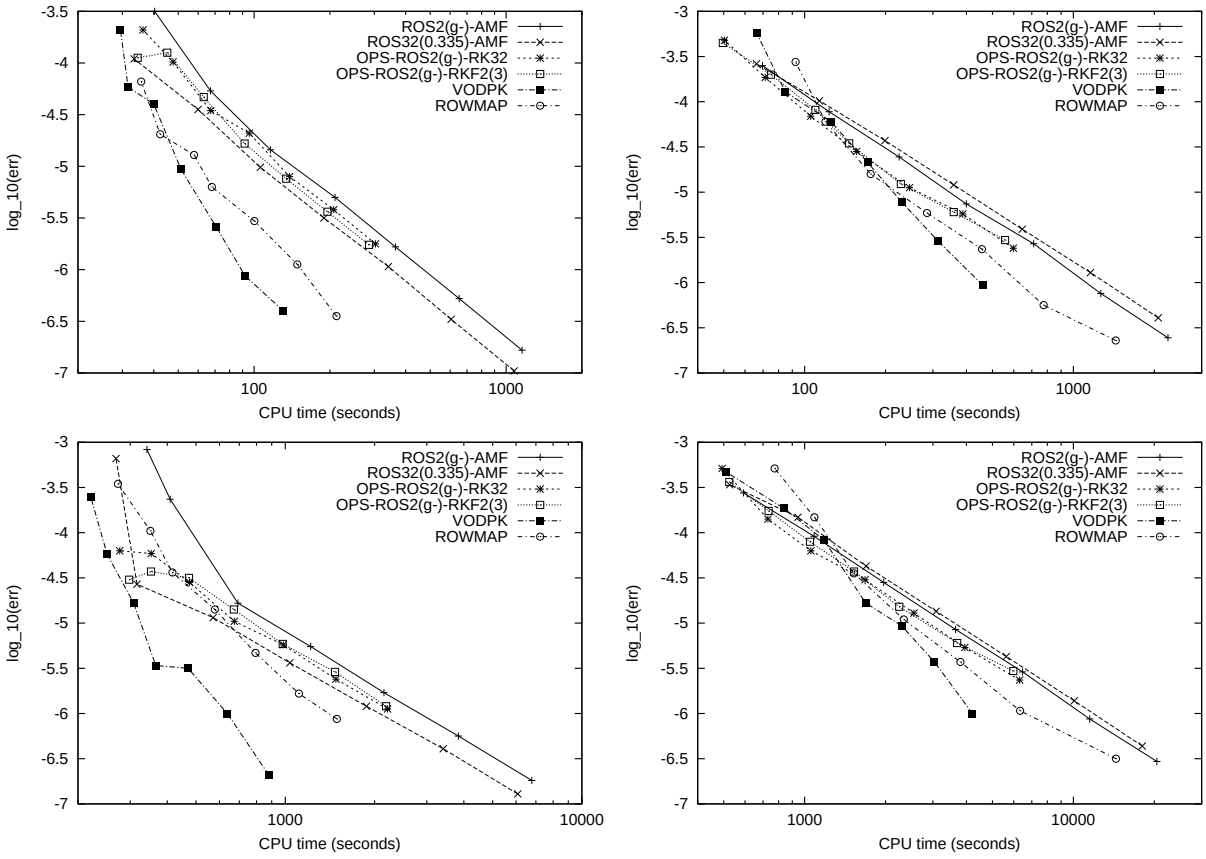

Figure 5.7: Error vs. CPU time plots of various methods applied to Model 3 with $\varepsilon = 0.00035$ (left column) and $\varepsilon = 0$ (right column), and spatial grid width $h = \frac{1}{100}$ (top row) and $h = \frac{1}{200}$ (bottom row). The final time is $T = 5$; all other parameters are as given in Sec. 2.3.2.1.

are also mass conservative and return nonnegative solutions under the same restrictions as stated for the AMF schemes in the previous paragraph.

From the error vs. CPU time plots in Fig. 5.7, we see that the AMF methods and the OPS schemes perform equally well for this example. Also, the reference methods VODPK and ROWMAP perform similarly in the lower and moderate accuracy range (except for $\varepsilon = 0.00035$ on the fine grid where VODPK outperforms ROWMAP). There is an advantage for the reference methods if $\varepsilon = 0.00035$ (not for ROWMAP on the fine grid), and a slight advantage for the splitting methods (AMF, OPS) if $\varepsilon = 0$. For higher accuracy demands the reference methods are superior (due to their higher order of accuracy). The reference methods have not as good mass conservation and positivity properties as the splitting schemes. The mass difference between the solutions at initial and final simulation time is in general in the order of 10^{-6} even for higher accuracy demands, and there are small in magnitude (but bigger than for the splitting methods) negative components in the solutions at final time for almost all tolerance requirements. These negative solution values are not harmful for the solution process of the MOL-ODE of this model.

We note that Model 3 is numerically much simpler to treat than Model 2: firstly, there is no reaction term in the taxis equation of Model 3 and such a term caused much of the trouble in the numerical simulation of Model 2 (blow-up due to negative solution components), and secondly, the boundary conditions in Model 3 are of zero-flux type whereas there is a combination of zero-flux and non-homogeneous Dirichlet boundary conditions in Model 2. Especially the inclusion of a reaction term in the taxis equation of Model 3 (e.g. in order to take EC proliferation into account) is expected to lead to a notably improved performance of the splitting schemes compared to the reference methods.

5.3 Tumour invasion — Model 4

Description of solution: The solution n of the equation describing the evolution of the tumour cell density of this model has an initial peak in the centre of the domain (representing the initially compact tumour mass). This peak spreads outward moving up gradients of the ECM density c_1 which is heterogeneous initially. This leads to a heterogeneous pattern in the cell density solution. These patterns are sharper if there is no cell diffusion (a break up of the initially compact cell mass can be observed) and more smeared with cell diffusion (the break up of cell mass is not so pronounced in this case). The total cell mass in the domain is a conserved quantity of the model. The tumour cells release MDE (c_2) which diffuses within the spatial domain. MDE in turn degrades ECM and hence leads to new gradients in the ECM density which give rise to further migration of the cells. The most interesting solution of this model is the cell density and Fig. 5.8 gives solution plots at three different output times for the cases with and without cell diffusion.

Comparison of selected AMF and OPS schemes with reference methods: We again start with looking at the performance of the AMF methods applied to the MOL-ODE obtained by the semi-discretization of the equations of Model 4. Again we note that ROS2(γ_+)-AMF requires significantly more CPU time than the other AMF methods to return a solution of the same accuracy. Also, the improved accuracy of ROS3-AMF, as observed in some of the numerical experiments with Model 2, does not show up. The remaining four methods of this type behave fairly similar and we single out ROS2(γ_-)-AMF and ROS32(0.335)-AMF for the numerical comparison with

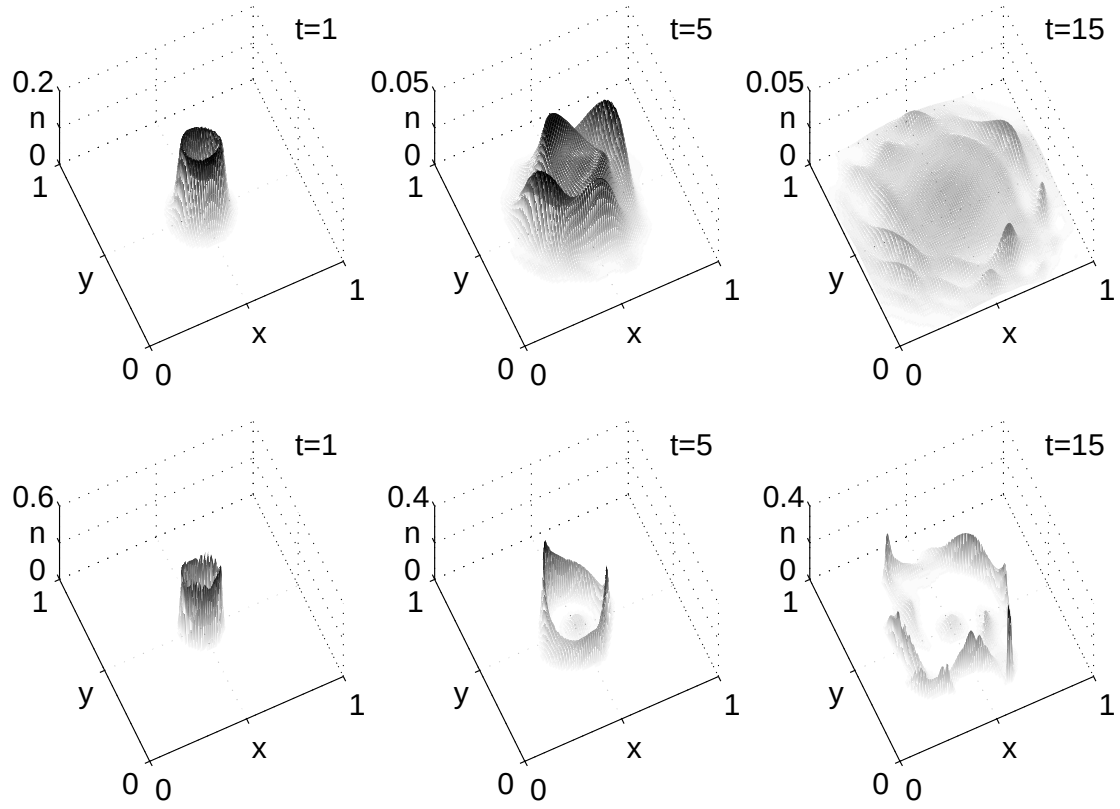

Figure 5.8: EC density n solutions of Model 4 for different simulations times and with or without random motility of the ECs. The random motility coefficient of the ECs is $\varepsilon = 0.001$ in the top row plots, and there is no EC random motility in the bottom row plots ($\varepsilon = 0$); all other parameters are as in Sec. 2.3.2.2.

the OPS schemes and the reference methods, see Fig. 5.9. There are virtually no rejected steps in the test runs with the AMF schemes. Also, the initial total mass of the tumour cells is conserved (up to machine precision) during the process of simulation until the final time T is reached. Finally, there are almost no negative solution components and if there are some (for very low accuracy requirements) then their order of magnitude is considerably smaller than machine precision.

We now turn to discuss the numerical experiments with the OPS schemes applied to this example. There are only small differences between all the methods and these are mainly due to the choice of the implicit scheme in the OPS method. If this implicit scheme is ROS3-AMF then the performance is worse than for all other methods in all four test cases. Concerning the other two implicit schemes tested, the performance depends on the choice of ε. If $\varepsilon = 0$ then the methods with implicit scheme ROS2(γ_-)-AMF have an advantage and if $\varepsilon = 0.001$ then the methods with implicit scheme ROS2(0.59)-AMF are slightly better (although the difference in this case is less than in the case with $\varepsilon = 0$). We observe slight stability problems in the case $\varepsilon = 0.001$ for very low tolerance requirements for all methods (rejected time steps in this case). These are the least pronounced for the method OPS-ROS2(γ_-)-RK32. The statements given above for the AMF methods concerning negative components in the solution and conservation of (tumour) cell mass also apply for the OPS schemes. Based on the observations we select the methods OPS-ROS2(γ_-)-RK32 and

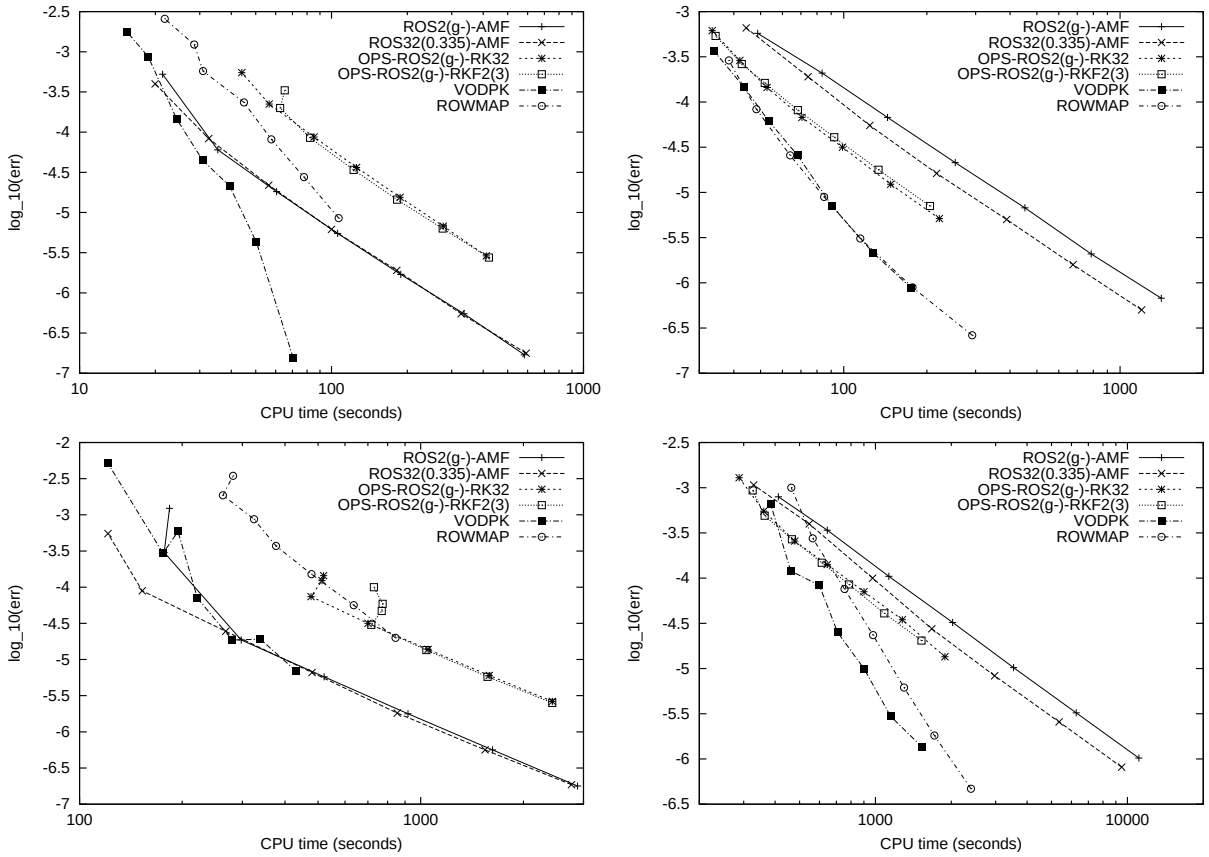

Figure 5.9: Error vs. CPU time plots of various methods applied to Model 4 with $\varepsilon = 0.001$ (left column) and $\varepsilon = 0$ (right column), and spatial grid width $h = \frac{1}{100}$ (top row) and $h = \frac{1}{200}$ (bottom row). The final time is $T = 5$; all other parameters are as given in Sec. 2.3.2.2.

OPS-ROS2(γ_-)-RKF2(3) for the numerical comparison in Fig. 5.9.

We see from the plots in Fig. 5.9 that VODPK turns out to be very efficient for this model. Due to increasing stiffness, this advantage of VODPK decreases for the finer grid resolution and more significantly, as reported in [16], if the (small) diffusion coefficient $d_2 = 0.001$ is enlarged by a factor of 10 or 100. We note that the AMF schemes can be applied with even less stringent tolerance requirements (e.g. up to $TOL = 10^{-2}$) in the case with cell diffusion ($\varepsilon = 0.001$) and then these schemes outperform VODPK (for consistency we do not plot these data points here).

In general, VODPK and ROWMAP preserve the cell mass well although, especially in the case $\varepsilon = 0$, not as good as the splitting schemes (mass conservation improves for increasing accuracy demands and eventually reaches the level of machine accuracy). VODPK and ROWMAP return nonnegative solutions if $\varepsilon = 0.001$ except for the weakest tolerance requirements $TOL = 10^{-3}, 10^{-3.5}$. In the case $\varepsilon = 0$, the solutions contain negative components for all tested tolerances TOL. The most negative values being around -10^{-5} for $TOL = 10^{-3}$ and reaching the level of machine accuracy for stricter tolerances TOL.

Returning to the plots in Fig. 5.9, we clearly see that the AMF schemes are more suitable than the OPS methods for the test case with cell diffusion ($\varepsilon = 0.001$); without cell diffusion ($\varepsilon = 0$), the situation is the opposite and the OPS schemes generally demonstrate a better performance.

It can also be observed that for cruder tolerances the methods based on the explicit method RK32

have a slightly improved behaviour compared to the corresponding methods based on the explicit methods ME or RKF2(3). We credit this advantage to the improved stability and positivity properties of RK32.
Finally, we mention that Model 4 (as Model 3 and in contrast to Model 2) is a TDR system without reaction term in the taxis equation. This serves to explain that slightly negative components in the numerical solution have no negative effect on the solution process. Also, the boundary conditions of Model 4 are of no-flux type and there are no inhomogeneous boundary conditions as in Model 2.

Chapter 6

Conclusions

In this thesis we have considered the numerical solution of taxis–diffusion–reaction (TDR) systems which arise in a variety of mathematical models of biological processes. The driving force for this work has been the aim to develop, implement, and test suitable numerical schemes for the simulation of the tumour-induced angiogenesis Model 2. The numerical technique derived (spatial discretization and solution of the MOL-ODE) has proved to be very appropriate for this model. We have also demonstrated that the splitting methods derived for the solution of the MOL-ODE (of various models) are at least competitive with standard integration schemes, especially for lower and moderate accuracy demands. For some models, especially Model 2, they are substantially more efficient. Lower or up to moderate accuracy demands (error in the range of 10^{-2} to 10^{-4}) are usually sufficient in biomathematical simulations of TDR systems.

We have followed the Method of Lines (MOL) approach to numerically find the solution of a TDR system. The finite volume approach used for the spatial discretization naturally respects the conservation of mass property of the TDR system. Special attention has been paid to select a spatial discretization which results in a MOL-ODE with nonnegative (analytic) solution. The careful discretization of the taxis term is especially important with respect to this. For this reason we have employed an upwinding technique in combination with limiter functions. The order of the approximation is two in general and results in an acceptable spatial error on fairly coarse grids already.

The second main part of this work is concerned with the solution of the initial value problem for the MOL-ODE. For this purpose we have employed two splitting techniques: approximate matrix factorization (AMF) and operator (Strang-) splitting (OPS). These splitting techniques are based on linearly-implicit W-methods and explicit Runge-Kutta methods. We have studied positivity properties of these methods and identified the optimal method RK32. A corresponding class of W-methods has been constructed. The resulting splitting schemes are of order two and from the variety of methods tested we especially recommend the schemes ROS2(γ_-)-AMF, ROS32(0.335)-AMF, OPS-ROS2(γ_-)-RK32, and OPS-ROS2(γ_-)-RKF2(3).

Maintenance of positivity of the solution during the solution process has been a major point in this work. Its significance has been clearly exemplified by the numerical experiments with Model 2. This example shows that a positive semi-discretization is not always sufficient for a successful simulation. Also for the time integration a suitable scheme with good positivity properties, e.g. the

recommended AMF and OPS schemes, should be used. This is especially important if zero is an unstable fixed point of the reaction term of the model (see the logistic growth term in Model 2). Then already slightly negative values in the numerical solution can lead to blow-up.

Altogether, the discretization of the spatial derivatives and the recommended splitting methods have shown to be very suitable for the simulation of TDR systems. The computer codes developed (see Sec. A.3) work reliably and robustly and are recommended for these simulations.

Looking ahead, interesting directions for further research in the field of numerical methods for the simulation of TDR systems are:

- The solution of linear systems with the AMF technique could be replaced by an iterative solution process: the coding of the spatial discretization would be simpler because a special order of the components in the MOL-ODE would no longer be necessary for efficiency reasons and more difficult geometries and non-regular meshes could be handled much easier.

- We have demonstrated the applicability of our approach for the simulation of TDR systems in two spatial dimensions. Going to three spatial dimensions will raise the question of parallelization of the scheme and also how the results can be suitably visualized. In [15] we have presented the first promising experiments with a parallel ODE solver.

- In all our models the organisms are assumed to be spread out continuously in space (i.e. represented by a continuous density function). However, in real life these organisms are discrete objects and for some modelling purposes (e.g. proliferation, loop formation during angiogenesis) it is more suitable to regard them as such. It would be a goal of future research to extend the methodology presented here so that it can be applied to models which include also discrete objects.

Appendix A

A.1 Solution of a first-order hyperbolic PDE related to Model 1

Let $4R > 0$ be an integer. Here we derive the analytical solution $u(t,r)$ of the following problem:

$$\partial_t u(t,r) + v(r)\partial_r u(t,r) = -\left(\frac{v(r)}{r} + w(r)\right) u \ , \ \text{for } t > 0, \ r \in (0,R) \ , \tag{A.1a}$$

$$u(0,r) = u_0(r) \ \text{ for } r \in [0,R] \ , \ u_0 \in C^1, \quad \text{and} \quad v(r) := 4\pi \sin(4\pi r) \ , w(r) := v'(r) \ . \tag{A.1b}$$

This problem arises in the derivation of the analytical solution of Model 1 in Section 2.3. The first order hyperbolic equation (A.1) can be solved by the Method of Characteristics. The initial curve is given by $\Gamma = \{(0, s, u_0(s)) \mid s \in [0,R]\}$, and the characteristic ODE system (with parameter s) by

$$\begin{aligned} &\Psi_1'(t) = 1 \ , && \Psi_1(0) = 0 \ , \\ &\Psi_2'(t) = v(\Psi_2(t)) \ , && \Psi_2(0) = s \ , \\ &\Psi_3'(t) = -\left(\frac{v(\Psi_2(t))}{\Psi_2(t)} + w(\Psi_2(t))\right)\Psi_3(t) \ , && \Psi_3(0) = u_0(s) \ . \end{aligned}$$

We obtain $\Psi_1(t;s) = t$, and, by using $v(\Psi_2)/\Psi_2 = \Psi_2'/\Psi_2$ and $w(\Psi_2) = \Psi_2''/\Psi_2'$,

$$\begin{aligned} \Psi_3(t;s) &= u_0(s) \exp\left(-\int_0^t \frac{\mathrm{d}}{\mathrm{d}\,\tau} \ln(\Psi_2(\tau)) + \frac{\mathrm{d}}{\mathrm{d}\,\tau} \ln(\Psi_2'(\tau)) \, d\tau\right) \\ &= u_0(s) \frac{s}{\Psi_2(t;s)} \ \cdot \ \frac{v(s)}{v(\Psi_2(t;s))} \ . \end{aligned}$$

We will give an expression for $\Psi_2(t;s)$ later.
Now, by the Method of Characteristics, we have $u(\Psi_1(t;s), \Psi_2(t;s)) = \Psi_3(t;s)$. Let $r = \Psi_2(t,s)$ and assume that we can solve this equation for s, i.e. $s = s(t,r)$, see Eq. (A.6). Then we can write down the solution of (A.1):

$$u(t,r) = u(t, \Psi_2(t,s)) = \Psi_3(t, s(t,r)) = u_0(s(t,r)) \frac{s(t,r)}{r} \ \cdot \ \frac{sin(4\pi s(t,r))}{sin(4\pi r)} \ . \tag{A.3}$$

We now derive the solution Ψ_2 of the second characteristic equation and the expression $s = s(t,r)$. Separation of variables formally leads to (let $C \in \mathbb{C}$ be an arbitrary constant)

$$\int \frac{1}{4\pi \sin(4\pi \Psi_2)} d\Psi_2 = \int dt, \quad \text{or} \quad \frac{1}{16\pi^2} \ln(\tan(2\pi \Psi_2)) = t + C \ . \tag{A.4}$$

Let $k = 0, 1, 2, \ldots$. We distinguish two cases: (*Case 1*) considers $\Psi_2 \in \left(\frac{k}{2}, \frac{k}{2} + \frac{1}{4}\right)$, i.e. $2\pi\Psi_2 \in \left(k\pi, k\pi + \frac{\pi}{2}\right)$, and (*Case 2*) considers $\Psi_2 \in \left(\frac{1}{4} + \frac{k}{2}, \frac{k+1}{2}\right)$, i.e. $2\pi\Psi_2 \in \left(\frac{\pi}{2} + k\pi, (k+1)\pi\right)$.
(*Case 1*) Consider $\Psi_2 \in \left(\frac{k}{2}, \frac{k}{2} + \frac{1}{4}\right)$: It is sufficient to consider $C \in \mathbb{R}$. Equivalent transformations of (A.4) lead to

$$\Psi_2(t) = \frac{1}{2\pi} \arctan\left(\exp\left(16\pi^2 t + C\right)\right) + \frac{k}{2} \in \left(\frac{k}{2}, \frac{k}{2} + \frac{1}{4}\right) .$$

Requiring $\Psi_2(0) = s$ for $s \in \left(\frac{k}{2}, \frac{k}{2} + \frac{1}{4}\right)$ determines $C = \ln\left(\tan\left(2\pi\left(s - \frac{k}{2}\right)\right)\right)$ and hence we have computed $\Psi_2(t; s)$ for values of $s \in \left(\frac{k}{2}, \frac{k}{2} + \frac{1}{4}\right)$. Now let $r = \Psi_2(t; s) \in \left(\frac{k}{2}, \frac{k}{2} + \frac{1}{4}\right)$. Then we can solve for s and obtain

$$s = s(t, r) = \frac{1}{2\pi} \arctan\left(\frac{\tan\left(2\pi\left(r - \frac{k}{2}\right)\right)}{\exp(16\pi^2 t)}\right) + \frac{k}{2} \ , \text{ for } r \in \left(\frac{k}{2}, \frac{k}{2} + \frac{1}{4}\right) . \tag{A.5a}$$

(*Case 2*) Consider $\Psi_2 \in \left(\frac{1}{4} + \frac{k}{2}, \frac{k+1}{2}\right)$: It is sufficient to consider $C = i\pi + \widetilde{C}$, $\widetilde{C} \in \mathbb{R}$. Then equivalent transformations of (A.4) lead to

$$\Psi_2(t) = \frac{1}{2\pi} \arctan\left(-\exp\left(16\pi^2 t + \widetilde{C}\right)\right) + \frac{k+1}{2} \in \left(\frac{k+1}{2} - \frac{1}{4}, \frac{k+1}{2}\right) .$$

Requiring $\Psi_2(0) = s$ for $s \in \left(\frac{1}{4} + \frac{k}{2}, \frac{k+1}{2}\right)$ determines $\widetilde{C} = \ln\left(-\tan\left(\pi(2s - (k+1))\right)\right)$ and hence we have computed $\Psi_2(t; s)$ for values of $s \in \left(\frac{1}{4} + \frac{k}{2}, \frac{k+1}{2}\right)$. Now let $r = \Psi_2(t; s) \in \left(\frac{1}{4} + \frac{k}{2}, \frac{k+1}{2}\right)$. Then we can solve for s and obtain

$$s = s(t, r) = \frac{1}{2\pi} \arctan\left(\frac{\tan\left(2\pi\left(r - \frac{k+1}{2}\right)\right)}{\exp(16\pi^2 t)}\right) + \frac{k+1}{2} \ , \text{ for } r \in \left(\frac{1}{4} + \frac{k}{2}, \frac{k+1}{2}\right) . \tag{A.5b}$$

The equations (A.5a) and (A.5b) can be simplified and written as one equation,

$$s = s(t, r) = \frac{1}{2\pi} \arctan\left(\frac{\tan(2\pi r)}{\exp(16\pi^2 t)}\right) + \frac{\mathrm{int}(4r) + (\mathrm{int}(4r) \mod 2)}{4} \ , \text{ for } r > 0, \ r \neq \frac{1}{4}k \ , \tag{A.6}$$

where $\mathrm{int} z$ is the integer part of $z \in \mathbb{R}_{+,0}$. This completes the derivation of the solution of (A.1) for $r \neq \frac{1}{4}k$. For $r = \frac{1}{4}k$ we obtain the solution by a limiting procedure $r \to \frac{1}{4}k$. For $k = 0, 1, 2$ we obtain

$$u(t, 0) = u_0(0) \exp\left(-16\pi^2 t\right) \ , \ u\left(t, \frac{1}{4}\right) = u_0\left(\frac{1}{4}\right) \exp\left(16\pi^2 t\right) \ , \ u\left(t, \frac{1}{2}\right) = u_0\left(\frac{1}{2}\right) \exp\left(-16\pi^2 t\right) .$$

A.2 Matrix functions — definition and properties

Definition 9 *[18, p. 381] Suppose $f(z),\ z \in \mathbb{C}$, is analytic inside and on a closed contour Γ which encircles the spectrum of a matrix $A \in \mathbb{R}^{m,m}$. We define $f(A)$ to be the matrix $f(A) = \frac{1}{2\pi i} \int_\Gamma f(z)(zI - A)^{-1} dz$ on an element-by-element basis.*

Lemma 20 *[18, p. 390] If $f(z),\ z \in \mathbb{C}$, has a power series representation $f(z) = \sum_{k=0}^{\infty} c_k z^k$ on an open disk containing the eigenvalues of the matrix $A \in \mathbb{R}^{m,m}$ then $f(A) = \sum_{k=0}^{\infty} c_k A^k$.*

Lemma 21 *Let the function $f(z),\ z \in \mathbb{C}$, be analytic in $-\mu$ for given $\mu \in \mathbb{R}_{+,0}$ and denote with R the radius of convergence of the Taylor series expansion of f around $-\mu$. For a given matrix $A \in \mathbb{R}^{m,m}$ define $B := \mu I + A$ and assume that the spectral radius $\rho(B)$ of B satisfies $\rho(B) < R$. Then we have*

$$f(A) = \sum_{k=0}^{\infty} \frac{f^{(k)}(-\mu)}{k!} B^k.$$

Proof f is analytic in $-\mu$ and this implies that the radius of convergence of the Taylor series expansion of f around $-\mu$ is greater than zero, $R > 0$, and that $f(z-\mu) = \sum_{k=0}^{\infty} \frac{f^{(k)}(-\mu)}{k!} z^k$ for all $|z| < R$.
We define $g(z) := f(z-\mu)$. Hence g has a power series expansion in the open disk $|z| < R$ which contains, by assumption, the spectrum of B. Therefore, by Lemma 20, holds

$$g(B) = \sum_{k=0}^{\infty} \frac{f^{(k)}(-\mu)}{k!} B^k.$$

It remains to show that $f(A) = g(B)$. Let Γ_A be the closed contour defined by $\Gamma_A = S_2(-\mu, (R + \rho(B))/2)$. Then f is analytic inside of and on Γ_A and the spectrum of A is inside of Γ_A. Let $\Gamma_B := S_2(0, (R + \rho(B))/2)$. Then the same statements hold with respect to the function g and the matrix B. By definition of matrix functions and substitution in the integral ($z \to \widetilde{z} - \mu \Rightarrow dz \to d\widetilde{z}, \Gamma_A \to \Gamma_B$) we obtain

$$\begin{aligned} f(A) &= \frac{1}{2\pi i} \int_{\Gamma_A} f(z)(zI - A)^{-1} dz \\ &= \frac{1}{2\pi i} \int_{\Gamma_B} f(\widetilde{z} - \mu)(\widetilde{z} I - (\mu I + A))^{-1} d\widetilde{z} \\ &= \frac{1}{2\pi i} \int_{\Gamma_B} g(\widetilde{z})(\widetilde{z} I - B)^{-1} d\widetilde{z} = g(B)\,. \end{aligned}$$

□

A.3 Computer programs

The programs written in preparation of this thesis are available from the author. The source files are documented such that users with some experience in FORTRAN77 can test the programs and modify them to their needs. Here we only give a short description on what is available.
We have implemented the semi-discretization of TDR systems in a collection of FORTRAN77 subroutines. These subroutines are subdivided in a set of model independent and a set of model dependent routines. The model dependent subroutines provide details about the initial data, the parameters, the taxis and reaction functions, and the boundary conditions of a specific TDR model. These subroutines can easily be modified by a user with some FORTRAN77 knowledge so that new models can be implemented.
The splitting schemes (AMF and OPS) for the solution of the MOL-ODE are also implemented in FORTRAN77. Their calling sequence follows the quasi-standard which is generally adopted in coding ODE integration methods. Hence they can also be used for the solution of ODE systems which do not specifically arise from semi-discretizations of TDR systems.
A command-line based simulation environment written in FORTRAN77 exists. This couples the semi-discretization with the time stepping schemes. The system allows to run simulations and comparisons easily by issuing a few commands or running a script file. Intermediate and final solutions can be saved in a format readable by Matlab. Some `.m` files are provided which can be used to read the output files of the simulation environment into Matlab and prepare the visualization of the data (for one and two spatial dimensions).

Bibliography

[1] H. Amann. *Gewöhnliche Differentialgleichungen*. de Gruyter Berlin–New York, 2. edition, 1995.

[2] A. R. A. Anderson and M. A. J. Chaplain. Continuous and Discrete Mathematical Models of Tumor-induced Angiogenesis. *Bull. Math. Biol.*, 60(5):857–899, 1998.

[3] A. R. A. Anderson, M. A. J. Chaplain, E. L. Newman, R. J. C. Steele, and A. M. Thompson. Mathematical modelling of tumour invasion and metastasis. *J. Theoret. Med.*, 2:129–154, 2000.

[4] L. Angermann. An introduction to finite volume methods for linear elliptic problems of second order. Bericht nr. 164, Institut für Angewandte Mathematik, Universität Erlangen-Nürnberg, February 1995.

[5] R. M. Beam and R. F. Warming. An implicit finite-difference algorithm for hyperbolic systems in conservation-law form. *J. Comput. Phys.*, 22:87–110, 1976.

[6] C. Bolley and M. Crouzeix. Conservation de la positivité lors de la discrétisation des problèmes d'évolution paraboliques. *RAIRO Anal. Numer.*, 12(3):237–245, 1978.

[7] G. D. Byrne. Pragmatic experiments with Krylov methods in the stiff ODE setting. In J. R. Cash and I. Gladwell, editors, *Computational Ordinary Differential Equations*, pages 323–356. Oxford University Press, New York, 1992.

[8] M. A. J. Chaplain and A. M. Stuart. A model mechanism for the chemotactic response of endothelial cells to tumour angiogenesis factor. *IMA J. Math. Appl. Med. Biol.*, 10:149–168, 1993.

[9] G. Dahlquist and R. Jeltsch. Shifted Runge–Kutta methods and transplanted differential equations. In K. Strehmel, editor, *Numerical Treatment of Differential Equations, Proceedings, Halle 1987*, pages 47–56. Teubner, Leipzig, 1988.

[10] K. Dekker and J. G. Verwer. *Stability of Runge-Kutta methods for stiff nonlinear differential equations*, volume 2 of *CWI monograph*. Elsevier–North Holland, Amsterdam, 1984.

[11] E. G. D'yakonov. Difference systems of second order accuracy with a divided operator for parabolic equations without mixed derivatives. *USSR Comput. Math. Math. Phys.*, 4(5):206–216, 1964.

[12] C. Eichler-Liebenow, N. H. Cong, R. Weiner, and K. Strehmel. Linearly implicit splitting methods for higher space-dimensional parabolic differential equations. *Appl. Numer. Math.*, 28(2–4):259–274, 1998.

[13] A. Gerisch. Finite difference methods for coupled nonlinear hyperbolic and parabolic partial differential equations in one and two dimensions. Master's thesis, University of Dundee, 1997.

[14] A. Gerisch, D. F. Griffiths, R. Weiner, and M. A. J. Chaplain. A positive splitting method for mixed hyperbolic-parabolic systems. *Numer. Methods Partial Differential Eq.*, 17:152–168, 2001.

[15] A. Gerisch and H. Podhaisky. Splitting methods for the simulation of tumor angiogenesis models. In *Proceedings of the 16th IMACS World Congress, Lausanne, Switzerland*. published on CD-ROM, 2000.

[16] A. Gerisch and J. G. Verwer. Operator splitting and approximate matrix factorization for taxis–diffusion–reaction models. Technical Report MAS-R0026, CWI (Amsterdam), 2000.

[17] A. Gerisch and R. Weiner. On the positivity of low order explicit Runge-Kutta schemes applied in splitting methods. *Computers and Mathematics with Applications*, in press, 2001.

[18] G. H. Golub and C. F. Van Loan. *Matrix Computations*. North Oxford Academic, 1986.

[19] J. A. van de Griend and J. F. B. M. Kraaijevanger. Absolute monotonicity of rational functions occurring in the numerical solution of initial value problems. *Numer. Math.*, 49:413 – 424, 1986.

[20] B. Gustafsson, H.-O. Kreiss, and J. Oliger. *Time dependent problems and difference methods*. John Wiley & Sons, Inc., 1995.

[21] Hackbusch. *Iterative Lösung großer schwachbesetzter Gleichungssysteme*. Number 69 in Leitfäden der angewandten Mathematik und Mechanik. Teubner Stuttgart, 1993.

[22] E. Hairer, S. P. Nørsett, and G. Wanner. *Solving ordinary differential equations. I. Nonstiff problems*. Number 8 in Springer Series in Computational Mathematics. Springer–Verlag, Berlin, 2nd edition, 1993.

[23] E. Hairer and G. Wanner. *Solving ordinary differential equations. II. Stiff and differentgial–algebraic problems*. Number 14 in Springer Series in Computational Mathematics. Springer–Verlag, Berlin, 2nd edition, 1996.

[24] D. J. Higham and L. N. Trefethen. Stiffness of ODEs. *BIT*, 33:285–303, 1993.

[25] Thomas Höfer. *Modelling Dictyostelium Aggregation*. PhD thesis, University of Oxford, 1996.

[26] Thomas Höfer, Jonathan A. Sherratt, and Philip K. Maini. Cellular pattern formation during Dictyostelium aggregation. *Physica D*, 85:425–444, 1995.

[27] Z. Horváth. Positivity of Runge–Kutta and diagonally split Runge–Kutta methods. *Appl. Numer. Math.*, 28(2–4):309–326, 1998.

[28] P. J. van der Houwen and B. P. Sommeijer. Approximate factorization for time-dependent partial differential equations. Report MAS-R9915, CWI, 1999.

[29] W. Hundsdorfer. Partial implicit BDF2 blends for convection dominated flows. Technical Report MAS-R9831, CWI, 1998.

[30] W. Hundsdorfer. Trapezoidal and midpoint splittings for initial-boundary value problems. *Math. Comp.*, 67(223):1047–1062, 1998.

[31] W. Hundsdorfer, B. Koren, M. van Loon, and J. G. Verwer. A positive finite-difference advection scheme. *J. Comput. Phys.*, 117(1):35–46, 1995.

[32] W. H. Hundsdorfer. Accuracy and stability of splitting with stabilizing corrections. Technical Report MAS-R9935, CWI, 1999.

[33] O. Knoth and R. Wolke. Implicit–explicit Runge–Kutta methods for computing atmospheric reactive flows. *Appl. Numer. Math.*, 28:327–341, 1998.

[34] A. L. Koch, A. Carr, and D. W. Ehrenfeld. The problem of open-sea navigation: The migration of the green turtle to ascension island. *J. Theor. Biol.*, 22:163–179, 1969.

[35] B. Koren. A robust upwind discretization method for advection, diffusion and source terms. In C. B. Vreugdenhill and B. Koren, editors, *Numerical Methods for Advection–Diffusion Problems*, volume 45 of *Notes on Numerical Fluid Mechanics*, chapter 5, pages 117–138. Vieweg, Braunschweig, 1993.

[36] J. F. B. M. Kraaijevanger. Absolute monotonicity of polynomials occurring in the numerical solution of initial value problems. *Numer. Math.*, 48(3):303–322, 1986.

[37] J. F. B. M. Kraaijevanger. Contractivity of Runge-Kutta methods. *BIT*, 31(3):482–528, 1991.

[38] D. Lanser, J. G. Blom, and J. G. Verwer. Time integration of the shallow water equations in spherical geometry. Technical Report MAS-R0021, CWI, 2000.

[39] D. Lanser and J. G. Verwer. Analysis of operator splitting for advection-diffusion-reaction problems from air pollution modelling. *J. Comput. Appl. Math.*, 111(1-2):201–216, 1999.

[40] R. J. LeVeque. CLAWPACK User Guide and Software. Available from `http://www.amath.washington.edu/~claw`.

[41] R. J. LeVeque. *Numerical Methods for Conservation Laws*. Lectures in Mathematics. Birkhäuser, 2nd edition, 1992.

[42] H. A. Levine, B. D. Sleeman, and M. Nilsen-Hamilton. Mathematical modeling of the onset of capillary formation initiating angiogenesis. *J. Math. Biol.*, 42:195–238, 2001.

[43] C. R. MacCluer. The many proofs and applications of Perrons's theorem. *SIAM Rev.*, 42(3):487–498, 2000.

[44] J. D. Murray. *Mathematical biology*. Springer–Verlag Berlin, 2nd edition, 1993.

[45] M. E. Orme and M. A. J. Chaplain. A mathematical model of the first steps of tumour-related angiogenesis: Capillary sprout formation and secondary branching. *IMA J. Math. Appl. Med. Biol.*, 13:73–98, 1996.

[46] Hans G. Othmer and Angela Stevens. Aggregation, blowup, and collapse: the ABC's of taxis in reinforced random walks. *SIAM J. Appl. Math.*, 57:1044–1081, 1997.

[47] K. J. Painter, P. K. Maini, and H. G. Othmer. Development and application of a model for cellular response to multiple chemotactic cues. *J. Math Biol.*, 41:285–314, 2000.

[48] D. W. Peaceman and Jr. H. H. Rachford. The numerical solution of parabolic and elliptic differential equations. *J. Soc. Indust. Appl. Math.*, 3:28–41, 1955.

[49] Y. Saad. *Iterative methods for sparse linear systems*. PWS Publishing Company, 1995.

[50] B. A. Schmitt and R. Weiner. Matrix-free W-methods using a multiple Arnoldi iteration. *Appl. Numer. Math.*, 18:307–320, 1995.

[51] L. F. Shampine. *Numerical Solution of Ordinary Differential Equations*. Chapman & Hall, New York, 1994.

[52] Chi-Wang Shu and Stanley Osher. Efficient Implementation of Essentially Non-oscillatory Shock-Capturing Schemes. *J. Comput. Phys.*, 77(2):439–471, 1988.

[53] T. Steihaug and A. Wolfbrandt. An attempt to avoid exact jacobian and nonlinear equations in the numerical solution of stiff differential equations. *Math. Comp.*, 33:521–534, 1979.

[54] G. Strang. On the construction and comparison of difference schemes. *SIAM J. Numer. Anal.*, 5:506–517, 1968.

[55] K. Strehmel and R. Weiner. *Linear-implizite Runge-Kutta-Methoden und ihre Anwendung*. Teubner, Leipzig, 1992.

[56] K. Strehmel and R. Weiner. *Numerik gewöhnlicher Differentialgleichungen*. B. G. Teubner, Stuttgart, 1995.

[57] P. K. Sweby. High resolution schemes using flux limiters for hyperbolic conservation laws. *SIAM J. Numer. Anal.*, 21:995–1011, 1984.

[58] R. Tyson, L. G. Stern, and R. J. LeVeque. Fractional step methods applied to a chemotaxis model. *J. Math. Biol.*, 41:455–475, 2000.

[59] Rebecca Tyson, S. R. Lubkin, and J. D. Murray. Model and analysis of chemotactic bacterial patterns in a liquid medium. *J. Math. Biol.*, 38:359–375, 1999.

[60] J. G. Verwer. S-stability properties for generalized Runge-Kutta methods. *Numer. Math.*, 27(359–370), 1977.

[61] J. G. Verwer, W. H. Hundsdorfer, and J. G. Blom. Numerical time integration for air pollution models. Report MAS-R9825, CWI, 1998.

[62] J. G. Verwer, E. J. Spee, J. G. Blom, and W. Hundsdorfer. A second-order Rosenbrock method applied to photochemical dispersion problems. *SIAM J. Sci. Comput.*, 20(4):1456–1480, 1999.

[63] R. Weiner and B. A. Schmitt. Order Results for Krylov-W-Methods. *Computing*, 61:69–89, 1998.

[64] R. Weiner, B. A. Schmitt, and H. Podhaisky. ROWMAP—a ROW-code with Krylov techniques for large stiff ODEs. *Appl. Numer. Math.*, 25:303–319, 1997.

Angaben zur Person und zum Bildungsgang

Persönliche Daten

Alf Gerisch

Geschwister-Scholl-Straße 2
06231 Bad Dürrenberg

geboren: am 2. September 1973 in Merseburg

Familienstand: ledig

Bildungsgang

1980 – 1990	Besuch der Polytechnischen Oberschule in Bad Dürrenberg.
1990 – 1991	Besuch der Spezialklassen für Mathematik und Physik der Martin-Luther-Universität Halle-Wittenberg (aufgelöst 1991).
1991 – 1992	Besuch des Landesgymnasiums Georg Cantor mathematisch-naturwissenschaftlich-technischer Richtung in Halle, Abitur.
1992 – 1993	Grundwehrdienst.
1993 – 1998	Studium im Diplomstudiengang Mathematik (Nebenfach Informatik) an der Martin-Luther-Universität Halle-Wittenberg, Diplom.
1996 – 1997	Studium an der University of Dundee (Schottland) in „Numerical Analysis and Programming“, Master of Science.
1998 – 2001	Promotionsstudent am Institut für Numerische Mathematik der Martin-Luther-Universität Halle-Wittenberg unterstützt durch ein Stipendium im Rahmen des von der DFG geförderten Graduiertenkollegs „Transport von Wirkstoffen in biologischen Systemen“ an der Martin-Luther-Universität Halle-Wittenberg.

Erklärung

Hiermit erkläre ich, Alf Gerisch, an Eides statt, daß ich die vorliegende Arbeit selbständig, ohne unzulässige Hilfe Dritter und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel angefertigt habe. Die aus anderen Werken wörtlich oder inhaltlich entnommenen Daten, Fakten und Konzepte sind unter Angabe der entsprechenden Quelle als solche gekennzeichnet.

Diese Arbeit wurde bisher weder im In- noch im Ausland in gleicher oder ähnlicher Form in einem anderen Prüfungsverfahren vorgelegt.

Halle (Saale), im Mai 2001

Alf Gerisch