Simulation of chemical systems with fast chemistry

Simulation von chemischen Reaktionssystemen mit schnellen chemischen Reaktionen
Dem Fachbereich Mathematik der Technischen Universität Darmstadt
zur Erlangung des Grades eines Doktors der Naturwissenschaften (Dr. rer. nat.)
genehmigte Dissertation von M.Sc. Axel Ariaan Lukassen aus Aachen
Tag der Einreichung: 07.08.2018, Tag der Prüfung: 23.10.2018
Darmstadt 2018 — D 17

1. Gutachten: Prof. Dr. Martin Kiehl
2. Gutachten: Prof. Dr. Jens Lang

TECHNISCHE UNIVERSITÄT DARMSTADT

Fachbereich Mathematik
AG Numerik und
wissenschaftliches Rechnen

Simulation of chemical systems with fast chemistry
Simulation von chemischen Reaktionssystemen mit schnellen chemischen Reaktionen

Genehmigte Dissertation von M.Sc. Axel Ariaan Lukassen aus Aachen

1. Gutachten: Prof. Dr. Martin Kiehl
2. Gutachten: Prof. Dr. Jens Lang

Tag der Einreichung: 07.08.2018
Tag der Prüfung: 23.10.2018

Darmstadt 2018 — D 17

Abstract

This PhD thesis deals with the numerical simulation of chemical reaction systems. Chemical reaction systems occur in many different areas of our lives. In some areas, such as biology, scientists try to analyze the occurring chemical reactions for better understanding of the underlying mechanisms. In other areas, such as the chemical industry, the focus is on optimization of reactor design in order to increase the productivity. All cases have in common that the temporal development of the chemical reaction systems is described by differential equations. The derivation of the corresponding differential equations is given in chapter 2. The resulting differential equations are generally large nonlinear systems. Therefore, analytical solutions are usually not available, and an approximation of the solution is calculated by numerical integration methods on a computer. The size and the stiffness of the corresponding differential equations often lead to an unfavorably long computing time or an exceeding of the available memory space. Hence, many scientists have developed reduction mechanisms for chemical reaction systems. These reduction mechanisms exploit that there are very slow as well as very fast processes in chemical reaction systems. Depending on the discretization, the timescales of the fast processes are much smaller than the used time step of the numerical integration method. Thus, the fast processes are approximated by their partial equilibrium. The relaxation assumption results in algebraic equations, which can be used to reduce the dimension of the differential equation. Furthermore, the reduced differential equations are often not or less stiff. A description of the partial equilibrium assumption of single reactions or the quasi-stationary state assumption of individual chemical species is given in chapter 3. In addition, frequently used automatic reduction mechanisms of different authors are summarized in the corresponding chapter. Despite the great popularity of reduction mechanisms, the reduction mechanisms listed in chapter 3 are not applicable for all chemical reaction systems. Most of them are based on the existence of a low-dimensional manifold which is approached by the state of the reaction system in a fraction of the considered time step. However, such a low-dimensional manifold does not always exist. Furthermore, the choice of fast processes depends on time and space. Therefore, the dimension of the reduced differential equation can vary. This leads to additional problems. In order to reduce the stiffness of the differential equation, a new approach is introduced in chapter 4. In opposition to many other methods, this new approach is also applicable if the relaxation of the fast processes does not restrict the state of the system onto a low-dimensional manifold, and if the number of fast processes changes in time and space. Reduction of stiffness can decrease the necessary computing time of the numerical solver. The computing time is particularly important in parameter identification. The rate of each chemical reaction is described by at least one parameter. In order to determine unknown parameters, the solution of the differential equation has to be computed for many different parameter sets. Hence, for parameter estimation it is desirable to reduce the computing time of each solution. Since the algebraic equations from chapter 3 also depend on the unknown parameters, it is not possible to replace the differential equation for all possible parameter sets by one reduced differential equation. However, the additional information from the partial equilibrium assumption or the quasi-stationary state assumption can be used to calculate some unknown parameters as a function of all other parameters. This reduces the dimension of the parameter space, and decreases the computing time of parameter identification. The procedure is described in chapter 5. For the sake of simplicity, ordinary differential equations are considered in the chapters 3 to 5. However, many chemical reaction systems that are not homogeneous in space are described by partial differential equations. In the case of splitting methods, a sequence of partial differential equations for the transport and ordinary differential equations for the chemical reaction steps is solved. Thus, if splitting methods are used, a homogeneous chemical reactor is considered for each spatial node in the chemical reaction step. Therefore, the results of the chapters 3 to 5 are applicable. However, an additional splitting error is introduced. Popular splitting methods are the first order Lie-Trotter splitting and the second order Strang splitting. Stiffness of the considered differential equation results in order reduction for the Strang splitting scheme. Therefore, Strang splitting is only a first order scheme for stiff chemical reaction systems. However, the extrapolated Lie-Trotter splitting is a second order scheme for stiff chemical reaction

systems. The analysis of the extrapolated Lie-Trotter splitting for chemical reaction systems is presented in chapter 6.

Zusammenfassung

Die vorliegende Promotion behandelt die numerische Simulation von chemischen Reaktionssystemen. Chemische Reaktionssysteme treten in vielen verschiedenen Bereichen unseres Lebens auf. In manchen Bereichen wie der Biologie versuchen Wissenschaftler die auftretenden chemischen Reaktionen zu analysieren, um sie besser zu verstehen. In anderen Bereichen wie der chemischen Industrie liegt der Fokus oft auf der Optimierung, im Sinne der Produktivität oder der Sicherheit. Alle Fälle haben gemeinsam, dass die zeitliche Entwicklung der chemischen Reaktionssysteme durch Differentialgleichungen beschrieben wird. Die Herleitung der entsprechenden Differentialgleichungen ist in Kapitel 2 gegeben. Die auftretenden Differentialgleichungen sind im Allgemeinen große nichtlineare Systeme. Deshalb sind analytische Lösungen meistens nicht verfügbar und eine Näherungslösung wird am Computer durch numerische Integrationsmethoden berechnet. Dabei führen die Größe und die Steifheit der betreffenden Differentialgleichungen häufig zu einer unvorteilhaft langen Rechenzeit oder einem Überschreiten des verfügbaren Speicherplatzes. Daher haben zahlreiche Wissenschaftler Reduktionsmechanismen für chemische Reaktionssysteme entwickelt. Dabei wird ausgenutzt, dass es in chemischen Reaktionssystemen sowohl sehr langsame als auch sehr schnelle Prozesse gibt. Abhängig von der Diskretisierung können die Zeitskalen der schnellen Prozesse dabei die verwendete Zeitschrittweite deutlich unterschreiten. Daher werden die schnellen Prozesse durch ihr Gleichgewicht approximiert. Es ergeben sich algebraische Gleichungen, welche den Zustand des Systems beschreiben. Diese Gleichungen können verwendet werden, um die Dimension der Differentialgleichung zu reduzieren. Desweiteren sind die reduzierten Differentialgleichungen in vielen Fällen weniger steif. Eine Beschreibung der Annahme des partiellen Gleichgewichts einzelner Reaktionen oder des quasi-stationären Zustands einzelner chemischer Spezies befindet sich in Kapitel 3. Außerdem werden in dem betreffenden Kapitel häufig verwendete automatische Reduktionsmechanismen verschiedener Autoren zusammengefasst. Trotz der großen Beliebtheit von Reduktionsmechanismen gibt es einige Szenarien, in denen die in Kapitel 3 gelisteten Reduktionsmechanismen nicht anwendbar sind. So basieren sie zum größten Teil auf der Existenz einer niedrig-dimensionalen Mannigfaltigkeit, welche das Gleichgewicht der schnellen Prozesse beschreibt. Die Mannigfaltigkeit ist jedoch nicht immer niedrig-dimensional. Desweiteren hängt die Auswahl an schnellen Prozessen manchmal von Zeit und Ort ab. Deshalb kann die Dimension der reduzierten Differentialgleichung variieren. Dies führt zu zusätzlichen Problemen. Um trotzdem die Steifheit der betrachteten Differentialgleichung zu reduzieren, wird in Kapitel 4 ein neuer Ansatz eingeführt. Dieser Ansatz ist auch anwendbar, falls der Zustand des chemischen Reaktionssystems nicht durch eine Approximation auf einer niedrig-dimensionalen Mannigfaltigkeit angenähert werden kann oder die Anzahl der schnellen Prozesse in Zeit und Ort variabel ist. Dies wird erreicht, indem die Geschwindigkeit der schnellen Prozesse reduziert wird. Dieses Vorgehen unterscheidet sich somit grundlegend von vielen anderen Methoden, die durch die sofortige Annahme des Gleichgewichts aller schneller Prozesse die betreffende Geschwindigkeiten auf unendlich erhöhen. Durch die Reduktion der Steifheit kann in vielen Fällen das Lösen der Differentialgleichungen beschleunigt werden. Ein Bereich, in dem die Rechenzeit besonders wichtig ist, ist die Parameteridentifizierung. Die Geschwindigkeit jeder chemischen Reaktion wird durch mindestens einen Parameter beschrieben. Um unbekannte Parameter zu bestimmen, ist im Allgemeinen die Lösung der Differentialgleichung für viele verschiedene Parametersätze nötig. Für die Parameteridentifizierung ist es daher wünschenswert die Rechenzeit der numerischen Integration zu reduzieren. Da die algebraischen Gleichungen aus Kapitel 3 ebenfalls von den unbekannten Parametern abhängen, ist es nicht möglich die Differentialgleichung für alle möglichen Parametersätze durch eine reduzierte Differentialgleichung zu ersetzen. Allerdings können die zusätzlichen Informationen aus der Annahme eines partiellen Gleichgewichts oder eines quasi-stationären Zustands benutzt werden, um einige unbekannte Parameter vorab in Abhängigkeit der anderen Parameter zu berechnen. Dadurch wird der Aufwand der Parameteridentifizierung verringert. Das betreffende Vorgehen wird in Kapitel 5 beschrieben. Der Einfachheit halber werden in den Kapiteln 3 bis 5 gewöhnliche Differentialgleichungen betrachtet. Jedoch werden viele chemische Reaktionssysteme, die nicht homogen im Ort sind, durch

partielle Differentialgleichungen beschrieben. Splitting-Verfahren können verwendet werden, um die erzielten Ergebnisse auf partielle Differentialgleichungen übertragen zu können. Bei Splitting-Verfahren werden abwechselnd partielle Differentialgleichungen für den Transport und gewöhnliche Differentialgleichungen für die chemischen Reaktionsterme betrachtet. Allerdings wird ein zusätzlicher Splitting-Fehler eingeführt. Dieser hängt von dem Splittingzeitschritt ab. Beliebte Splitting Methoden sind das Lie-Trotter Splitting mit Ordnung 1 und das Strang Splitting mit Ordnung 2. Ordnungsreduktion kann jedoch im Falle des Strang Splittings durch die Steifheit der betrachteten Systeme auftreten. Im Gegensatz dazu hat das extrapolierte Lie-Trotter Splitting auch für steife Reaktionssysteme mit langsamem Transport Ordnung 2. Eine Analyse des extrapolierten Lie-Trotter Splittings für chemische Reaktionssysteme befindet sich in Kapitel 6.

Acknowledgements

First of all, I thank my supervisor Professor Doctor Martin Kiehl for his support in recent years. It was always possible to discuss current problems in research as well as teaching. Thereby numerous new impulses have emerged for my work.

I also thank the co-referee Professor Doctor Jens Lang. Furthermore, I thank the members of the examining board Professor Doctor Reinhard Farwig, Professor Doctor Steffen Roch, and Professor Doctor Stefan Ulbrich.

Moreover, I thank my family. Even in hard times my family motivated me to continue my work, and to achieve my goals.

The pleasant working atmosphere has been another important aspect in recent years. My colleagues always had time to discuss current research results. I have found not only knowledge, but also friends. Furthermore, I would like to mention the secretariat, which has handled all administrative duties.

Contents

List of Figures 1

List of Tables 3

Nomenclature 5

1 Introduction 9
1.1 Outline of the present work . . . 10

2 Modelling chemical reaction systems 13
2.1 Homogeneous chemical reaction system . . . 13
2.2 Advection-diffusion-reaction system . . . 15
2.3 Boundary conditions . . . 17

3 Reduction of chemical reaction systems 19
3.1 Partial equilibrium assumption (PEA) and quasi-steady state assumption (QSSA) . . . 19
3.2 Introduction of automatic reduction mechanisms . . . 25
3.3 Illustration of the low-dimensional manifold . . . 28
3.4 Low-dimensional manifold and transport processes . . . 29
3.5 Recommended reduction mechanisms for chemical reaction systems . . . 30
3.5.1 Intrinsic Low-Dimensional Manifold method . . . 30
3.5.2 Differential-Algebraic Equation and thermodynamics . . . 31
3.5.3 Computational Singular Perturbation method . . . 32
3.5.4 Flamelet-Generated Manifold method . . . 33
3.5.5 Reaction-Diffusion Manifold method . . . 33
3.5.6 Global Quasi-Linearisation method . . . 34
3.6 Validity of reduction mechanisms based on singular perturbation theory . . . 35

4 Reduction of stiffness-induced round-off errors 39
4.1 Introduction . . . 39
4.2 New method . . . 42
4.2.1 Motivation . . . 42
4.2.2 The modified problem . . . 43
4.3 Analysis . . . 45
4.3.1 Linear functions . . . 46
4.3.2 Symmetric linear functions . . . 48
4.3.3 Non-symmetric linear functions . . . 49
4.3.4 Nonlinear functions . . . 52
4.3.5 Choice of the parameter λ_g . . . 53
4.3.6 Computational costs . . . 54
4.4 Numerical test case . . . 54
4.4.1 Numerical Jacobian matrix . . . 55
4.4.2 Analytical Jacobian matrix . . . 57
4.5 Conclusions and summary . . . 58

5 Parameter identification for chemical reaction systems 59
5.1 Introduction . . . 59
5.2 Parameter-dependence of the low-dimensional manifold . . . 61

5.3 Thermodynamic description of partial equilibrium . . . 63
5.4 Exploiting different timescales for parameter identification . . . 67
5.4.1 Determining PEA and QSSA . . . 68
5.4.2 Using thermodynamic description for PEA . . . 69
5.4.3 Using QSSA . . . 71
5.5 Numerical test cases . . . 73
5.5.1 Fictitious linear system . . . 74
5.5.2 Zeldovich mechanism - QSSA . . . 75
5.5.3 Dissociation of dinitrogen pentoxide - PEA . . . 77
5.6 Conclusions and summary . . . 78

6 Operator splitting for stiff differential equations 79
6.1 Introduction . . . 79
6.2 Splitting methods . . . 81
6.3 Convergence order for (non-stiff) problems . . . 82
6.4 Splitting methods for stiff differential equations . . . 84
6.5 Stability of the extrapolated Lie-Trotter splitting . . . 91
6.6 Numerical examples . . . 99
6.6.1 Linear example . . . 99
6.6.2 Slow dimerisation . . . 101
6.6.3 Fast dimerisation . . . 102
6.6.4 Extremely stiff reaction system . . . 103
6.7 Conclusions and summary . . . 105

7 Summary and outlook 107
7.1 Summary . . . 107
7.2 Rise to future work . . . 108

References 111

Appendix 121

List of Figures

1 Temporal development of the concentration $[N_2O_5]$ . . . 24
2 Relative error for the QSSA of $[NO_3]$ . . . 24
3 Temporal development of the concentration for stiff chemical reaction systems . . . 28
4 Temporal development of the concentration for non-stiff chemical reaction systems . . . 29
5 Time step size for RADAU5 and the modified RADAU5 for a numerically computed Jacobian matrix . . . 56
6 Error for RADAU5 and the modified RADAU5 for a numerically computed Jacobian matrix 56
7 Logarithmic time step size for RADAU5 and the modified RADAU5 in case of an analytical Jacobian matrix . . . 58
8 Error for RADAU5 and the modified RADAU5 in case of an analytical Jacobian matrix . . . 58
9 Temporal development of the concentration $[N]$ . . . 76
10 Relative error for the QSSA of $[N]$ . . . 76
11 Splitting error $e_{(.)}(\Delta t)$ plotted against the splitting time step Δt for the stiffness parameter $\epsilon \approx 5 \cdot 10^{-4}$ . . . 101
12 Splitting error $e_{(.)}(\Delta t)$ plotted against the splitting time step Δt for the stiffness parameter $\epsilon \approx 5 \cdot 10^{-5}$ . . . 101
13 Splitting error $e_{(.)}(\Delta t)$ plotted against the splitting time step Δt for the slow dimerisation 102
14 Splitting error $e_{(.)}(\Delta t)$ plotted against the splitting time step Δt for the fast dimerisation . 102
15 Fast transient phase of the species H_2, O_2 and H_2O . . . 103
16 Time evolution of the species H_2, O_2 and H_2O . . . 103
17 Splitting error $e_{(.)}(\Delta t)$ plotted against splitting time step Δt for inflow (6.58) . . . 104
18 Splitting error $e_{(.)}(\Delta t)$ plotted against splitting time step Δt for inflow (6.59) . . . 104

List of Tables

1 Costs for the simulation of Hydrogen Oxygen Combustion in case of a numerically computed Jacobian matrix 57
2 Costs for the simulation of Hydrogen Oxygen Combustion in case of an analytical Jacobian matrix 57
3 Chemical potential of all species in Example 8 65
4 Average computational costs and average error of parameter identification of a fictitious example with different dimension of the parameter space 75
5 Average computational costs and average error of parameter identification of the Zeldovich mechanism with different dimension of the parameter space 77
6 Chemical potential of necessary species in the dissociation of dinitrogen pentoxide 78
7 Average error for different fixed parameter estimations of reaction rate constant k_2 78

Nomenclature

α	Thermal diffusivity
β_j	Reaction-dependent Arrhenius parameter of the jth reaction
ΔC_i	The Laplacian of the concentration of the ith species
$\Delta C_{(\mathrm{i})}$	The ith step of the Newton method
Δt	Time step of a splitting method
ϵ	Small parameter, which describes the stiffness of the system
ϵ_m	Machine epsilon
Λ	Matrix such that the entry Λ_{ij} describes the influence of the jth mode on the ith mode
λ_g	Threshold $\ll 0$ for eigenvalues
λ_i	The ith eigenvalue
μ_i	The chemical potential of the ith species
$\mu_i(T)$	The concentration-independent part of the chemical potential of the ith species
$\mu_p[\cdot]$	The logarithmic matrix norm that is associated to the matrix p-norm
Ω	Spatial domain
$\partial\Omega$	Boundary of the spatial domain
σ_i	The ith singular value of a given matrix
τ	Transformed time variable $\tau = t/\epsilon$
$\tilde{\tau}$	Local discretization error of a Runge-Kutta method
τ_i	Timescale of the ith reaction
θ	Local coordinates on the low-dimensional manifold
A_i	The ith chemical species
A_j^{Arr}	Reaction-dependent Arrhenius parameter of the jth reaction
C	Concentration vector $(C_1, ..., C_{m_s})^T$ of all chemical species
$\tilde{C}$	Solution of the modified (e.g. by a reduction mechanism) chemical reaction system
C_i	Concentration of the ith chemical species
C^i	Exact or approximated concentration $C(t_i)$
$C_{(i)}$	The ith iteration of the Newton method
c_{ij}	State dependent coefficient
C_{LT}	Approximation of C that is obtained by Lie-Trotter splitting
C^m	Measurement of the concentration vector of all chemical species

C_{RE} Approximation of C that is obtained by extrapolated Lie-Trotter splitting

C_{ref} Reference solution that is computed with a very small step size

C_S Approximation of C that is obtained by Strang splitting

C_0 Initial values of the ODE (2.4)

C_∞ Steady state of the ODE (2.4)

D Diagonal matrix

D_i Diffusion coefficient of the ith chemical species

e Error of the computed solution

$e_{\cdot j}$ Educts of the jth chemical reaction

$E_{A,j}$ Reaction-dependent Arrhenius parameter, which describes the activation energy

F Chemical source term, which depends on the concentration vector C, the temperature T and (sometimes) a parameter p

F_C Jacobian matrix of the chemical source term F regarding the concentration C

$f(z^1, z^2, \epsilon)$ Function in a singularly perturbed differential equation

G Transport term

$g(z^1, z^2, \epsilon)$ Function in a singularly perturbed differential equation

h Step size of a numerical integration method

I Identity matrix. Thereby the dimension is not given explicitly

I_c Set of eigenvalues, which fulfil $\lambda_i \leq c \ll 0$

I_n Identity matrix in $\mathbb{R}^{n \times n}$

I_{PEA} The subset of chemical reactions in partial equilibrium

J_i Total flux of the ith chemical species

J_{QSSA} The subset of species in quasi-steady state

$k_j(T)$ Temperature-dependent reaction rate constant

M Low-dimensional manifold

m_g Number of spatial grid points

m_r Number of chemical reactions

m_s Number of species

n Vector containing the amount of all chemical species, which is composed of the n_i

n_f The difference $m_s - n_s$ between the full dimension and the reduced dimension

n_i Amount of the ith chemical species

n_j^{Arr}	Reaction-dependent Arrhenius parameter
n_m	Number of given measurements
n_p	Dimension of the unknown parameter p^*
n_s	Reduced dimension of the differential equation
$\tilde{n}_s$	$\tilde{n}_s = n_0 + Q_s y_s$
n_0	Vector containing the amount of all chemical species at the initial time t_0
$n_\perp$	Outer normal vector
P	Mapping of the state of the system C onto the local coordinates θ
p	Parameter that is used for unknown reaction rate constants p^*
$p_{\cdot j}$	Products of the jth chemical reaction
P_f	Projection on the subspace spanned by Q_f
p_{press}	Pressure, which depends on the concentration and the temperature (ideal gas law)
P_s	Projection on the subspace spanned by Q_s
p_s	Standard pressure $p_s = 1\text{bar}$
p^*	Unknown reaction rate constants
Q	$Q = [Q_s, Q_f]$ is the orthogonal matrix of the Schur-decomposition $Q\tilde{T}Q^T$
q	The vector $q \in \mathbb{R}^{m_s}$ is a direction in the reaction space
Q_f	Matrix, which spans the equilibrium subspace (**f**ast reactions)
Q_s	Matrix, which spans the complement of the equilibrium subspace (**s**low reactions)
Q_T	Source term of the temperature
R	Stoichiometric matrix, which is composed of the reaction vectors r_j
$R(\cdot)$	Stability function of a numerical integration method
r_j	Reaction vector
R_m	Gas constant
T	Temperature
$\tilde{T}$	Triangular matrix of the Schur-decomposition
t	Time
t_e	Final time
t_i	Instant of time $t_i = t_0 + ih$ or $t_i = t_0 + i\Delta t$
T_s	Standard temperature $T_s = 298.15\text{K}$
TOL.	Absolute or relative tolerance

t_0	Initial time
u	Flow field
V	Volume of the chemical reaction system
$v(\cdot)$	Reaction rate, which depends on the concentration C and the temperature T
v_c	The vector v_c characterizes a conserved quantity
V_D	Eigenvector matrix
$v_{D,i}$	Eigenvector to the eigenvalue λ_i
v_i^{∞}	Contribution of all chemical reactions that are not in partial equilibrium
w	Transformation $w := V^{-1}C$ of the concentration
x	Position in space
x_f	Slack variable
y_e	Extent of the reaction
y_f	Partial extent of the fast reactions
y_s	Partial extent of the slow reactions
z	$z := \lambda \Delta t$ is the argument of the stability function
z^i	Variables in a singularly perturbed differential equation

1 Introduction

The chemical industry is one of the most important branches of industry. Among other things, plastics, fertilizer, pesticides, and pharmaceuticals are manufactured in the chemical industry. Therefore, chemical industry is necessary for the fabrication of most everyday products. An example for the importance of chemical industry is given by the Haber-Bosch process, which is an industrial process for ammonia synthesis. Ammonia is a worldwide used fertilizer. Hence, it serves for the global production of food. This chemical process is so important that three Nobel prizes have been awarded to Fritz Haber, Carl Bosch, and Gerhard Ertl for developing and investigating the Haber-Bosch process.

There are numerous challenges, such as the development of new energy sources, the food supply of the growing world population or resource shortages, which demand the enhancement of existing chemical processes or the development of new chemical processes. Thereby numerical simulations of the chemical reaction systems are used for the analysis. In a numerical simulation various parameters can be changed freely. Hence, the influence of different chemical reactions or species can easily be investigated. This is an enormous advantage over the complex analysis by means of real experiments. Furthermore, chemical reaction systems cover several different timescales. The fastest chemical reactions are completed after a fraction of a second. Hence, it is difficult to observe these fast reactions in actual measurements. However, if the reaction rate constants are known from other chemical reaction settings (e.g. a chemical reactor with less species), very fast processes can be modelled with numerical simulations. Thus, the simulation of chemical reaction systems results in lower costs and better results than experimenting with different conditions.

For known chemical processes, the production of chemicals is then carried out in chemical reactors. The most common vessel types are tank reactors and pipe reactors. Furthermore, it is distinguished between continuous reactors and batch reactors. A continuous reactor has an inflow and an outflow. Therefore, educts are added continuously, and a corresponding amount of products is removed. In comparison a batch reactor is filled once, and all products are removed after the terminated chemical process. Thus, the geometry, the inflow, and the outflow can be varied. Moreover, parameters like temperature and pressure can be controlled. In total there are unlimited design possibilities for chemical reactors. In order to guarantee the most efficient mode, the reactor design is adapted to the chemical process. One possibility to optimize the design is to manufacture reactor prototypes and to test them. However, this is both cost-intensive and time-consuming. Therefore, the numerical simulation of chemical reactors is now used in the development of new reactors because it is cost-efficient and fast.

Hence, the numerical simulation of chemical reaction systems has a wide field of applications, and numerous publications appear in this branch of science. For example, in 2013 Martin Karplus, Michael Levitt, and Arieh Warshel have been awarded with the Nobel Prize in Chemistry for developing computer models that simulate chemical reactions. Ongoing interest by scientists of different fields indicates that several problems regarding the simulation of chemical reaction systems are still unsolved. For example, different timescales of chemical reactions result in stiff systems of differential equations, whose numerical solution requires implicit solution methods. However, the number of degrees of freedom is very large for chemical reaction systems because the number of involved chemical species and the discrete number of points over the considered domain tend to be large. Therefore, computing the numerical solution for the next time step involves the solution a high-dimensional, highly nonlinear

equation system. Hence, direct numerical simulation (DNS) is computational demanding or not feasible due to limited computational resources. However, additional information is available due to the existing of very fast chemical reactions, and this additional information can be used for the reduction of the dimension of the differential equation. Then the reduced differential equation can be solved much faster. Moreover, in some applications it is necessary to solve the corresponding differential equation for many different parameter settings. E.g., parameter estimation requires many forward solutions of the chemical reaction system. Thereby the existence of very fast chemical reactions might be used for the reduction of the dimension of the parameter space. Thus, using the occurrence of different timescales for parameter estimation of chemical reaction systems is a possible field of research. In the following PhD thesis some of these aspects are examined thoroughly.

1.1 Outline of the present work

In this work numerical aspects of the simulation of chemical reaction systems are elaborated. In Chapter 2, a chemical reaction system is converted to a differential equation. Thereby spatially homogeneous reactors, which result in ordinary differential equations (ODEs), and spatially heterogeneous reactors, which result in partial differential equations (PDEs), are considered. Afterwards the obtained differential equation can be solved with a numerical integration method. However, the dimension of the occurring differential equation is often very high, and the computational costs of the implicit integration of the high-dimensional differential equation might be prohibitive. Moreover, the occurrence of many different timescales results in very stiff differential equations. Chapter 3 treats reduction mechanisms for chemical reaction systems. The state of the system approaches a low-dimensional manifold due to the presence of fast chemical reactions. The state of the system can be approximated by a point on the low-dimensional manifold. Thus, the dimension of the describing differential equation can be reduced. However, additional algebraic equations are necessary for the characterization of the low-dimensional manifold. The algebraic constraints follow from the partial equilibrium assumption (PEA) and the quasi-steady state assumption (QSSA). These are introduced in Section 3.1. In Section 3.5, recommended reduction mechanisms are presented, and in Section 3.6 the validity of reduction mechanisms that are based on timescale separation is examined. In order to save computing time with a reduction mechanism, the state of the system has to be restricted to a low-dimensional manifold, which can be precomputed and stored in a look-up table. Although, the state of the system might be restricted to a manifold, whose dimension is too large for a look-up table (curse of dimensionality). Furthermore, in case of a partial differential equation the dimension of the low-dimensional manifold might change in space, which results in a discontinuity. Therefore, reduction mechanisms are not always applicable. In Chapter 4, a new approach [85] is derived, that reduces the stiffness of the system and is also applicable if the state of the system does not lie on a low-dimensional manifold. The proposed new method reduces the range of the occurring timescales, and thereby reduces the stiffness of the considered differential equation. In contrast to the QSSA and the PEA, the reaction rate of the fast processes is not set to infinity but is decreased. Afterwards in Chapter 5, parameter estimation of unknown reaction rate constants is examined. Thereby the usage of QSSA and PEA for the identification of unknown parameters is illustrated. It is shown that the additional information that is obtained by QSSA and PEA can be used for the reduction of the dimension of the parameter space. Thereby ODEs are considered for simplicity in Chapters 3 to 5. However, most

chemical reactor settings are described by a PDE. Operator splitting can be used to obtain a subproblem that only considers the transport operator (advection and diffusion) and a subproblem that only considers chemistry. The transport-only equation can be solved with specialised solvers. Furthermore, in case of a chemistry-only subproblem the considered PDE decomposes to one system of ordinary differential equations for each spatial gridpoint. Thus, the results from the Chapters 3 to 5 are applicable in case of PDEs if operator splitting is used. Recommended operator splitting methods are the Lie-Trotter splitting and the Strang splitting. The Lie-Trotter splitting has order one, and the Strang splitting has order two. However, in case of stiff differential equations order reduction occurs for the Strang splitting scheme. The extrapolated Lie-Trotter splitting is investigated in Chapter 6. It is shown that the extrapolated Lie-Trotter splitting has order two for chemical reaction systems with fast chemistry. Thus, the extrapolated Lie-Trotter splitting should be used in case of fast chemistry.

2 Modelling chemical reaction systems

In this chapter, the basic principles of the modelling of chemical reaction systems are illustrated. In Section 2.1, the chemical source term is introduced by the discussion of a homogeneous chemical reaction system. Thereby four different reaction conditions usually are examined in case of a gas-phase chemical reaction system. The four setups are an adiabatic system with constant volume, a chemical reaction system with constant volume and constant temperature, an adiabatic system with constant pressure, as well as a system with constant pressure and constant temperature. Other applications occur in the simulation of chemical reactions in a liquid solution.
However, in most applications the considered chemical reaction systems are heterogenous in space. Hence, advection and diffusion are examined for a spatial inhomogeneous reactor with a given flow field in Section 2.2.

2.1 Homogeneous chemical reaction system

Firstly, a homogeneous chemical reaction system is examined. In a homogeneous chemical reaction system, the concentrations of the different chemical species are constant in space. Therefore, any transport process can be neglected, and changes in the amount of a species are solely caused by chemical reactions. The chemical reaction system contains m_s different species $A_i,\ i \in \{1,2,...,m_s\}$. The amount of the species A_i is denoted by n_i, and the amount of all species is a vector $n \in \mathbb{R}^{m_s}$. Furthermore, assume that the number of chemical reactions is m_r. Each reaction transforms some educts into some products, which results in

$$\sum_{k=1}^{m_s} e_{kj} A_k \rightarrow \sum_{l=1}^{m_s} p_{lj} A_l, \qquad j \in \{1,2,...,m_r\}\,.$$

Thus, in the jth reaction e_{kj} mol of the species $A_k,\ k \in \{1,2,...,m_s\}$, which are the educts, are transformed into p_{lj} mol of the species $A_l,\ l \in \{1,2,...,m_s\}$, which are the products. Thereby the direction of each reaction can be written as a reaction vector

$$r_j = \begin{bmatrix} p_{1j} - e_{1j} \\ \vdots \\ p_{m_s j} - e_{m_s j} \end{bmatrix}.$$

A reaction vector r_j shows the direction of possible changes in the amount of all species due to the jth reaction. However, the progress of the chemical reaction in time t is not characterized by the reaction vector r_j because the velocity of the chemical reaction is not part of the reaction vector. In the following all reaction vectors are combined in a stoichiometric matrix R, whose columns are the single reaction vectors. The stoichiometric matrix is

$$R = (r_1, r_2, ..., r_{m_r}) \in \mathbb{R}^{m_s \times m_r}\,. \tag{2.1}$$

Moreover, the state of the system n fulfils

$$n = n_0 + Ry_e\,. \tag{2.2}$$

Thereby the vector $n_0 \in \mathbb{R}^{m_s}$ contains the initial values for all species at initial time t_0, and $y_e \in \mathbb{R}^{m_r}$ is the extent of reactions. The volume of the reactor is denoted by V. Then the reaction rate is defined as

$$v := \frac{\partial (y_e/V)}{\partial t} \in \mathbb{R}^{m_r}\,.$$

Hence, the reaction rate v is the temporal change of the extent of reaction per volume V. Instead of the vector n, which is an extensive property (i.e. is additive for subsystems), the concentrations $C_i := [A_i]$ of the different species A_i can be considered. The concentration of the different species is an intensive property (i.e. does not depend on the system size). Thus, the concentration $C := (C_1, ...C_{m_s})^T$ of all chemical species is locally defined. Therefore, it is possible to examine chemical reaction systems, whose different chemical species are not constant in space. However, the quotient rule is necessary in the time derivative if the volume changes with time. In order to simplify the problem chemical reaction systems with a fixed volume are examined in the following (e.g., a gas-phase chemical reaction system with constant volume and constant temperature, an adiabatic system with constant volume, or a chemical reaction system in a liquid solution with constant density). The concentration of all species is described by

$$C = \frac{n}{V} = \frac{n_0}{V} + R\frac{y_e}{V}\,. \tag{2.3}$$

Computing the time derivative of (2.3) and inserting the definition of the reaction rate $v = \frac{\partial (y_e/V)}{\partial t} \in \mathbb{R}^{m_r}$ results in

$$\frac{\partial C}{\partial t} = Rv(C,T), \qquad C(0) = C_0 = \frac{n_0}{V}\,. \tag{2.4}$$

Note that equation (2.4) only holds for a homogeneous chemical reaction system with constant volume. If the volume is not constant, the species mass fractions or the specific mole numbers should be considered (detailed discussion in [98] or [92]). The reaction rate v is necessary for solving the differential equation (2.4). Therefore, elementary reactions are introduced. An elementary reaction is a substep of the reaction mechanism that is not divisible. Every chemical reaction is the sum of elementary reactions. The reaction rate of an elementary reaction can easily be computed because it is proportional to the number of collisions of all educts. Moreover, the number of collisions is proportional to the product of the concentrations of all educts. Hence, the reaction rate of the jth elementary reaction fulfils

$$v_j(C,T) = k_j(T)\prod_{k=1}^{m_s}\left(C_k\right)^{e_{kj}}\,. \tag{2.5}$$

Thereby $k_j(T)$ is the temperature-dependent reaction rate constant of the jth reaction. The reaction rate constant can be computed with the Arrhenius equation [5]. Arrhenius stated that the jth reaction takes place if the corresponding educt molecules are colliding and the available energy is larger than a necessary activation energy $E_{A,j}$. According to the Maxwell-Boltzmann distribution, the fraction of molecules that possess a larger energy than $E_{A,j}$ is proportional to $e^{E_{A,j}/(R_m \cdot T)}$. Note that R_m is not the stoichiometric matrix, but it is the gas constant. However, the obtained rate does not describe the actual reaction rate constant for all reactions. Hence, an empirical factor T^{β_j} is augmented. Finally, the reaction rate constant $k_j(T)$ is given by

$$k_j(T) = A_j^{Arr} \cdot T^{\beta_j} \cdot e^{\frac{E_{A,j}}{R_m \cdot T}} \tag{2.6}$$

with reaction-dependent parameters A_j^{Arr}, β_j, and $E_{A,j}$. The parameters of the Arrhenius equation are determined empirically. For some reactions they are not known or only the ratio of the reaction rate constants of the forward and the backward reaction is given. However, for most elementary reactions the parameters are listed in databases like NIST Standard Reference Database Number 17 [93]. If elementary reactions are considered and the corresponding parameters are known, the source term of the chemical reaction system only depends on the concentrations of all species and the temperature. Thus, the differential equation (2.4) can be solved. For advection-diffusion-reaction systems the change of concentration due to chemical reactions can also be described similar to (2.4). However, in general additional terms are necessary in order to account for transport processes like diffusion or advection. In Section 2.2, the differential equation (2.4) will be expanded to advection-diffusion-reaction systems.

Remark 2.1. *The introduced variables n and C are variables in the positive real numbers $\mathbb{R}^{m_s}$. However, molecules come in whole numbers. Therefore, the assumption of continuous variables is a relaxation, which results in a simplified description* (2.4) *of the state of the system. If the size of the considered reactor is large enough, the differential equation* (2.4) *works very well. However, if the considered reactor is very small (e.g., a living cell), discreteness and stochasticity have to be considered. Approaches for the stochastic simulation of chemical kinetics are listed in [55]. Furthermore, a stochastic approach that uses the occurrence of different timescales in chemistry is introduced in [23].*

2.2 Advection-diffusion-reaction system

In this section, a heterogenous chemical reactor is described. For simplification a liquid solution with constant density is modelled. This assumption corresponds to a constant volume of a homogeneous chemical reactor. Consider the concentration C_i of the chemical species A_i. The concentration depends on the position $x \in \mathbb{R}^d$ and the time t. Thereby d is the dimension of the chemical reactor. Thus, it holds $C_i := C_i(x,t)$. Furthermore, assume that $\Omega \subset \mathbb{R}^d$ is a bounded region in the chemical reactor, and $\partial\Omega$ is its boundary. Thereby Ω does not change in time. Now the amount $n_i(\Omega,t)$ of the species A_i in the domain Ω at time t is examined. One obtains

$$n_i(\Omega,t) = \int_\Omega C_i(x,t)\mathrm{d}x\,.$$

The time derivative of $n_i(\Omega, t)$ is given by the negative flux of C_i through the boundary $\partial\Omega$ plus the change $F(C,T)$ due to chemical reactions. Thus, it holds

$$\frac{\partial}{\partial t} n_i(\Omega, t) = \frac{\partial}{\partial t} \int_\Omega C_i(x,t)\mathrm{d}x = -\int_{\partial\Omega} J_i \cdot n_\perp \mathrm{d}s + \int_\Omega \big(F(C,T)\big)_i \mathrm{d}x\,. \tag{2.7}$$

Thereby $n_\perp$ is the outer normal vector, and J_i is the total flux. The total flux J_i measures the amount of the species A_i that flows through a unit area in a unit time interval. Due to Gauss' theorem, equation (2.7) can be transformed to

$$\int_\Omega \frac{\mathrm{d}C_i}{\mathrm{d}t}\mathrm{d}x = -\int_\Omega \mathrm{div} J_i \mathrm{d}x + \int_\Omega \big(F(C,T)\big)_i \mathrm{d}x\,. \tag{2.8}$$

The domain Ω is arbitrary in equation (2.8). Hence, it follows

$$\frac{\mathrm{d}C_i}{\mathrm{d}t} = -\mathrm{div} J_i + \big(F(C,T)\big)_i\,. \tag{2.9}$$

The total flux J_i is composed of the diffusion flux and the advection flux. The diffusion flux arises due to diffusion. It can be approximated by Fick's first law. The advection flux arises due to a flow field $u(x,t)$. Due to the constant density of the liquid solution, the flow field is independent of the concentrations of the different chemical species. With the given simplification the incompressible Navier-Stokes equation can be used in order to precompute the flow field $u(x,t)$. Afterwards, the flow field u is available for the computation of the concentrations of all chemical species. Thus, the total flux is given by

$$J_i = -D_i \nabla C_i + u C_i\,. \tag{2.10}$$

Thereby D_i is the diffusion coefficient of C_i. Note that several simplifications are used. First of all, a liquid solution with constant density is considered. The density of an ideal gas is not constant. In this case the species mass fractions or the specific mole numbers are modelled [92, 98]. Moreover, it is assumed that the flow field u does not depend on the concentration C_i. If the flow field depends on the concentration C_i (for example, the density is not constant due to temperature changes), a coupled system of differential equations is derived. (e.g., a detailed discussion for a gas-phase chemical reaction system is given in [98]). Furthermore, the diffusion is assumed to be isotrop. Hence, the diffusion is uniformly in all orientations. However, this is not exact due to some effects like thermophoresis. Thermophoresis describes the effect of particles moving from a hot to a cold region due to a temperature gradient. Moreover, the applicability of the Navier-Stokes equation is limited to continuous fluids (see Remark 2.1) that are Newtonian fluids. Some of these effects can be modelled by additional equations or an additional source term. However, some simplifications (like the assumption of a continuous fluid) are necessary for a reasonable computing time of the numerical integration method.

The differential equation for C_i is obtained by inserting equation (2.10) into equation (2.9). Moreover, in case of a liquid solution with constant density the chemical source term is $F(C,T) = R\nu(C,T)$. Thus, it holds

$$\frac{\mathrm{d}C_i}{\mathrm{d}t} = -\operatorname{div}\big(-D_i\nabla C_i + uC_i\big) + \big(R\nu(C,T)\big)_i\,. \tag{2.11}$$

The velocity field of an incompressible flow is a divergence free vector field. Hence, we obtain

$$\frac{\mathrm{d}C_i}{\mathrm{d}t} = \operatorname{div}\big(D_i\nabla C_i\big) - u\cdot\nabla C_i + \big(R\nu(C,T)\big)_i\,. \tag{2.12}$$

In case of a constant diffusion coefficient D_i the differential equation for C_i is

$$\frac{\mathrm{d}C_i}{\mathrm{d}t} = D_i\Delta C_i - u\cdot\nabla C_i + \big(R\nu(C,T)\big)_i\,. \tag{2.13}$$

Note that the differential equations for all C_i are coupled by the chemical source term $R\nu(C,T)$. Furthermore, the spatial coupling is generated by the diffusion and advection term. Thus, if the transport term is zero (no diffusion and no advection), equation (2.13) results in an ODE with dimension m_s for each spatial grid point. Moreover, if the chemical source term is zero, a PDE is obtained for each species A_i. A similar differential equation holds for the temperature T. The temperature is described by

$$\frac{\mathrm{d}T}{\mathrm{d}t} = \alpha\Delta T - u\cdot\nabla T + Q_T\,. \tag{2.14}$$

Thereby α is the thermal diffusivity, and Q_T is the source term due to chemical reactions. The source term Q_T can be computed from the heat capacity of the chemical mixture and the standard enthalpy of reaction [51]. Thereby the standard enthalpy of reaction is listed in databases like NIST Standard Reference Database Number 69 [83]. Note that the enthalpy of a chemical reaction system often is modelled instead of the temperature T. The corresponding differential equation is given in [13, 115].

2.3 Boundary conditions

Boundary conditions are required for the considered PDE. The used boundary conditions are the Neumann-condition and the Dirichlet-condition. The Dirichlet-condition provides the values at the boundary of the domain. It is used for the concentration at the inflow. Furthermore, it is used for the temperature at the inflow and heated or cooled walls. The Neumann-condition provides the derivative in direction of the normal vector $n_\perp$. Regarding the temperature, it describes the heat flow at the boundary. Thus, an inhomogeneous Neumann-condition is given for a wall that is non-isolated. Furthermore, adiabatic walls and the outflow of the domain are characterized by the homogeneous Neumann-condition. Regarding the concentration of a chemical species, the homogeneous Neumann-condition is used for all boundaries except the inflow of the domain.

Remark 2.2. *A general discussion of gas-phase chemical reaction systems is given in [98]. Thereby the species mass fraction, the temperature, and the pressure are the unknown variables.*

Remark 2.3. *In the following chapters, all considered homogeneous chemical reaction systems have constant volume. Hence, equation* (2.4) *describes the considered homogeneous chemical reaction systems. Furthermore, all considered heterogenous chemical reaction systems are in the form* (2.13)*. The only reason for the restriction to equations* (2.4) *and* (2.13) *is the simple form of the chemical reaction rate (reaction rate constant times a product of concentrations, which are the unknown variables). However, the derived results are also valid for other systems.*

3 Reduction of chemical reaction systems

The numerical simulation of chemical reaction systems has a wide field of applications (see Chapter 1). However, the occurring systems of differential equations are very large and strongly nonlinear. Moreover, typical chemical reaction systems contain several different timescales, ranging from 10^{-9} seconds to 10^{2} seconds [92]. Hence, the resulting systems of differential equations are very stiff. Stiff differential equations cause very small step sizes for explicit integration schemes. Therefore, implicit integration schemes are usually used [28]. Every step of an implicit integration method requires the solution of a nonlinear equation system. Thereby the size of the equation system is equal to the number of unknown variables. Thus, the numerical solution of large reaction mechanisms in complex geometry is computationally demanding or unfeasible. Furthermore, large rounding errors occur in the evaluation of the chemical source term due to very fast reactions. E.g., consider a fast chemical reaction in partial equilibrium. It is assembled by a forward and a backward reaction, which cancel each other out. Hence, it has a small net reaction rate, but the forward and backward reaction rates are very large. Therefore, considerable round-off errors by cancellation occur. However, the described problems can be avoided if a gap in the timescales of the processes exists. In this case the state of the system evolves close to or on a low-dimensional manifold that is defined by the partial equilibrium of the fast processes. The state of the system changes on this low-dimensional manifold due to slow chemical reactions and transport phenomena. In the following the existence of an attracting low-dimensional manifold is explained by the partial equilibrium assumption (PEA) [22, 56, 57, 97, 110] and the quasi-steady state assumption (QSSA) [11, 54, 97, 105, 131]. PEA and QSSA are introduced in Section 3.1. Applying the PEA on a single chemical reaction or the QSSA on a single species relies on experience and intuition of the chemist. However, in case of large reaction systems, the examination of all reactions and all species is very time consuming or impossible. Hence, the automatic determination of fast processes by an eigendecomposition is examined in Section 3.2. In Section 3.3, the existence of the low-dimensional manifold is illustrated by a (fictious) small-scale example. Afterwards, the incorporation of transport processes is examined in Section 3.4. Moreover, recommended reduction mechanisms are introduced in Section 3.5. Finally, the validity of reduction mechanisms is investigated by singular perturbation theory in Section 3.6.

3.1 Partial equilibrium assumption (PEA) and quasi-steady state assumption (QSSA)

Most reduction mechanisms are based on partial equilibrium assumption and quasi-steady state assumption. These approaches are presented in this section. For simplicity a spatially homogeneous reactor is examined. The corresponding ODE is (2.4). The chemical source term $F(C,T)$ occurs naturally in the following form [8]:

$$F(C,T) = R\nu(C,T) = \sum_{i=1}^{m_r} r_i \nu_i(C,T). \tag{3.1}$$

Thereby all reactions are elementary reactions. Furthermore, there is a forward and a backward reaction for every reaction direction. The reactions are sorted such that $r_{2i-1} = -r_{2i}$ for $i = \{1,2,...,m_r/2\}$. Hence, the ith reaction direction has contributions $r_{2i-1}\nu_{2i-1}$ and $-r_{2i-1}\nu_{2i}$.

Example 1. *Consider the chemical system*

$$HCl + H\cdot \rightarrow_{k_1} H_2 + Cl\cdot$$
$$H_2 + Cl\cdot \rightarrow_{k_2} HCl + H\cdot\,.$$

The state of the system is $C := \Big([HCl],\quad [H\cdot],\quad [H_2],\quad [Cl\cdot]\Big)^T$. *Then the corresponding differential equation is*

$$\begin{aligned} C' &= \begin{pmatrix} -1 \\ -1 \\ +1 \\ +1 \end{pmatrix} k_1(T)\cdot[HCl]\cdot[H\cdot] + \begin{pmatrix} +1 \\ +1 \\ -1 \\ -1 \end{pmatrix} k_2(T)\cdot[H_2]\cdot[Cl\cdot] \\ &= \sum_{i=1}^{2} r_i\, v_i(C,T) \end{aligned}$$

In the following the PEA is introduced. The PEA is used for directions of reactions with a small timescale that is smaller than the time period of interest. If a fast chemical reaction is exhausted, it is in partial equilibrium. Thus, its net reaction rate is approximately equal to zero. Hence, the PEA for the ith pair of forward and backward reaction is

$$v_{2i-1}(C,T) - v_{2i}(C,T) = 0\,. \tag{3.2}$$

This approximation can be used in order to solve for one of the involved species in terms of the other involved species.

Example 2. *The most familiar example of a partial equilibrium is the ionization of water. Therefore, a dilute aqueous solution is considered. The occurring chemical reaction system includes the self-ionization reaction of water, which is*

$$2H_2O \rightarrow_{k_1} H_3O^+ + OH^-$$
$$H_3O^+ + OH^- \rightarrow_{k_2} 2H_2O\,.$$

The corresponding forward/backward reaction of the self-ionization is in partial equilibrium (potential of hydrogen pH). Therefore, it holds $k_1[H_2O][H_2O] \approx k_2[H_3O^+][OH^-]$. *For a dilute aqueous solution it holds* $[H_2O] \approx 55.5\text{mol/l}$ *and the PEA results in*

$$\frac{[H_3O^+]}{\text{mol/l}}\frac{[OH^-]}{\text{mol/l}} = 10^{-14}\,. \tag{3.3}$$

The approximation error of (3.3) *is very small.*

In spite of giving a good approximation for a species, the PEA (3.2) cannot be used in order to eliminate $v_{2i-1}(C,T)$ and $v_{2i}(C,T)$ from the differential equation because the small net reaction rate of an exhausted fast chemical reaction is not negligible in comparison to an active slow chemical reaction. Example 3 clearifies this problem. In the following the unknown solution of the modified differential equation is denoted by $\tilde{C}$. Thereby the modification can be obtained by the PEA, the QSSA, or any other reduction mechanism.

Example 3. *Consider the reaction system*

$$\begin{aligned} A_1 &\to A_2, & k_1 &= 10^3, \\ A_2 &\to A_1, & k_2 &= 10^3, \\ A_3 &\to A_1, & k_3 &= 1\,. \end{aligned}$$

Thereby the reaction rate constants k_i, $i \in \{1,2,3\}$, do not depend on the temperature T. The reaction system results in the following differential equation:

$$\begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix}' = \begin{pmatrix} -1 \\ 1 \\ 0 \end{pmatrix} k_1 C_1 + \begin{pmatrix} 1 \\ -1 \\ 0 \end{pmatrix} k_2 C_2 + \begin{pmatrix} 1 \\ 0 \\ -1 \end{pmatrix} k_3 C_3 + \begin{pmatrix} -1 \\ 0 \\ 1 \end{pmatrix} 0\,.$$

The sum $C_1 + C_2 + C_3$ is a conserved quantity of the full reaction mechanism, because the stoichiometric matrix R fulfils

$$\begin{pmatrix} 1 & 1 & 1 \end{pmatrix} \cdot R = \begin{pmatrix} 0 & 0 & 0 & 0 \end{pmatrix}.$$

Assume that the first reaction is in partial equilibrium. Therefore, $k_1\tilde{C}_1 = k_2\tilde{C}_2$ is used, which results in $\tilde{C}_1 = \frac{k_2}{k_1}\tilde{C}_2 = \tilde{C}_2$. The solution C of the full differential equation fulfils $C_1 \approx C_2$. Hence, the PEA results in a good approximation for the considered ODE. However, if $\tilde{C}_1 = \tilde{C}_2$ is substituted into the differential equation, the reduced model is

$$\begin{aligned} \begin{pmatrix} \tilde{C}_1 \\ \tilde{C}_3 \end{pmatrix}' &= \begin{pmatrix} 1 \\ -1 \end{pmatrix} k_3 \tilde{C}_3, \\ \tilde{C}_2 &= \tilde{C}_1\,. \end{aligned}$$

Obviously, the sum $\tilde{C}_1 + \tilde{C}_3 = \tilde{C}_1(t_0) + \tilde{C}_3(t_0)$ is constant. It follows $\tilde{C}_1(t) + \tilde{C}_2(t) + \tilde{C}_3(t) = \tilde{C}_1(t_0) + \tilde{C}_3(t_0) + \tilde{C}_2(t) = \tilde{C}_2(t) + const.$ Thus, the quantity $\tilde{C}_1 + \tilde{C}_2 + \tilde{C}_3$ is not conserved. The approximation error of the reduced model is at least equal to the error in the conserved quantity. Hence, the approximation error of the reduced model is large. The reason is that the small net reaction rate of an exhausted fast chemical reaction is not negligible in comparison to an active slow chemical reaction.

An approach for the elimination of the exhausted reaction rates from the differential equation is given in [78]. Define a set $I_{PEA} \subset \{1,2,...,m_r/2\}$ that is the subset of all chemical reaction directions in partial equilibrium. For $i \in I_{PEA}$ the net reaction rates $\nu_{2i-1}(C,T) - \nu_{2i}(C,T)$ are differentiated with respect to time. Differential equations for $\nu_{2i-1}(C,T) - \nu_{2i}(C,T),\ i \in I_{PEA}$, are obtained as a result. Following [78] these differential equations can be transformed to the following form:

$$\frac{\mathrm{d}(\nu_{2i-1} - \nu_{2i})}{\mathrm{d}t} = \frac{-1}{\tau_i}\Big[(\nu_{2i-1} - \nu_{2i}) + \sum_{j \in I_{PEA}, j \neq i} c_{ij}\big(\nu_{2j-1} - \nu_{2j}\big) - \nu_i^{\infty} \Big]$$

with temperature and state dependent coefficients $c_{ij},\ i \neq j \in I_{PEA}$. Moreover, $\nu_i^{\infty},\ i \in I_{PEA}$, is the contribution of all chemical reactions that are not in partial equilibrium. If the timescales $|\tau_i|,\ i \in I_{PEA}$, of the partial equilibrium reactions converge against zero, the terms in squared brackets also converge against zero, which results in a system of linear equations for $(\nu_{2i-1}(\tilde{C},T) - \nu_{2i}(\tilde{C},T)),\ i \in I_{PEA}$. The approach is illustrated in the following example.

Example 4. *Consider the reaction system from Example 3. The first reaction is in partial equilibrium. One obtains*

$$\begin{aligned}
\nu_1 - \nu_2 &= k_1 C_1 - k_2 C_2, \\
\frac{d(\nu_1 - \nu_2)}{dt} &= k_1 C_1' - k_2 C_2' \\
&= k_1\big(-\nu_1 + \nu_2 + \nu_3 - \nu_4\big) - k_2\,(\nu_1 - \nu_2) \\
&= (-k_1 - k_2)\Big[(\nu_1 - \nu_2) - \frac{k_1}{k_1 + k_2}(\nu_3 - \nu_4) \Big] \\
&\Longrightarrow \big(\nu_1(\tilde{C},T) - \nu_2(\tilde{C},T)\big) \approx \frac{k_1}{k_1 + k_2}\big(\nu_3(\tilde{C},T) - \nu_4(\tilde{C},T)\big).
\end{aligned}$$

Substituting $\nu_1(\tilde{C},T) - \nu_2(\tilde{C},T) = \frac{k_1}{k_1+k_2}\big(\nu_3(\tilde{C},T) - \nu_4(\tilde{C},T)\big)$ *and* $\tilde{C}_2 = k_1/k_2\tilde{C}_1$ *into the differential equation results in the approximation*

$$\begin{aligned}
\begin{pmatrix} \tilde{C}_1 \\ \tilde{C}_3 \end{pmatrix}' &= \begin{pmatrix} -1 \\ 0 \end{pmatrix}\left(\frac{k_1 k_3}{k_1 + k_2}\tilde{C}_3\right) + \begin{pmatrix} 1 \\ -1 \end{pmatrix} k_3 \tilde{C}_3 \\
\tilde{C}_2 &= k_1/k_2 \tilde{C}_1 .
\end{aligned}$$

Then the solution of the reduced mechanism fulfils

$$\begin{aligned}
(\tilde{C}_1 + \tilde{C}_2 + \tilde{C}_3)' &= \big(-\frac{k_1 k_3}{k_1 + k_2}\tilde{C}_3 + k_3\tilde{C}_3\big) + k_1/k_2\big(-\frac{k_1 k_3}{k_1 + k_2}\tilde{C}_3 + k_3\tilde{C}_3\big) + \big(-k_3\tilde{C}_3\big) \\
&= \big(-\frac{1}{2}\tilde{C}_3 + \tilde{C}_3\big) + 1 \cdot \big(-\frac{1}{2}\tilde{C}_3 + \tilde{C}_3\big) + \big(-\tilde{C}_3\big) = 0 .
\end{aligned}$$

Therefore, $\tilde{C}_1+\tilde{C}_2+\tilde{C}_3$ is a conserved quantity. Hence, this approximation does not violate any conservation law.

Remark 3.1. (Franz [51]) *A chemical reaction system with reactions in partial equilibrium can also be modelled by defining an equilibrium reaction subspace, spanned by an orthogonal matrix Q_f, and its complement, spanned by an orthogonal matrix Q_s. The fast processes are in the equilibrium reaction subspace. We define the projections P_f on the equilibrium reaction subspace and $P_s = I - P_f$ on its complement. A slack variable x_f restricts the state of the system onto the manifold. Then the PEA of the fast chemical reactions for the differential equation $C' = F(C,T)$ results in*

$$\begin{aligned} \tilde{C}' &= P_s F(\tilde{C},T) + Q_f x_f, \\ 0 &= Q_f^T F(\tilde{C},T). \end{aligned} \tag{3.4}$$

Hence, the reaction velocities must not be differentiated. The differential-algebraic equation (DAE) (3.4) *has index* 2*. It is also possible to obtain a DAE* (3.5) *with index* 1*.*

$$\begin{aligned} C_s' &= P_s F(C_s + Q_f x_f, T), \\ 0 &= Q_f^T F(C_s + Q_f x_f, T), \\ \tilde{C} &= C_s + Q_f x_f. \end{aligned} \tag{3.5}$$

However, an additional variable is introduced for each reaction in partial equilibrium. Thus, the dimension of both DAEs is larger than the dimension of the original ODE (2.4)*.*

In opposition to the PEA, which is an assumption for some chemical reactions, the QSSA is an assumption about some species. It assumes that some chemical species reach steady state in a time span that is smaller than the time period of interest. Assume that $J_{QSSA} \subset \{1,2,...,m_s\}$ is the subset of the chemical species that are in steady state. Note that a chemical species $A_j,\ j \in J_{QSSA}$, is a radical in many cases. The change of the concentration of a species in quasi-steady state is very low. Therefore, the QSSA results in

$$\tilde{C}_j' = 0 \qquad \forall j \in J_{QSSA}. \tag{3.6}$$

Equation (3.6) can be used in order to solve for $\tilde{C}_j,\ j \in J_{QSSA}$, and in order to eliminate one of the involved reaction rates $v_i,\ 1 \le i \le m_r$. The QSSA is clearified with an example.

Example 5. *The dissociation of dinitrogen pentoxide [52] is considered in order to illustrate the QSSA. Thereby dinitrogen pentoxide (N_2O_5) is transformed into nitrogen dioxide (NO_2) and oxygen molecules (O_2). A simplified mechanism is given by*

$$\begin{aligned} N_2O_5 &\to_{k_1} NO_2 + NO_3, & NO_2 + NO_3 &\to_{k_2} N_2O_5, \\ NO_3 + NO_2 &\to_{k_3} NO + NO_2 + O_2, & NO_3 + NO &\to_{k_4} 2NO_2. \end{aligned}$$

Thereby the reaction rate constants at $T = 298K$ are taken from the NIST Chemical Kinetics Database [93] and are given by

$$k = \begin{pmatrix} k_1 \\ k_2 \\ k_3 \\ k_4 \end{pmatrix} = \begin{pmatrix} 0.10\text{s}^{-1} \\ 1.90 \cdot 10^{-12} \frac{\text{cm}^3}{\text{molecules} \cdot \text{s}} \\ 6.56 \cdot 10^{-16} \frac{\text{cm}^3}{\text{molecules} \cdot \text{s}} \\ 2.60 \cdot 10^{-11} \frac{\text{cm}^3}{\text{molecules} \cdot \text{s}} \end{pmatrix}.$$

Furthermore, the initial values are

$$[N_2O_5](0) = 1.6 \cdot 10^{18} \frac{\text{molecules}}{\text{cm}^3}, \qquad [NO_2](0) = 10^{16} \frac{\text{molecules}}{\text{cm}^3}, \qquad [O_2](0) = 0$$
$$[NO_3](0) = 0, \qquad [NO](0) = 0\,.$$

In the considered mechanism the species NO_3 is an intermediate species and the concentration of NO_3 can be approximated by the QSSA. It holds

$$[NO_3]' = k_1[N_2O_5] - k_2[NO_2][NO_3] - k_3[NO_2][NO_3] - k_4[NO][NO_3] \overset{!}{\approx} 0,$$
$$[NO_3] \overset{!}{\approx} \frac{k_1[N_2O_5]}{k_2[NO_2] + k_3[NO_2] + k_4[NO]}\,.$$

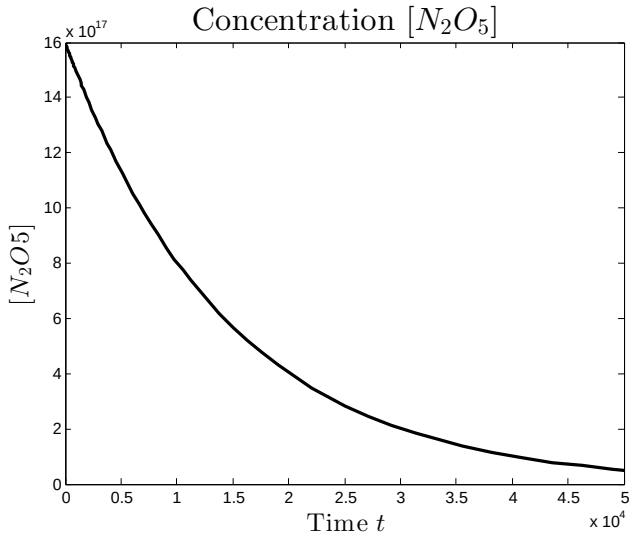

Figure 1: Temporal development of the concentration $[N_2O_5]$

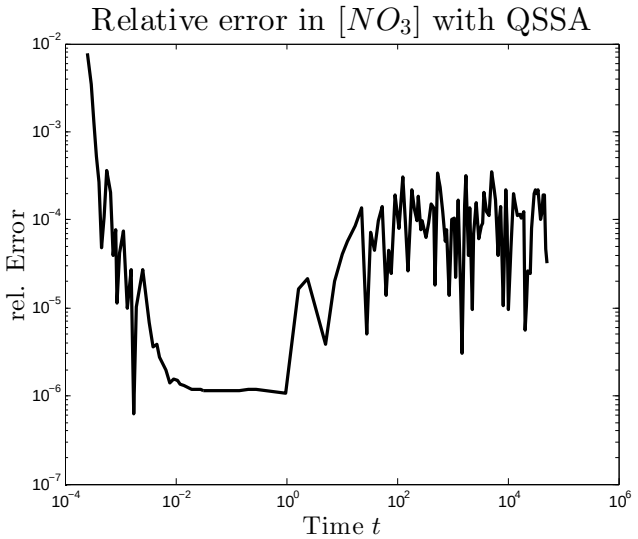

Figure 2: Relative error for the QSSA of $[NO_3]$

In Figure 1 the temporal development of the species N_2O_5 is given. The temporal development of N_2O_5 shows that the considered time interval $[0,\ 5\cdot 10^4]$ covers the dissociation of dinitrogen pentoxide. Moreover, in Figure 2 the relative error

$$\frac{\left|[NO_3]-\frac{k_1[N_2O_5]}{k_2[NO_2]+k_3[NO_2]+k_4[NO]}\right|}{[NO_3]}$$

is plotted. Figure 2 shows that the QSSA results in a very good approximation of $[NO_3]$ after the fast transient phase.

PEA and the QSSA have different areas of application, and PEA and QSSA complement one another. Furthermore, experience and intuition of the investigator are necessary in order to choose reactions in partial equilibrium and species in quasi-steady state. Thereby all reactions and all species are examined separately. Some selection rules for reactions in partial equilibrium and species in quasi-steady state are described in Subsection 5.4.1.

3.2 Introduction of automatic reduction mechanisms

If the reaction mechanism is very complex, the determination of a speed ranking of the chemical reactions or the determination of all species in quasi-steady state is hardly possible. Hence, automatic reduction mechanisms are applied for large reaction systems. Some automatic reduction mechanisms are introduced in Section 3.5. Moreover, a straightforward approach using an eigendecomposition is illustrated in the following. Many reduction mechanisms like the Intrinsic Low Dimensional Manifold method (see Section 3.5) are based on the eigendecomposition (or Schur-decomposition) of the Jacobian matrix of the chemical source term. In order to exemplify this approach, let C_∞ be the steady state of (2.4). Furthermore, let the initial value of (2.4) be a slightly perturbed state $C_0 \approx C_\infty$. Then, due to Taylor series it holds

$$\begin{aligned} C' &= F(C,T), \\ C'_\infty &= F(C_\infty,T) = F(C,T) + F_C(C,T)(C_\infty - C) + \mathcal{O}\left(\|C-C_\infty\|^2\right) \\ &\approx F(C,T) + F_C(C,T)(C_\infty - C)\,. \end{aligned}$$

Thus, one obtains

$$\begin{aligned} \frac{\mathrm{d}(C-C_\infty)}{\mathrm{d}t} &= F_C(C,T)(C-C_\infty) + \mathcal{O}\left(\|C-C_\infty\|^2\right) \\ &\approx F_C(C,T)(C-C_\infty)\,. \end{aligned}$$

According to this, the discrepancy of the actual state to the steady state can be modelled by a linear differential equation with coefficient $F_C(C,T)$. Assume that the coefficient $F_C(C,T)$ is constant and that there exists the following eigendecomposition with a diagonal matrix D and eigenvector matrix V_D:

$$F_C(C,T) = V_D D {V_D}^{-1}$$

If $F_C(C,T)$ is not diagonalizable, a similar result is obtained by using the Jordan normal form. Furthermore, the eigenvalues of the matrix D at position D_{ii} are denoted by λ_i, and the corresponding eigenvectors are $v_{D,i}$. Then a new variable is introduced

$$w = {V_D}^{-1}(C - C_\infty)\,.$$

It holds

$$\begin{aligned} w' &= {V_D}^{-1}\frac{\mathrm{d}(C - C_\infty)}{\mathrm{d}t}, \\ &= Dw\,. \end{aligned} \tag{3.7}$$

Thus, components corresponding to negative eigenvalues with large absolute values decay very fast. A set $I_c = \{i \mid \lambda_i \leq c, \quad 1 \leq i \leq m_s\}$ is defined. Thereby $c \ll 0$ is a predefined constant. Hence, $w_i \approx 0$ for $i \in I_c$ during the time period of interest. Then the state of the system evolves close to or on a low-dimensional manifold M that is defined by

$$\tilde{w}_i' = \left({V_D}^{-1}\right)_{i\cdot} F(\tilde{C},T) \stackrel{!}{=} 0 \qquad \forall i \in I_c\,. \tag{3.8}$$

Note that the movement on the manifold is caused by the slow processes. Therefore, the stiffness of the system is induced by the movement perpendicular to the manifold. Moreover, equation (3.8) reduces the degrees of freedom. Thus, there exists a low-dimensional parameter θ that parameterizes the manifold M such that $M = \{\tilde{C}(\theta) \in \mathbb{R}^{m_s} \;:\; \theta \in \mathbb{R}^{n_s}\}$. Thereby the mapping $\theta \to \tilde{C}(\theta)$ has to be injective and well-conditioned. Many authors assume that the parameter θ is a linear combination of the variable $\tilde{C}$. Therefore, the parametrization is defined by

$$P\tilde{C} = \theta \qquad \text{with } P \in \mathbb{R}^{n_s \times m_s}, \tilde{C} \in M \subset \mathbb{R}^{m_s} \text{ and } \theta \in \mathbb{R}^{n_s}\,.$$

Thereby every θ determines a $\tilde{C} \in M$ uniquely. The state of the chemical reaction system evolves close to or on the low-dimensional manifold. Hence, an approximation $\tilde{C}$ of the state C of the system can be characterized by the low-dimensional parameter θ, which is called the local coordinates or the reduced set of variables. Therefore, the differential equation (2.4) can be replaced by a low-dimensional ODE for the local coordinates θ (e.g., the following PhD theses [13, 98, 115]). However, the full trajectory of the state of the system has to be computed from the solution trajectory of θ after solving the low-

dimensional ODE. Thereby the computation of the source term of the low-dimensional ODE as well as the computation of $\tilde{C}(\theta)$ depend on quantities with dimension $m_s \gg n_s$. Thus, the corresponding data are precomputed and stored in look-up tables. The computation and the storage of the look-up table is only realizable for a very low dimension of the considered manifold. If the computation and the storage can be executed, then the simulation of the chemical reaction system by a mechanism with reduced dimension is very cheap.

Remark 3.2. *A closed, homogeneous chemical reaction system has several conserved quantities. Every conserved quantity can be used in order to generate an additional algebraic equation. E.g., conservation of elements holds. The amount of each element is not changed by chemical reactions. Hence, conservation of elements is characterized by a vector v_c that fulfils*

$$R^T v_c \stackrel{!}{=} 0 \text{ with the stoichiometric matrix } R\,.$$

Note that conserved quantities are connected to zero eigenvalues in equation (3.7)*. Hence, an eigendecomposition of the Jacobian of the chemical source term also detects conserved quantities. The following algebraic equation holds:*

$$v_c^T C \stackrel{!}{=} const\,. \tag{3.9}$$

Equation (3.9) *gives additional constraints for the state of the system, which can be used in addition to PEA and QSSA. Thus, the existence of conserved quantities reduces the dimension of the low-dimensional manifold $M = \{\tilde{C}(\theta) \in \mathbb{R}^{m_s} \;:\; \theta \in \mathbb{R}^{n_s}\}$. Therefore, the dimension of the reduced set of variables is decreased, and the applicability of a look-up table is increased (curse of dimensionality). Note that similar algebraic equations are not used for an inhomogeneous reaction system. E.g., consider the conservation of elements in an arbitrary domain in the inhomogeneous chemical reactor. Then molecules pass the boundary of the domain by advection and diffusion. Thus, the amount of each element changes in the considered domain, and it is not conserved. If the chosen domain is around a spatial gridpoint, it follows that each conserved quantity does not result in an algebraic equation for each spatial grid point but in an algebraic equation for the full system. Therefore, conserved quantities produce an insignificant number of algebraic equations for an inhomogeneous reaction system. Furthermore, open reactor systems do not have conserved quantities because molecules as well as energy pass the boundary of the chemical reactor.*

3.3 Illustration of the low-dimensional manifold

In this section, a (fictitious) chemical reaction system is used to illustrate the concept of a low-dimensional manifold. A modification of Example 3 is considered. The chemical reaction system is given by

$$\begin{aligned} A_1 &\to A_2, \qquad k_1 = 1/\epsilon, \\ A_2 &\to A_1, \qquad k_2 = 1/\epsilon, \\ A_3 &\to A_1, \qquad k_3 = 1 \end{aligned}$$

on the time interval $[0, 2]$. Thereby the stiffness parameter ϵ is varied in order to demonstrate the relation between the timescale of the fast processes and the existence of the low-dimensional manifold M. Furthermore, the initial time $t_0 \in [0, 2]$, and the initial values are not fixed. However, the initial values are restricted such that $C_3(t_0) = \exp(-t_0)$ and that the conserved quantity $C_1(t) + C_2(t) + C_3(t)$ is equal to 2. If the reaction $A_1 \longleftrightarrow A_2$ is in partial equilibrium, the concentrations of the species A_1 and A_2 are equal. With the given assumptions, the PEA of the reaction $A_1 \longleftrightarrow A_2$ results in the same trajectory for all initial values $C(t_0)$. The corresponding trajectory (with PEA for the reaction $A_1 \longleftrightarrow A_2$) is

$$\tilde{C}_1(t) = \tilde{C}_2(t) = \frac{2 - \exp(-t)}{2}, \qquad \tilde{C}_3(t) = \exp(-t) \text{ for } t \geq t_0 \,.$$

However, the PEA is only valid for $\epsilon \ll 1$. In order to illustrate the existence of a low-dimensional manifold, the concentration $C_2(t)$ is plotted for different initial values, which fulfil the given restrictions, and for different values $\epsilon \in \{0.01,\ 0.05,\ 0.15,\ 0.5\}$. The minimal ϵ is 0.01. Hence, the considered system of differential equations is not very stiff.

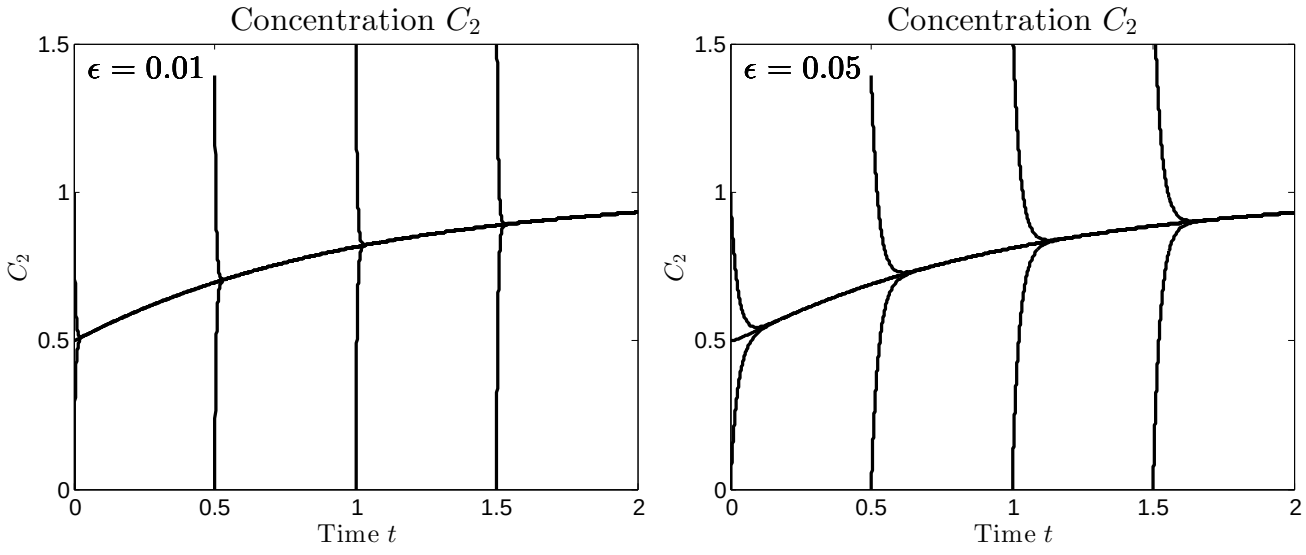

Figure 3: Concentration C_2 for different initial values and $\epsilon = 0.01$ as well as $\epsilon = 0.05$

According to Figure 3, the concentration $C_2(t)$ rapidly converges to $\tilde{C}_2(t) = \tilde{C}_1(t)$ for $\epsilon < 0.05$. Therefore, the approximation error of the PEA is small. Thus, the PEA can be used to obtain a modified differential equation with dimension $m_s - 1$. The conservation of $C_1(t) + C_2(t) + C_3(t)$ results in one additional algebraic equation. Hence, the state of the system lies close to a one-dimensional manifold. Thereby perturbations of the manifold are damped rapidly for $\epsilon < 0.05$ (see Figure 3). However, the approximation error of the reduced mechanism increases for larger parameter ϵ. According to Figure 4, the state of the system is not close to the low-dimensional manifold for a parameter ϵ that is larger than 0.15. Therefore, the quality of the reduced mechanism depends on the parameter ϵ.

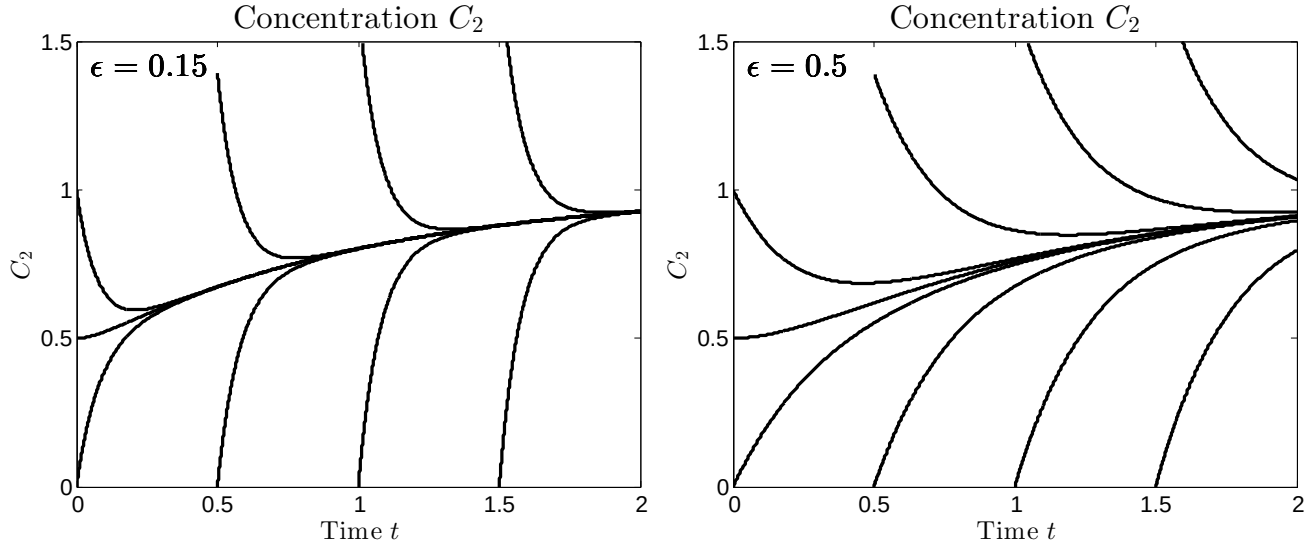

Figure 4: Concentration C_2 for different initial values and $\epsilon = 0.15$ as well as $\epsilon = 0.50$

3.4 Low-dimensional manifold and transport processes

In the previous Sections 3.1–3.3, reduction mechanisms for a homogeneous chemical reaction system are examined. However, in general chemical reactors are not spatially homogeneous. Thus, chemical reaction systems are described by equation (2.13), which includes advection and diffusion. Assume that the state of the system is described by the differential equation

$$\frac{\mathrm{d}C}{\mathrm{d}t} = G(C, \nabla C, \Delta C, T) + F(C, T). \tag{3.10}$$

Thereby $C(x,t)$ is the concentration vector, which depends on position x and time t. Moreover, G is the transport term, which includes advection and diffusion, and F is the chemical source term, which is given by $R\nu(C,T)$. If the timescales of the chemical reaction system are separated, the state of the system lies close to or on a low-dimensional manifold (see Section 3.3). The existence of the low-dimensional manifold is justified by the timescale separation of fast and slow processes, and the corresponding algebraic equations are defined by the equilibration of these fast processes. An approximation of the manifold is given by the algebraic equations (3.8), which originate from a linearization of differential equation $C' = F(C,T)$. Similar to the derivation of equation (3.8), many reduction mechanisms are based on the analysis of the chemical source term $F(C,T) = R\nu(C,T)$. Hence, the computation of the low-

dimensional manifold does often not consider transport processes. Thus, all transport processes have to be slow processes in order to legitimate the derivation of the algebraic equations by an analysis of the chemical source term. Nevertheless, the occurrence of (slow or fast) transport processes enlarges the approximation error in the occurring algebraic equations like (3.8). The transport processes move the state of the system away from the low-dimensional manifold. However, the fast processes counteract the movement by a force in the opposed direction. The absolute value of the counteracting force is zero on the low-dimensional manifold, and it increases with a larger distance to the low-dimensional manifold. Therefore, the state of the system remains close to the manifold for large timescale separation, and it can be approximated by points on the low-dimensional manifold. Hence, the transport processes are split up into a part on the low-dimensional manifold and a part perpendicular to the low-dimensional manifold. The perpendicular part is relaxed by the fast chemical processes. The other part moves the state of the system within the manifold because it couples with the slow chemical reactions. Thus, the transport term $G(C, \nabla C, \Delta C)$ is projected onto the low-dimensional manifold in order to model the contribution of the transport processes to the modified differential equation with reduced dimension. An example of the detailed treatment of transport processes for reduction mechanisms is given in [13, 89, 98, 115], which examine the Intrinsic Low Dimensional Manifold method (ILDM method).

3.5 Recommended reduction mechanisms for chemical reaction systems

In this section, some reduction mechanisms for chemical reaction systems are introduced. The selection of these mechanisms is based on personal preferences and does not claim completeness.

3.5.1 Intrinsic Low-Dimensional Manifold method

The Intrinsic Low-Dimensional Manifold (ILDM) method [90, 91, 92] by Maas and Pope is an approach for the reduction of the number of unknown variables of the differential equation. For illustration of the method, a homogeneous reactor, which is modelled by an ODE, is considered. The chemical system is modelled by equation (2.4). We assume that the system has constant volume V and constant temperature T. The chemical source term is denoted by $F(C,T) := R\nu(C,t)$. Firstly, a Schur-decomposition of the Jacobian $F_C(C,T)$ is used in order to identify the low-dimensional manifold (see Section 3.1). The Schur-decomposition of the Jacobian matrix F_C is given by $F_C = Q\tilde{T}Q^T$ with the orthogonal matrix Q and the triangular matrix $\tilde{T}$. Assume that the eigenvalues are negative. Furthermore, they are sorted in a descending order on the diagonal of $\tilde{T}$. Then it holds $Q = [Q_s, Q_f]$ (compare Remark 3.1). Thereby Q_f defines the equilibrium reaction subspace, which contains the fast processes. The equilibrium of the fast processes restricts the state of the system. These restrictions are given by

$$Q_f^T F(\tilde{C}, T) = 0\,.$$

Additional algebraic equations can be obtained due to conserved quantities like conservation of elements. The algebraic equations reduce the degrees of freedom. Hence, the equilibrium of the fast processes and the conservation of some quantities define a manifold with dimension $n_s \ll m_s$, and the state of the system is forced close to or onto this manifold. Then a differential equation for the local coordinates θ, which are a reduced set of variables, is generated. This reduced set of variables defines an approximation

$\tilde{C}$, which lays on the low-dimensional manifold, of the variable C. One obtains a differential equation for n_s variables that describe the movement on the manifold. In order to save computation time the obtained low-dimensional differential equation should not depend on high-dimensional (m_s-dimensional) quantities. In general the computation of the right hand side of this reduced differential equation involves the full chemical source term $F(\tilde{C}, T) \in \mathbb{R}^{m_s}$. Hence, the evaluation of the right hand side depends on a high-dimensional quantity, and it is computationally expensive. In order to prevent the evaluation of high-dimensional quantities in the numerical integration of the reduced differential equation, the right hand side of the reduced differential equation as well as $\tilde{C}(\theta)$ are precomputed for a mesh over the accessible domain of the reduced set of variables. Due to the low dimension the memory requirement of the look-up table is feasible. Therefore, the right hand side of the reduced differential equation can be looked up in a table, and the evaluation is independent of any high-dimensional quantities. Thus, the corresponding differential equation can be solved cheaply. Afterwards the approximation $\tilde{C}$ of the full solution C can be ontained by the look-up table cheaply. This approach relies on the feasibility of a look-up table. Otherwise, the low-dimensional differential equation depends on high-dimensional quantities, and the decrease in computing time is not very large. Although, if a look-up table is feasible (the manifold has a very small dimension), the ILDM method leads to a great saving of computing time.

3.5.2 Differential-Algebraic Equation and thermodynamics

In the PhD thesis [51], changes due to fast processes are modelled by algebraic equations. For simplicity the main idea is illustrated for a chemical reactor with constant temperature T and constant volume V. The state of the system is described by equation (2.2). The reaction space spanned by R can be decomposed into two orthogonal subspaces. Similar to the ILDM method, the fast subspace is spanned by the matrix Q_f, and the slow subspace is spanned by the matrix Q_s. Then equation (2.2) can be transformed to

$$n(t) = n_0 + Q_s y_s(t) + Q_f y_f(t)\,.$$

Thereby y_s and y_f are the extent of slow and fast processes. Multiplication of (2.4) by the volume V results in

$$n' = VR\nu(n/V, T) = Q_s y_s' + Q_f y_f'\,.$$

Further transformation yields

$$\begin{aligned} y_s' &= VQ_s^T R\nu(n/V, T), \\ y_f' &= VQ_f^T R\nu(n/V, T)\,. \end{aligned}$$

All processes in the subspace spanned by Q_f are very fast processes. If the corresponding timescales are smaller than the step size of the numerical integration method, the differential equation for y_f can be

replaced by an algebraic equation. Define $\tilde{n}_s(t) := n_0 + Q_s y_s(t)$. Then an approximation $\tilde{n}$ of the state of the system can be computed by

$$\begin{aligned} \tilde{n}_s' &= V(I - Q_f Q_f^T) R\nu\left(\left(\tilde{n}_s + Q_f y_f\right)/V, T\right), \\ 0 &= Q_f^T R\nu\left(\left(\tilde{n}_s + Q_f y_f\right)/V, T\right), \\ \tilde{n}_s(0) &= n_0, \\ \tilde{n} &= \tilde{n}_s + Q_f y_f \end{aligned} \tag{3.11}$$

or

$$\begin{aligned} \tilde{n}' &= V(I - Q_f Q_f^T) R\nu\left(\tilde{n}/V, T\right) + Q_f x_f, \\ 0 &= Q_f^T R\nu\left(\tilde{n}/V, T\right), \\ \tilde{n}(0) &= n_0\,. \end{aligned} \tag{3.12}$$

Equation (3.11) is a differential-algebraic equation with differential index 1, and equation (3.12) is a differential-algebraic equation with differential index 2. The advantage of this approach is that it improves the convergence of the Newton method for the arising nonlinear equation systems. A further aspect, which is discussed in [51], is that unknown reaction rate constants can be substituted by the thermodynamic description of the partial equilibrium (further discussion in Section 5.3). However, fast processes due to species in quasi-steady state still require the unknown reaction rate constants (further discussion in Section 5.3).

3.5.3 Computational Singular Perturbation method

The Computational Singular Perturbation (CSP) [77, 78, 132] approach provides an iterative method for the computation of a basis of the fast reaction subspace. Thereby the initial basis can be computed with an eigendecomposition. Assume that a basis $\tilde{V} = \left(\tilde{v}_1, ..., \tilde{v}_{m_s}\right)$ of $\mathbb{R}^{m_s}$ (e.g., the eigenvector matrix V_D) and its inverse $\tilde{W} := \tilde{V}^{-1}$ with row vectors $\tilde{w}_i$ are given. Furthermore, for simplicity equation (2.4) is considered, and the chemical source term is denoted by $F(C, T)$. Then the chemical source term can be represented in the basis $\tilde{V}$

$$F(C,T) = \sum_{i=1}^{m_s} \tilde{v}_i f_i \qquad \text{with } f_i = \tilde{w}_i F(C,T)\,.$$

The time derivative of f_i is given by

$$\frac{\mathrm{d}f_i}{\mathrm{d}t} = \sum_{j=1}^{m_s} \Lambda_{ij} f_j \qquad \text{with } \Lambda_{ij} = \left(\frac{\mathrm{d}\tilde{w}_i}{\mathrm{d}t} + \tilde{w}_i F_C\right)\tilde{v}_j\,.$$

The different modes are decoupled if the matrix Λ is a diagonal matrix. Furthermore, if the matrix Λ is a block-diagonal matrix, the first block contains all fast modes, and the second block all slow modes,

fast and slow processes are decoupled. Therefore, the choice of the basis $\tilde{V}$ is crucial for decoupling of different timescales. Note that other approaches like the ILDM method use the eigendecomposition of the Jacobian F_C in order to decouple the different modes. However, the eigenvalues and eigenvectors are time-dependent. Thus, the different modes are not decoupled which results in an approximation error of the reduced model. In comparison to the previous methods the CSP method contains a refinement strategy of the corresponding basis. Starting with an initial basis (for example, the eigendecomposition) the basis can be improved. Hence, a threshold is defined, and the CSP basis is updated until the error due to equilibrating the fast chemical processes is below the threshold. Afterwards the obtained basis can be used for the reduction of stiffness [78] or the reduction of the dimension of (2.4).

3.5.4 Flamelet-Generated Manifold method

The Flamelet-Generated Manifold (FGM) method [136] combines the main ideas of a manifold method (e.g., ILDM method) and of a flamelet approach [26]. In a manifold method it is assumed that the state of the system evolves on a low-dimensional manifold, which is precomputed in order to reduce the necessary computing time. In a flamelet approach a multi-dimensional flame is considered as a collection of one-dimensional flames. Hence, chemistry of one-dimensional flames can be used to generate and tabulate a low-dimensional manifold. Similar to the ILDM method, a look-up table is precomputed, and a small equation system for the control variables has to be solved instead of a large nonlinear system of differential equations. Moreover, solving the small system is independent of the high dimension of the full system. Note that transport processes like convection and diffusion exist in one-dimensional flames. Hence, the Flamelet-Generated Manifold gives a better approximation of the influence of transport on the low-dimensional manifold in comparison to the ILDM method. In particular, in regions with a low temperature the obtained results are improved because convection and diffusion provide essential contributions in case of low temperature.

3.5.5 Reaction-Diffusion Manifold method

Most reduction mechanisms for chemical reaction systems examine the chemical source term in order to determine the fast and the slow subspaces. However, transport processes may have an influence on the slow manifold. Furthermore, the transport term has to be modified in the reduced model. The REaction-DIffusion Manifold (REDIM) method [21] considers the chemical source term as well as the transport term for the computation of the low-dimensional manifold. For simplicity in this explanation it is assumed that the temperature and the volume are constant in the chemical reactor and that the state of the system is defined by equation (2.13). The right hand side of equation (2.13) includes the chemical reaction term $F(C)$ and the transport term $G(C, \nabla C, \Delta C)$. Assume that the state of the system is close to a low-dimensional manifold M with dimension $n_s \ll m_s$ and that the manifold is parameterized by a parameter θ such that $M = \{\tilde{C}(\theta) \in \mathbb{R}^{m_s} \ : \ \theta \in \mathbb{R}^{n_s}\}$. Thereby the paramter θ is the local coordinates

or the reduced set of variables. The manifold M is invariant if the right hand side $F(\tilde{C}) + G(\tilde{C}, \nabla\tilde{C}, \Delta\tilde{C})$ belongs to the tangent space of M. Thus, the invariant manifold is defined by the invariance condition

$$\left(I - \tilde{C}_\theta \tilde{C}_\theta^+\right)\left(F\left(\tilde{C}(\theta)\right) + G\left(\tilde{C}(\theta), \nabla\tilde{C}(\theta), \Delta\tilde{C}(\theta)\right)\right) = 0\,. \tag{3.13}$$

Thereby $\tilde{C}_\theta$ is the derivative with respect to the parameter θ and $\tilde{C}_\theta^+$ is the Moore-Penrose pseudoinverse of $\tilde{C}_\theta$ [104]. Note that equation (3.13) describes a low-dimensional manifold by the invariance of the full right hand side. Thus, it is taken account of the transport processes $G(\tilde{C}, \nabla\tilde{C}, \Delta\tilde{C})$. In order to solve equation (3.13) the stationary solution of the following PDE is computed:

$$\tilde{C}(\theta)' = \left(I - \tilde{C}_\theta \tilde{C}_\theta^+\right)\left(F\left(\tilde{C}(\theta)\right) + G\left(\tilde{C}(\theta), \nabla\tilde{C}(\theta), \Delta\tilde{C}(\theta)\right)\right).$$

Therefore, in opposition to other approaches the invariant manifold M of the REDIM method captures the system dynamics of (2.13) including the transport processes.

3.5.6 Global Quasi-Linearisation method

Another obstacle of the mentioned methods is the local character of the manifold. The low-dimensional manifold of many manifold methods is obtained by an eigenvalue analysis of the Jacobian of the chemical source term. Hence, a linearisation at one specific state is considered. However, the state of the system (including concentrations and temperature) can change in time and space. Therefore, the validity of the eigenvalue analysis is limited. In opposition to local methods the Global Quasi-Linearisation (GQL) [19] identifies a fast/slow decomposition globally. For simplicity the approach is introduced with a spatially homogeneous chemical reactor. The corresponding differential equation is (2.4). Assume that the conserved quantities have been removed. Thus, it is possible to find m_s different concentrations $C^{(i)}$, $1 \leq i \leq m_s$, such that the vectors $F(C^{(i)})$ are linearly independent. The concentration vectors $C^{(i)}$ are saved in a matrix $C^* := [C^{(1)}, ..., C^{(m_s)}] \in \mathbb{R}^{m_s \times m_s}$. Furthermore, a matrix $F^* = [F(C^{(1)}), ..., F(C^{(m_s)})]$ is computed, and the linear function that maps C^* onto F^* is given by

$$T^* := F^* \cdot (C^*)^{-1}\,.$$

The matrix T^* defines the GQL of the chemical reaction system (2.4). Thereafter the slow and the fast subspaces can be determined by the following decomposition of the matrix T^*:

$$T^* = \begin{pmatrix} Z_s & Z_f \end{pmatrix} \begin{pmatrix} N_s & 0 \\ 0 & N_f \end{pmatrix} \begin{pmatrix} \tilde{Z}_s \\ \tilde{Z}_f \end{pmatrix}.$$

Thereby N_s and N_f are triangular matrices, and the eigenvalues are sorted in decreasing order. The corresponding decomposition of T^* can be obtained by Schur-decomposition and solving the Sylvester equation [36]. Then Z_f defines a permanent basis of the fast subspace (compare to Q_f in Section 3.1).

Afterwards the reduced differential equation is generated similar to the ILDM approach.

There are much more methods which exploit the occurrence of different timescales (e.g., proper orthogonal decomposition (POD) [29, 38, 123] or lumping [9, 81, 82]). Furthermore, there are countless expansions of the described methods. E.g., published expansions of the ILDM method are the ILDM method extended with diffusion [16], an extension of the ILDM method to the domain of slow chemistry [20], and hierarchically extended ILDMs [75]. Thus, Section 3.5 does not provide a complete listing of reduction mechanisms for chemical reaction systems. However, a complete list is not possible and would be beyond the scope of this work.

3.6 Validity of reduction mechanisms based on singular perturbation theory

In this section, the validity of reduction mechanisms is examined. The reduction of chemical reaction systems is based on the separation of the occurring timescales. Hence, a singular perturbed problem is studied. The occurring system is considered after spatial discretization in order to simplify the analysis. Thus, an ODE describing the chemical reaction system is given. Furthermore, the following special idealization of the system is considered:

$$\begin{aligned} (z^1)' &= f(z^1, z^2, \epsilon), \\ \epsilon (z^2)' &= g(z^1, z^2, \epsilon), \qquad 0 < \epsilon \ll 1\,. \end{aligned} \tag{3.14}$$

Thereby it holds $z^1 \in \mathbb{R}^{n_s}$, $z^2 \in \mathbb{R}^{m_s - n_s}$, and the functions $f : D_{z^1} \times D_{z^2} \times D_\epsilon \to \mathbb{R}^{n_s}$ and $g : D_{z^1} \times D_{z^2} \times D_\epsilon \to \mathbb{R}^{m_s - n_s}$ as well as their derivatives are elements of $\mathcal{O}(1)$. Furthermore, all functions are as regular as needed. It follows that the variable z^1 is slowly evolving, and the variable z^2 is changing fast. The initial values are given by $z_0^1 = z^1(t_0)$ and $z_0^2 = z^2(t_0)$. In general the corresponding ODE has not the special form (3.14). However, equation (3.14) is adequate for analysis, and it covers a wide range of different timescales, which is the main property of fast chemical reaction system. The transformation of an autonomous linear ODE with timescale separation into a system (3.14) is given in [19]. The existence of a transformation of a nonstandard form of two-timescale systems into the standard form is examined in [48]. In equation (3.14) the transformation of the time variable to $\tau = t/\epsilon$ results in the system

$$\begin{aligned} \frac{\mathrm{d}z^1}{\mathrm{d}\tau} &= \epsilon f(z^1, z^2, \epsilon), \\ \frac{\mathrm{d}z^2}{\mathrm{d}\tau} &= g(z^1, z^2, \epsilon), \qquad \tau_0 = t_0/\epsilon\,. \end{aligned} \tag{3.15}$$

If the parameter ϵ decreases, the timescale separation becomes larger. In the limit $\epsilon \to 0$, the separation of fast and slow processes is infinite, and the system (3.14) reduces to

$$\begin{aligned} (z^1)' &= f(z^1, z^2, \epsilon), \\ 0 &= g(z^1, z^2, \epsilon)\,. \end{aligned} \tag{3.16}$$

The reduced system (3.16) is a differential-algebraic equation, and it describes the state of the system after the fast transient. The fast variable z^2 is characterized by the slow variable z^1 and algebraic equations, which are given by the partial equilibrium of the fast dynamics. In comparison equation (3.15) converges to

$$\begin{aligned} \frac{\mathrm{d}z^1}{\mathrm{d}\tau} &= 0, \\ \frac{\mathrm{d}z^2}{\mathrm{d}\tau} &= g(z^1, z^2, \epsilon). \end{aligned} \tag{3.17}$$

The differential equation (3.17) specifies the fast transient behaviour of the fast variable z^2. At the initial time t_0 the fast processes are not in partial equilibrium. Hence, the corresponding variable z^2 changes rapidly with a timescale of size ϵ, while the slow variable z^1 is constant on that timescale.

In general ϵ is not equal to zero, but it is small. Thus, further analysis is necessary. The analysis of (3.14) is given in [48, 71, 101]. For completeness it is summarized in this section. The following assumptions are necessary for the analysis of equation (3.14):

A1 There exists a function h_0, such that $g(z^1, h_0(z^1), 0) = 0 \ \forall z^1 \in D_{z^1} \subset \mathbb{R}^{n_s}$. Furthermore, D_{z^1} is a compact domain.

A2 The eigenvalues of the Jacobian matrix $g_{z^2}(z^1, h_0(z^1), 0)$ are smaller than a constant $-c_1 < 0$ for all $z^1 \in D_{z^1}$. Therefore, $h_0(z^1_*)$ is an asymptotically stable solution of $(z^2)' = g(z^1_*, z^2, 0)$ for fixed $z^1_* \in D_{z^1}$.

A3 The solution of the differential equation $(z^2)' = g(z^1_0, z^2, 0),\ z^2(t_0) = z^2_0$ converges towards the steady state $h_0(z^1_0)$.

A4 The reduced differential equation

$$(z^1)' = f(z^1, h_0(z^1), 0), \qquad z^1(t_0) = z^1_0 \tag{3.18}$$

has a solution $Z^1(t) \in D_{z^1}$ for $t \in [t_0, t_0 + t_e],\ t_e > 0$.

Theorem 3.3 (Nipp [101])**.** *Let the previous assumptions **A1**-**A4** be fulfilled. Then for every $\delta > 0$ there is a constant $\epsilon_\delta > 0$ such that for $\epsilon \le \epsilon_\delta \in D_\epsilon$ the solution $z^1(t, \epsilon)$, $z^2(t, \epsilon)$ of differential equation* (3.14) *exists for $t \in [t_0, t_0 + t_e]$, and for a constant $c_2 > 0$ it holds*

$$\begin{aligned} \|z^1(t, \epsilon) - Z^1(t)\| &\le c_2 \epsilon \qquad \text{for } t \in [t_0, t_0 + t_e], \\ \|z^2(t, \epsilon) - Z^2(t)\| &\le c_2 \epsilon \qquad \text{for } t \in [t_0 + \delta, t_0 + t_e]. \end{aligned} \tag{3.19}$$

Thereby $Z^1(t)$ is the solution of equation (3.18)*, and $Z^2(t)$ is defined by $Z^2(t) = h_0(Z^1(t))$.*

Note that $Z^2(t)$ is defined by the equilibrium of the fast processes. Thus, Theorem 3.3 results in an error bound for a reduction mechanism of equation (3.14). However, the special form of equation (3.14) provides the splitting in fast and slow processes. In general an important part of each reduction

mechanism is the determination of the fast processes.
Furthermore, if assumptions **A1** and **A2** hold, there exists a reduced manifold that is invariant under the dynamics of (3.14) and is ϵ-close to the manifold $M_0 = \{(z^1, z^2) \in \mathbb{R}^{m_s} \ : \ z^2 = h_0(z^1),\ z^1 \in D_{z^1}\}$.

Theorem 3.4 (Kaper [71])**.** *Let assumptions **A1** and **A2** be fulfilled. For any sufficiently small $\epsilon > 0$ a manifold M_ϵ that is locally invariant under the dynamics of* (3.14) *exists. Thereby the manifold M_ϵ is given by*

$$M_\epsilon = \{(z^1, z^2) \in \mathbb{R}^{m_s} \ : \ z^2 = h_\epsilon(x), z^1 \in D_{z^1}\},$$

where the function h_ϵ has the expansion

$$h_\epsilon(z^1) = h_0(z^1) + \epsilon h_1(z^1) + \epsilon^2 h_2(z^1) + \mathcal{O}(\epsilon^3) \text{ as } \epsilon \to 0\,.$$

A recursive formula for the coefficient functions $h_i,\ i \in \mathbb{N}_0$, is given in [71].

Hence, due to Theorem 3.3 the solution of (3.14) is approximated by a solution of the reduced system (3.16) after a short transient phase. The approximation error is in $\mathcal{O}(\epsilon)$. Moreover, according to Theorem 3.4, for suited initial values the solution of (3.14) lays on a low-dimensional manifold M_ϵ, which is defined by $h_\epsilon(z^1)$. Every reduction mechanism of a chemical reaction system defines a reduced system, whose solution lays on an approximation M_{app} of this low-dimensional manifold M_ϵ. The total approximation error between the solution of the reduced mechanism and the solution of the full system is caused by the approximation error $dist(M_{app}, M_\epsilon)$ of the low-dimensional manifold and by initial values, which do not lay on the invariant manifold. Moreover, the state of the system relaxes exponentially fast towards the invariant manifold M_ϵ (see Remark 3.5).

Remark 3.5. *[Nipp [100]] Under stricter assumptions one can prove that there are positive constants c_1, c_2, c_3 such that every solution $(z^1(t), z^2(t))$ of* (3.14) *with $|z_0^2 - h_\epsilon(z_0^1)| \leq c_2$ satisifies*

$$|z^2(t) - h_\epsilon(z^1(t))| \leq c_1 |z_0^2 - h_\epsilon(z_0^1)| e^{-c_3 t/\epsilon}\,.$$

The solution of (3.14) *is exponentially close to the manifold M_ϵ.*

4 Reduction of stiffness-induced round-off errors

Acknowledgement: This chapter is (mainly) the accepted manuscript of an article [85] by Axel Lukassen and Martin Kiehl published as the version of record in *Combustion Theory and Modelling* 2017 (copyright Taylor & Francis) on 11th August 2016, available online:

http://www.tandfonline.com/doi/full/10.1080/13647830.2016.1213427

4.1 Introduction

In this chapter, a new procedure to simplify ordinary differential equations for the modelling of chemical reaction systems is introduced. The mathematical model of a chemical reaction system is a partial differential equation. But in order to simplify the development of the method, a spatially homogeneous gas reaction system with constant temperature and constant volume is examined in this chapter. Thus, in the following the considered ordinary differential equation is

$$\begin{aligned} C' &= F(C), \\ C(0) &= C_0 \in \mathbb{R}^{m_s}. \end{aligned} \tag{4.1}$$

The timescales of chemical processes usually cover a range from 10^{-9} seconds to 10^{2} seconds. Therefore, the considered problem is a stiff system of differential equations. So, (4.1) causes problems for explicit integration schemes, and implicit integration schemes are usually used [28]. Implicit methods require the solution of a nonlinear equation system. Usually, Newton's method is used for solving these equation systems. Anyhow, if Newton's method fails due to the stiffness of the system, integration methods for computing the numerical solution of (4.1) may require very small step sizes [96]. The reason for failing of Newton's method often is the occurrence of round-off errors in the evaluation of the nonlinear chemical source term due to very fast processes (other reasons also exist). These round-off errors appear if the range of the occurring timescales gets very large [62]. This point is illustrated in Section 4.2. A straightforward technique to support convergence of Newton's method is to reduce the step size. But this results in an increased computing time.

Another possibility to avoid failure of Newton's method is to reduce the occurring round-off errors. As already stated, different timescales of the chemical system lead to the round-off errors. Therefore, the stiffness of the system results in failure of Newton's method, and reducing the stiffness supports the convergence of Newton's method. There are many methods that reduce the dimension of the system of differential equations (compare to Section 3.5). Furthermore, they eliminate stiffness of (4.1), and they reduce the round-off errors in the modified source term as a spin-off. Most of them use the quasi-steady state approximation (QSSA) [11, 97, 105, 131] or the partial equilibrium approximation (PEA) [22, 56, 57, 97, 110]. Thereby the PEA for some reactions or the QSSA for some species matches to set the corresponding reaction velocity to infinity. However, the reduction mechanisms from Section 3.5 have several drawbacks. In the following these drawbacks will be sketched. The Intrinsic Low-Dimensional Manifold method (ILDM) [92], the REaction-DIffusion Manifold method (REDIM) [21], the Global Quasi-Linearization method (GQL) [19], and the Flamelet-Generated Manifold method (FGM) [136] are well-known reduction mechanisms, which eliminate the stiffness of the ordinary dif-

ferential equation. However, the state variables have to be determined, using a reduced set of variables and a cheap way to achieve this is needed. A common procedure is to create look-up tables [90, 91]. Obviously, the dimension of a look-up table should not be too large, therefore, the trajectory of the state variable has to lay on a low-dimensional manifold (usually between one and three dimensions). This request has several major drawbacks. First, such a low-dimensional manifold has to exist. If there are only a few fast processes but a lot of slow processes, the dimension of the manifold gets too large, and these methods cannot be used anymore. Second, the dimension of the manifold has to be constant in time, and in case of a partial differential equation the dimension has to be constant in time and space. Otherwise, the worst possible (the largest) dimension has to be used. Therefore, in case of areas with slow chemistry (low temperature, mixing zone) the dimension of the manifold depends on this area, and gets too large. Some extensions of the mentioned methods exist. For example, there is an extension of the ILDM method to areas with slow chemistry [20]. This extension splits the area in an area of fast chemistry, an area with negligible chemistry, and a transition area. Then, this method requires that the transition area is insignificant. Afterwards the remaining parts can be handled separatly. An obvious drawback of the method is the requirement of a negligible transition area. Also this procedure uses the ILDM method in the area of fast chemistry, and thereby, requires only a few slow processes in this area (in order to obtain a low-dimensional look-up table). Another technique to reduce the stiffness of the considered ordinary differential equation (4.1) is introduced by Franz [51]. This method offers a time variable number of fast processes (variable dimension of the manifold). But the number of variables is enlarged by the number of fast processes locally, and therefore, this method leads to an increased number of variables in comparison to the newly introduced method, and is difficult to adapt to a partial differential equation.

In total the above-named methods have to be modified, if the dimension of the subspace, defined by the partial equilibrium of the fast processes, is variable. Besides, some of the methods require a low-dimensional manifold, namely that the subspace spanned by the slow processes is low-dimensional. On the other hand, if it is possible to use these methods, many of them reduce the number of variables drastically, and therefore, lead to a great saving of computing time.

Another method to simplify stiff differential equations is operator splitting (see Chapter 6). Operator splitting is often used for advection-diffusion-reaction systems. For that matter two main approaches are used. First, it is possible to split between transport processes and chemistry [116, 118]. This approach transforms the high-dimensional problem into a lot of low-dimensional problems, which can be solved more easily. As a result one obtains ordinary differential equations, which describe the chemical reactions, but still include the described difficulties for chemistry including processes with different timescales. Furthermore, the coupling between slow chemical processes and slow transport phenomena is ignored. Second, the splitting between stiff and non-stiff processes is frequently used. Generally, this means transport processes and slow chemistry are separated from the fast chemical processes. In case of partial differential equations, the second splitting approach can generate close-by nodes with a discontinuous source term. Moreover, a splitting error is introduced. Due to the stiffness of the differential equation, this splitting error cannot be estimated with the classical analysis of the splitting error [121] (detailed discussion in Chapter 6).

Further recommended methods for reducing the stiffness of the system of differential equations are the Computational Singular Perturbation method (CSP) [78, 132] and the G-scheme [133, 134]. The CSP by Lam and Goussis was developed in order to enlarge the comprehension of the chemical system by simplifying the chemical source term and to reduce the stiffness of the system. However, if the slow and fast processes are changing in space, the CSP method can generate close-by nodes with a discontinuous chemical source term. In this case the solution is not continuous in space. However, the error (thus, the discontinuity) will be small. The G-scheme by Valorani and Paolucci is an adaptive model reduction method, which was developed for systems of ordinary differential equations. By using the method of lines it is also applicable to partial differential equations. The G-scheme results in a substantially smaller and non-stiff system of differential equations. Thereby, the G-scheme decomposes the tangent space as the sum of four subspaces, namely the active subspace, the invariant subspace, the fast and the slow subspace. The classification of the subspaces is done by an eigenvalue analysis of the Jacobian of the source term. Hence, for partial differential equations the eigendecomposition of a very large matrix is required.

In order to avoid the described drawbacks of all these methods a new approach is used. The proposed new method reduces the range of the occurring timescales, and thereby reduces the occurring round-off errors. In contrast to the QSSA and the PEA the reaction velocity of the fast processes is not set to infinity but is decreased. Thereby it is used that the exact reaction velocity of very fast processes is not important. It is important that the assumed reaction velocity guarantees partial equilibrium as far as the tolerance in the shortest considered time interval of interest. Thus, the reaction velocity of very fast processes with a timescale shorter than the minimal examined time period can be reduced until the timescale of the process is equal to the shortest considered time span. The stiffness of the ordinary differential equation is connected with fast processes in partial equilibrium. Evaluation of the chemical source term requests subtraction of very large numbers due to these fast processes, and therefore, stiffness leads to large round-off errors due to cancellation. These round-off errors can lead to failure of Newton's method. Then, smaller step sizes are required, and thus, the computing time is increased. The described approach reduces the stiffness of the system, and consequently, reduces the occurring round-off errors. This avoids failure of Newton's method, and hence, leads to a reduction of the computing time. Additionally, the dimension of the modified differential equation is not changed, and therefore, the procedure can easily be adapted for partial differential equations with a spatial changing number of fast processes. Furthermore, this method does not depend on the existence of a low-dimensional manifold that contains the trajectories of the state variables after a short transient phase, and thereby, the technique is also applicable in case of a large number of remaining slow processes.

The structure of this chapter is the following. First, in Section 4.2 it is shown that very stiff differential equations lead to implicit equation systems, which may cause numerical problems due to round-off errors. Furthermore, it is shown that the reason for these round-off errors are the short timescales of the very fast processes. Then, a new method to reduce this effect is described and analyzed for linear differential equations. Finally, in Section 4.4 a numerical example is considered. This example shows that the introduced method is well suited for solving chemical reaction systems, which result in very stiff differential equations.

4.2 New method

4.2.1 Motivation

As already discussed, the differential equation (4.1) that describes a chemical reaction system often is a stiff differential equation. Hence, implicit methods are usually used to compute a numerical solution of (4.1). If an implicit integration method is used, an equation system has to be solved in each time step. The equation system of implicit Runge-Kutta methods is usually solved with Newton's method. Therefore, the Jacobian matrix of the chemical source term is needed in order to solve the equation system. If the Jacobian matrix is not available, the Jacobian matrix is computed numerically. Thus, round-off errors in the computer evaluation of the chemical source term result in a perturbed Jacobian matrix. This might cause step size reductions in the integration method [96]. Furthermore, if an analytical expression of the Jacobian matrix of the chemical source term is available, round-off errors in the computation of the chemical source term can cause convergence problems and step size reductions. This point is illustrated for the implicit Euler scheme applied to (4.1). Thereby a numerical approximation C^i of the concentration $C(t_i) = C(t_0 + ih)$ for a given step size h is computed. After the initialization

$$t_0 = 0, \qquad C^0 := C(t_0) \tag{4.2}$$

the implicit Euler scheme is given by

$$\begin{aligned} t_{i+1} &= t_i + h, \\ C^{i+1} &= C^i + hF(C^{i+1}). \end{aligned} \tag{4.3}$$

Within each integration step the nonlinear equation system

$$g_*(C^{i+1}) := C^{i+1} - C^i - hF(C^{i+1}) \stackrel{!}{=} 0 \tag{4.4}$$

has to be solved. If the simplified Newton's method is used to solve (4.4), each Newton iteration reads

$$[I_{m_s} - hF_C(C_{(0)}^{i+1})]\Delta C_{(j)}^{i+1} = g_*(C_{(j)}^{i+1}).$$

Thereby I_{m_s} is the $(m_s \times m_s)$-identity matrix, the index (j) in $C_{(j)}^{i+1}$ is the counter of the Newton iteration, and $\Delta C_{(j)}^{i+1}$ is the Newton step. After solving this linear equation for $\Delta C_{(j)}^{i+1}$ the next Newton iterate follows at once. It holds

$$C_{(j+1)}^{i+1} = C_{(j)}^{i+1} - \Delta C_{(j)}^{i+1}.$$

Note that damped Newton methods are unusual in the context of integration methods, as convergence problems are avoided by a step size reduction in (4.3). A chemical reaction system usually holds some conservation laws like conservation of mass, conservation of energy or conservation of elements. These

conserved quantities are connected to the eigenvalues λ_k equal to zero of the Jacobian matrix $F_C(\cdot)$. On the other hand, fast processes with a short timescale are connected to very small negative eigenvalues $\lambda_k \ll 0$. Furthermore, slow processes with a large timescale result in a large integration time period, and after a short transient phase these slow processes determine the used step sizes h, thus, the step size h becomes very large. Therefore, the matrix $[I_{m_s} - hF_C(C_{(0)}^{i+1})]$ is an ill-conditioned matrix with eigenvalues in $[1, 1+h\max_k|\lambda_k|]$. As already mentioned, the evaluation of $g_*(C_{(j)}^{i+1})$ requests subtraction and addition of very large numbers due to the fast processes, which results in round-off errors by cancellation. The effect of these round-off errors on the Newton correction is damped in directions of fast processes, but it is retained in directions of slow processes or conserved quantities. This can lead to large relative errors in $\Delta C_{(j)}^{i+1}$, and hence might cause non-convergence of Newton's method.

In Section 3.1, the temporal development of a perturbation of the stationary state C_∞ is examined. Thereby the different modes are decoupled, and the ODE (3.7) is derived. In equation (3.7) the absolute value of each eigenvector component $w_k(t) = \left(V_D^{-1}\right)_{(k,\cdot)} \cdot (C(t) - C_\infty)$ develops like $e^{\lambda_k \cdot t}$. Note that D is a diagonal matrix with eigenvalues λ_k at position D_{kk}, and $v_{D,k}$ is the corresponding eigenvector. Thus, components corresponding to negative eigenvalues with large absolute values decay very fast. If a component's "time of decay" is much smaller than the used time step h, the system instantaneously reaches equilibrium in the direction of the eigenvector $v_{D,k}$ within the required tolerance. Therefore, there exists a bound λ_g such that for all $\lambda_k \leq \lambda_g \ll 0$ and all $t \geq h$ it holds

$$0 < |w_k(t)| = |w_k(0)| \cdot e^{\lambda_k \cdot t} \leq |w_k(0)| \cdot e^{\lambda_g \cdot t} \leq |w_k(0)| \cdot e^{\lambda_g \cdot h} \leq Tol\,. \tag{4.5}$$

This equation shows that for very fast processes the absolute value of λ_k does not matter as long as $\lambda_k \leq \lambda_g$. It is possible to replace every very small negative eigenvalue λ_k by λ_g, and at the same time not to change the solution more than the required tolerance. Thus, it is possible to modify the differential equation (3.7) such that the maximal absolute value of the eigenvalues decreases drastically, and thereby to improve the solvability of the problem.

Remark 4.1. *Fixed-point iteration is usually used for multistep methods. Therefore, multistep methods are not considered in this chapter.*

4.2.2 The modified problem

In order to replace every small eigenvalue λ_k by λ_g, the equation (3.7) can be modified as

$$\tilde{w}' = \mathrm{diag}_{k \leq m_s}(a_k) \cdot D \cdot \tilde{w}, \qquad a_k = \begin{cases} 1 & \lambda_k > \lambda_g \\ \frac{\lambda_g}{\lambda_k} & \lambda_k \leq \lambda_g \end{cases}. \tag{4.6}$$

In the case $\lambda_k \ll \lambda_g$, the maximal absolute value of the eigenvalues is reduced drastically. An arbitrary linear differential equation

$$C' = A \cdot C \tag{4.7}$$

can be modified to

$$\tilde{C}' = V_D \cdot \mathrm{diag}_{k \leq m_s}(a_k) \cdot D \cdot V_D^{-1} \cdot \tilde{C}\,. \tag{4.8}$$

And the nonlinear differential equation (4.1) can be modified using the eigendecomposition of the Jacobian matrix F_C. It follows

$$\tilde{C}' = V_D \cdot \mathrm{diag}(a_k) \cdot V_D^{-1} \cdot F(\tilde{C}) =: \tilde{F}(\tilde{C})\,. \tag{4.9}$$

However, this modification requires the inverse of the eigenvector matrix V_D, which leads to several drawbacks. First, all eigenvectors of the Jacobian matrix have to be computed in order to determine the inverse of the eigenvector matrix. Thereby the evaluation of the eigenvector matrix V_D and its inverse are two computationally expensive operations. Second, the eigenvector matrix can be ill-conditioned. An ill-conditioned matrix results in large computation errors for the inverse matrix V_D^{-1}. In order to avoid the drawbacks of the eigendecomposition, the Schur-decomposition can be used. The Schur-decomposition of a matrix F_C is given by

$$F_C = Q \cdot \tilde{T} \cdot Q^T \in \mathbb{R}^{m_s \times m_s}$$

In case of real eigenvalues, $\tilde{T}$ is an upper triangular matrix, and the diagonal elements of the matrix $\tilde{T}$ are the eigenvalues of the matrix F_C. In case of complex pairs of eigenvalues, the eigenvalues appear as two by two blocks in $\tilde{T}$. Furthermore, Q is an orthogonal matrix. Among others, Franz [51] as well as Pope and Maas [92] use the Schur-decomposition instead of the eigendecomposition of the Jacobian matrix $F_C(\cdot)$. The eigenvalues of $\tilde{T}$ can be placed on the diagonal in any order. Without loss of generality the Schur-decomposition is ordered such that the eigenvalues are increasing on the diagonal of $\tilde{T}$. Then, the Schur-decomposition has the representation

$$\begin{aligned} A &= Q \cdot \tilde{T} \cdot Q^T = \begin{pmatrix} Q_f & Q_s \end{pmatrix} \cdot \tilde{T} \cdot \begin{pmatrix} Q_f & Q_s \end{pmatrix}^T, \\ Q_s &\in \mathbb{R}^{m_s \times n_s}, \qquad Q_f \in \mathbb{R}^{m_s \times (m_s - n_s)}. \end{aligned}$$

Thereby Q_f is a basis of the $(m_s - n_s)$-dimensional space spanned by the eigenvectors corresponding to eigenvalues smaller than $\lambda_g \ll 0$. If the dimension of the subspace spanning the space of fast processes is small, the column vectors of Q_f can be computed with power iteration effectively. We define $n_f := m_s - n_s$. Then problem (4.1) can be modified to

$$\begin{aligned} \tilde{C}' &= \left[I_{m_s} - Q_f \cdot \mathrm{diag}_{k \leq n_f} \left(1 - \frac{\lambda_g}{\tilde{T}_{kk}} \right) \cdot Q_f^T \right] \cdot F(\tilde{C}), \\ &\mathrm{diag}_{k \leq n_f} \left(1 - \frac{\lambda_g}{\tilde{T}_{kk}} \right) \in \mathbb{R}^{n_f \times n_f}. \end{aligned} \tag{4.10}$$

Remark 4.2. *The modification of the differential equation requires a decomposition of the Jacobian matrix F_C. As already mentioned, the numerically computed Jacobian matrix is perturbed by round-off errors. However, the relative perturbations of the large eigenvalues are not severe [10]. Furthermore, the perturbation of a subspace spanned by some eigenvectors decreases with a larger gap between the corresponding eigenvalues and the other eigenvalues [3, 30]. This gap often increases with the magnitude of the considered eigenvalues. Therefore, the subspace spanned by fast processes usually is not perturbed seriously.*

We will show for linear differential equations that the modification of the system reduces the round-off error in the evaluation of the source term. Nevertheless, the accuracy of the method is not improved. The accuracy is controlled by the error control of the numerical method. If the error is too large, the step size is reduced. However, if errors in the evaluation of the chemical source term are too large, Newton's method can fail. Then, additional step size reductions are needed, and additional computational effort is necessary. The advantage of the modified problem is the reduction of these round-off errors, and hence, avoiding failure of Newton's method. Thus, if Newton's method converges, the modification of the system of differential equations does not result in any speed-up. On the contrary, the modification needs additional computing time, and it introduces an approximation error. Therefore, the differential equation (4.1) is only modified if Newton's method does not converge and the initial transient phase is over. Hence, the purpose of the modification is not to generate a smaller total error in the solution of (4.1). The purpose is solely to reduce the necessary number of integration steps. In summary the modified problem introduces an additional approximation error, which is controlled by the choice of λ_g, and reduces the round-off error in the source term, which prevents failure of Newton's method, and thus, allows larger step sizes.

4.3 Analysis

In this section, the influence of the described modification on the round-off error of the source term is analyzed. Large round-off errors in the evaluation of the source term $F(C)$ occur for all integration methods. Therefore, it is possible to consider only the round-off error of the chemical source term $F(C)$. However, if the integration method is fixed and the exact Jacobian matrix F_C is given, the influence of the round-off error on the Newton correction can be examined, and thus, the influence on the state of the system. For example, if the implicit Euler scheme and simplified Newton's method is used to solve the equation system (4.1), one obtains

$$\begin{aligned}\left[I_{m_s} - hF_C(C_{(0)}^{i+1})\right]\Delta C_{(j)}^{i+1} &= g_*(C_{(j)}^{i+1}) \\ \Longrightarrow A_* \cdot \Delta C_{(j)} &= g_* \text{ with } A_* := \left[I_{m_s} - hF_C(C_{(0)}^{i+1})\right].\end{aligned}$$

The absolute round-off error in $\tilde{g}$ (see (4.4)) is approximately equal to the round-off error in $(h \cdot F)$. Furthermore, the eigenvalues corresponding to conserved quantities are equal to zero. Thus, in the computed $\Delta C_{(j)}$ this round-off error can be amplified by a factor

$$\|A_*^{-1}\|_2 \geq \frac{1}{1 + h \min_i |\lambda_i|} \tag{4.11}$$

Hence, if a conserved quantity with an eigenvalue equal to 0 exists, a round-off error in the computer evaluation of F can be amplified by $1 \cdot h$ in the computer evaluation of the Newton correction $\Delta C_{(j)}$. Therefore, the Newton correction is perturbed seriously, which can result in the non-convergence of Newton's method. In order to reduce the difference between the exact Newton correction and the computed Newton correction, the modified differential equation (4.9) or (4.10) is used, and thus, the round-off error in F is decreased. In the following terms $(\cdot)$ that originate from the modification are marked as $(\tilde{\cdot})$, and the computer evaluation of a function is marked by $(\cdot)^{\epsilon}$. Then, the modified differential equation is

$$\tilde{C}' = \tilde{F}(\tilde{C}) .$$

Using the implicite Euler scheme and Newton's method for this modified differential equation provides the following implicit equation for the modified Newton correction:

$$\tilde{g}_*\left(\tilde{C}^{i+1}\right) = \tilde{C}^{i+1} - \tilde{C}^{i} - h \cdot \tilde{F}\left(\tilde{C}^{i+1}\right) \stackrel{!}{=} 0 .$$

Use of Newton's method results in

$$\begin{aligned} \left[I_{m_s} - h \cdot \tilde{F}_C\left(\tilde{C}_{(0)}^{i+1}\right)\right] \cdot \Delta\tilde{C}_{(j)}^{i+1} &= \tilde{g}_*\left(\tilde{C}_{(j)}^{i+1}\right), \\ \Longrightarrow \tilde{A}_* \cdot \Delta\tilde{C}_{(j)} &= \tilde{g}_* . \end{aligned}$$

Note that the spectral norm of the matrix $(\tilde{A}_*)^{-1}$ is equal to the spectral norm of A_*^{-1}. In the further progress, it is assumed that all eigenvalues of the matrix F_C are not positive. In case of a permanent positive eigenvalue λ_k of the matrix F_C the corresponding component $(V_D^{-1} \cdot C)_k$ increases without any limit. Therefore, the system does not reach any steady state. Given that the described problems of non-convergence of Newton's method occurs for very large step sizes, this case can be neglected.

4.3.1 Linear functions

As first step of the analysis, a linear differential equation (4.7) is considered. Hence, an upper bound for the round-off error in the computer evaluation of a linear function is generated.

Remark 4.3. *According to [37, 129], the absolute condition number according to the spectral norm can be estimated by*

$$\max_k |\lambda_k| \leq \max_k \sigma_k = \|A\|_2 =: \kappa_{A,abs,2} .$$

Thereby σ_k, $k = 1, 2, ..,$ are the singular values of the matrix A. Furthermore, a normal matrix A fulfils

$$\max_k |\lambda_k| = \max_k \sigma_k .$$

Lemma 4.4. *Let*

$$F(C) = A \cdot C \qquad \text{with } A \in \mathbb{R}^{m_s \times m_s} \text{ and } C \in \mathbb{R}^{m_s}$$

be a linear function and ϵ_m be the machine epsilon. The difference between the exact evaluation of the linear function $F(C)$ and the machine evaluation $[F(C)]^\epsilon$ can be estimated by

$$\begin{aligned} \| F(C) - [F(C)]^\epsilon \|_\infty &\leq m_s \cdot \kappa_{A,abs,2} \cdot \epsilon_m \cdot \|C\|_2 + \mathcal{O}(\epsilon_m^2) \\ &\approx m_s \cdot \kappa_{A,abs,2} \cdot \epsilon_m \cdot \|C\|_2 . \end{aligned}$$

Proof. Define

$$|A| := \left(|a_{ij}|\right) .$$

According to [58], the matrix

$$E := (A \cdot C)^\epsilon - A \cdot C$$

fulfils

$$|E| \leq m_s \cdot \epsilon_m \cdot |A| \cdot |C| + \mathcal{O}(\epsilon_m^2) .$$

It follows

$$\begin{aligned} \| A \cdot C - (A \cdot C)^\epsilon \|_\infty = \|E\|_\infty &\leq \| m_s \cdot \epsilon_m \cdot |A| \cdot |C| + \mathcal{O}(\epsilon_m^2) \|_\infty \\ &\leq m_s \cdot \epsilon_m \cdot \| \left(|A| \cdot |C|\right) \|_\infty + \mathcal{O}(\epsilon_m^2) \\ &\leq m_s \cdot \epsilon_m \cdot \max_{\|x\|_2 = \|C\|_2} \| \left(|A| \cdot |x|\right) \|_\infty + \mathcal{O}(\epsilon_m^2) \\ &= m_s \cdot \epsilon_m \cdot \max_{\|x\|_2 = \|C\|_2} \|A \cdot x\|_\infty + \mathcal{O}(\epsilon_m^2) \\ &\leq m_s \cdot \epsilon_m \cdot \max_{\|x\|_2 = \|C\|_2} \|A \cdot x\|_2 + \mathcal{O}(\epsilon_m^2) \\ &\leq m_s \cdot \epsilon_m \cdot \kappa_{A,\text{abs},2} \cdot \|C\|_2 + \mathcal{O}(\epsilon_m^2) . \end{aligned}$$

□

This lemma shows that the maximal round-off error in $F(C) = A \cdot C$ is connected to the maximal singular value of A. Thus, reduction of the maximal singular value results in reduction of the occurring round-off error. Nevertheless, the direction of the round-off error is random, and the magnitude as well as the direction depend on the order of summation.

4.3.2 Symmetric linear functions

A linear differential equation (4.7) with a symmetric matrix A is considered. A symmetric Jacobian matrix is very unlikely for a chemical reaction system. Nevertheless, in this case the maximal singular value is equal to the maximal absolute value of all eigenvalues. Hence, the analysis of this case clarifies the idea of reducing the maximal singular value by reducing the maximal absolute value of all eigenvalues. Furthermore, in case of a symmetric matrix A, the triangular matrix $\tilde{T}$ of the Schur-decomposition is a diagonal matrix, and therefore, Schur-decomposition and eigendecomposition are the same. Omitting higher order terms the round-off error in F can be estimated by

$$\begin{aligned}\|F - F^{\epsilon}\|_2 &\leq \sqrt{m_s} \cdot \|F - F^{\epsilon}\|_{\infty} \\ &\leq \sqrt{m_s} \cdot m_s \cdot \kappa_{A,\mathrm{abs},2} \cdot \epsilon_m \cdot \|C\|_2 \\ &= \sqrt{m_s} \cdot m_s \cdot \max_i |\lambda_i| \cdot \epsilon_m \cdot \|C\|_2 .\end{aligned}$$

It is also possible to estimate the round-off error in the computer evaluation of $\tilde{F}$ of the modified differential equation for the Schur-decomposition, and accordingly the eigendecomposition. It holds

$$\begin{aligned}\|\tilde{F} - \left(\tilde{F}\right)^{\epsilon}\|_2 &\leq \sqrt{m_s} \cdot m_s \cdot \kappa_{\tilde{A},\mathrm{abs},2} \cdot \epsilon_m \cdot \|C\|_2 \\ &= \sqrt{m_s} \cdot m_s \cdot |\lambda_g| \cdot \epsilon_m \cdot \|C\|_2 \\ &\leq \sqrt{m_s} \cdot m_s \cdot \max_i |\lambda_i| \cdot \epsilon_m \cdot \|C\|_2 .\end{aligned} \tag{4.12}$$

As a consequence a reduction of the parameter λ_g also reduces the error in the Newton correction. Thus, if Newton's method fails due to round-off errors, the modification is an option in order to prevent failure of Newton's method. However, this leads to an increase in the approximation error. The approximation error is the difference between the solution of the original differential equation and the solution of the modified differential equation. According to equation (4.5), the components connected with negative eigenvalues with large absolute value decay very fast. Using the eigenvector basis V_D, one obtains

$$w = V_D^{-1} \cdot C, \qquad w_i(t) = w_i(0) \cdot e^{\lambda_i \cdot t} .$$

Note that V_D is orthogonal for symmetric matrices A. Hence, it holds $\|V_D\|_2 = \|V_D^{-1}\|_2 = 1$. Consequently, the estimation

$$\begin{aligned}\|C(h) - \tilde{C}(h)\|_2 &\leq \|V_D\|_2 \cdot \|w(h) - \tilde{w}(h)\|_2 && (4.13)\\ &\leq \|V_D\|_2 \cdot \|w(0) \cdot \mathrm{diag}\left(e^{\lambda_i \cdot h} - e^{\lambda_g \cdot h}\right)\|_2 \\ &\leq \|V_D\|_2 \cdot \|w(0)\|_2 \cdot e^{\lambda_g \cdot h} \\ &\leq \|V_D\|_2 \cdot \|V_D^{-1}\|_2 \cdot \|C(0)\|_2 \cdot e^{\lambda_g \cdot h} && (4.14)\\ &= \|C(0)\|_2 \cdot e^{\lambda_g \cdot h} && (4.15)\end{aligned}$$

follows from equation (4.6) for all eigenvalues λ_i that are smaller than $\lambda_g \ll 0$. If $\lambda_g \ll 0$ is chosen very small, the considered difference in (4.13) is surely very small (and much smaller than the given tolerance). Note that the nodes of a Runge-Kutta method can be located between t_i and t_{i+1}. Thus, the estimation also has to be fulfilled for a fraction of the step size h.

4.3.3 Non-symmetric linear functions

Now a linear differential equation of the form (4.7) with a non-symmetric matrix A is examined. Again it is assumed that the eigenvalues are not positive, and additionally, it is assumed that the matrix A has m_s linear independent real eigenvectors. For determining the maximal round-off error in the evaluation of $F(\cdot)$, one can use the maximal singular value of the matrix A together with Lemma 4.4. But in this case the eigenvalue with the maximal absolute value is not necessarily equal to the maximal singular value of A. Therefore, it is not possible to use the estimation (4.12). If the eigendecomposition is used instead of the Schur-decomposition the estimation

$$\begin{aligned} \kappa_{A,\mathrm{abs},2} &= \max_i \sigma_i = \|A\|_2 = \|V_D \cdot D \cdot V_D^{-1}\|_2 \\ &\leq \|V_D\|_2 \cdot \|D\|_2 \cdot \|V_D^{-1}\|_2 = \max_i |\lambda_i| \cdot \|V_D\|_2 \cdot \|V_D^{-1}\|_2 \end{aligned} \tag{4.16}$$

is derived easily. The modification does not change the eigenvectors. Thus, using the modification reduces the upper bound (4.16). However this bound is not sharp. Therefore, an increase of the maximal singular value is also possible. In order to bound a worst case scenario of an increasing maximal singular value, the Schur-decomposition is used. With the Schur-decomposition instead of the eigendecomposition one obtains the following lemma:

Lemma 4.5. *Using the Schur-decomposition, the modified reaction mechanism results in a decrease of the maximal singular value*

$$\max_i \tilde{\sigma}_i \leq \max_i \sigma_i \,. \tag{4.17}$$

Thereby $\tilde{\sigma}_i$ are the singular values of the modified matrix $\tilde{A}$.

Proof. The maximal singular value $\max_i \sigma_i$ of the matrix A is equal to the 2-norm $\|A\|_2$ of A. According to [58], orthogonal matrixes $Q_1, Q_2 \in \mathbb{R}^{m_s \times m_s}$ fulfil

$$\|Q_1 \cdot A \cdot Q_2\|_2 = \|A\|_2 \,.$$

Let $A = Q \cdot \tilde{T} \cdot Q^T$ be the Schur-decomposition of the matrix A. Then it follows immediately

$$\max_i \sigma_i = \|A\|_2 = \|Q \cdot \tilde{T} \cdot Q^T\|_2 = \|\tilde{T}\|_2 \,.$$

Next, the effect of the modification on the 2-norm of the Schur-decomposition is considered.

$$\begin{aligned}\tilde{A} &= \left[I_{m_s} - Q_f \cdot \mathrm{diag}_{k \le n_f} \left(1 - \frac{\lambda_g}{\tilde{T}_{kk}} \right) \cdot Q_f^T \right] \cdot A \\ &= \left[I_{m_s} - Q_f \cdot \mathrm{diag}_{k \le n_f} \left(1 - \frac{\lambda_g}{\tilde{T}_{kk}} \right) \cdot Q_f^T \right] \cdot \begin{pmatrix} Q_f & Q_s \end{pmatrix} \cdot \tilde{T} \cdot \begin{pmatrix} Q_f^T \\ Q_s^T \end{pmatrix} \\ &= \begin{pmatrix} Q_f & Q_s \end{pmatrix} \cdot \begin{pmatrix} \mathrm{diag}_{k \le n_f} (\frac{\lambda_g}{\tilde{T}_{kk}}) & 0 \\ 0 & I_{n_s} \end{pmatrix} \cdot \tilde{T} \cdot \begin{pmatrix} Q_f^T \\ Q_s^T \end{pmatrix}.\end{aligned}$$

The absolute values of all elements of the diagonal matrix are smaller or equal to one. Thus, one obtains equation (4.17) by

$$\begin{aligned}\max_i \tilde{\sigma}_i &= \|\tilde{A}\|_2 \\ &= \left\| \begin{pmatrix} \mathrm{diag}_{k \le n_f} (\frac{\lambda_g}{\tilde{T}_{kk}}) & 0 \\ 0 & I_{n_s} \end{pmatrix} \cdot \tilde{T} \right\|_2 \\ &\le \left\| \begin{pmatrix} \mathrm{diag}_{k \le n_f} (\frac{\lambda_g}{\tilde{T}_{kk}}) & 0 \\ 0 & I_{n_s} \end{pmatrix} \right\|_2 \cdot \left\| \tilde{T} \right\|_2 \\ &\le \|\tilde{T}\|_2 = \max_i \sigma_i \,.\end{aligned}$$

□

Hence, the round-off error in the source term can be decreased with the modification. Although, the modification leads to an approximation error. If the modification method using the eigendecomposition is used, the approximation error can be estimated by equation (4.13) to (4.14). In case of the Schur-decomposition, the transformed variable $w = Q^T \cdot C$ is considered. Then, the considered differential equations are

$$\begin{aligned} w' &= \tilde{T} \cdot w, \qquad w(t_0) = w^0, \\ \tilde{w}' &= \begin{pmatrix} \mathrm{diag}_{k \le n_f} (\frac{\lambda_g}{\tilde{T}_{kk}}) & 0 \\ 0 & I_{n_s} \end{pmatrix} \cdot \tilde{T} \cdot \tilde{w}, \qquad \tilde{w}(t_0) = w^0 \,. \end{aligned}$$

The difference between w and $\tilde{w}$ is given by

$$\begin{aligned}\|w(t) - \tilde{w}(t)\| &= \left\| \begin{pmatrix} I_{n_f} & \\ & 0 \end{pmatrix} \cdot [w(t) - \tilde{w}(t)] + \begin{pmatrix} 0 & \\ & I_{n_s} \end{pmatrix} \cdot [w(t) - \tilde{w}(t)] \right\| \\ &= \left\| \begin{pmatrix} I_{n_f} & \\ & 0 \end{pmatrix} \cdot [w(t) - \tilde{w}(t)] \right\|.\end{aligned}$$

Thereby it is used that the last n_s rows of the triangular matrix $\tilde{T}$ and its modification are equal. Now, assume that the slow processes are very slow, and that therefore, the slow variables are almost constant in each time step such that

$$w_i = \tilde{w}_i = \text{const}, \qquad i \in \{n_f + 1, ..., m_s\}.$$

This is a restriction, which can be avoided by using the eigendecomposition because then the error in the modified model just depends on the length of the time step. With this restriction, we obtain differential equations for the first n_f components of w and $\tilde{w}$. It holds

$$\begin{aligned} w'_{(1,n_f)} &= \tilde{T}_{(1,n_f),(1,n_f)} \cdot w_{(1,n_f)} + c_1, & w_{(1,n_f)}(t_0) &= w^0_{(1,n_f)}, \\ \tilde{w}'_{(1,n_f)} &= \text{diag}_{k \le n_f}\left(\frac{\lambda_g}{\tilde{T}_{kk}}\right) \cdot \tilde{T}_{(1,n_f),(1,n_f)} \cdot \tilde{w}_{(1,n_f)} + c_2, & \tilde{w}_{(1,n_f)}(t_0) &= w^0_{(1,n_f)}. \end{aligned}$$

Here $c_1, c_2 \in \mathbb{R}^{n_f}$ are constants. Furthermore, define the state w_{eq} as the equilibrium state of both these differential equations. Finally, the difference between w and $\tilde{w}$ can be estimated by

$$\begin{aligned} \|w(t) - \tilde{w}(t)\| &= \|w_{(1,n_f)}(t) - \tilde{w}_{(1,n_f)}(t)\| \\ &\le \|w_{(1,n_f)}(t) - w_{eq}\| + \|w_{eq} - \tilde{w}_{(1,n_f)}(t)\| \\ &\le e^{\mu_2[\tilde{T}_{(1,n_f),(1,n_f)}]\cdot t} \cdot \|w_{(1,n_f)}(t_0) - w_{eq}\| \\ &+ e^{\mu_2[\text{diag}_{k \le n_f}(\lambda_g/\tilde{T}_{kk})\cdot \tilde{T}_{(1,n_f),(1,n_f)}]\cdot t} \cdot \|\tilde{w}_{(1,n_f)}(t_0) - w_{eq}\|. \end{aligned}$$

Thereby, $\mu_2[\cdot]$ is the logarithmic matrix norm. According to [120, 125, 126], it holds for a matrix $A \in \mathbb{R}^{m_s \times m_s}$

$$\begin{aligned} \mu_2[A] &= \lambda_{max}\left[\frac{1}{2} \cdot \left(A + A^T\right)\right] \\ &\le \frac{1}{2} \cdot \left[\max_j \left(a_{jj} + \sum_{i=1, i\neq j}^{m_s} |a_{ij}|\right) + \max_i \left(a_{ii} + \sum_{j=1, j\neq i}^{m_s} |a_{ij}|\right)\right]. \end{aligned}$$

Hence, the usage of the Schur-decomposition is limited by the size of the off-diagonal elements of $\tilde{T}_{(1,n_f),(1,n_f)}$. If the diagonal elements are negative and much larger than the absolute values of the other elements and the slow processes are very slow in comparison to the fast processes, the approximation error is small and the modification of the differential equation can be used in order to decrease the occurring round-off errors in the source term. The analysis of the linear case is an argument to use the method for nonlinear differential equations as well. This argument is supported by the fact that the main reason for the usage of the eigendecomposition for finding processes with short timescales is a linearization of the differential equation.

4.3.4 Nonlinear functions

Finally, a nonlinear differential equation is considered. The nonlinear differential equation is

$$C' = F(C), \tag{4.18}$$

$$F(C) : \mathbb{R}^{m_s} \to \mathbb{R}^{m_s}. \tag{4.19}$$

The differential equation models a chemical reaction system, and the source term $F(C)$ contains processes with several different timescales. The Jacobian matrix of the nonlinear function $F(C)$ can be used to approximate the directions of fast progress, and the reaction velocity in these directions can be decreased (see (4.9) or (4.10)). Obviously, the application of the modification on the already evaluated function $F(C)$ does not reduce the round-off error and effects no improvement. On the contrary, an approximation error is introduced in addition to the round-off error. But the differential equation (4.18) originates from the modelling of a chemical reaction sytem. Hence, the function $F(C)$ occurs naturally in the form (see Section 2.1 and [8])

$$\begin{aligned} F(C) &= R \cdot \nu(C), \\ R &\in \mathbb{Z}^{m_s \times m_r}, \qquad \nu(C) : \mathbb{R}^{m_s} \to \mathbb{R}^{m_r}. \end{aligned}$$

In this equation the matrix R is the stoichiometric matrix, and the vector $\nu(C)$ contains the reaction rates of all reactions. With this representation the modification can be applied to the stoichiometric matrix R and the modification can reduce the resulting round-off error. The modified source term is

$$\begin{aligned} \tilde{F}(C) &= \left[\left(V_D \cdot \operatorname{diag}(a_i) \cdot V_D^{-1}\right) \cdot R\right] \cdot \nu(C) \\ &=: \tilde{R} \cdot \nu(C). \end{aligned} \tag{4.20}$$

If the eigenvector-matrix V_D is ill-conditioned or only a few fast processes occur, using the Schur-decomposition provides the alternative modified source term

$$\begin{aligned} \tilde{F}(C) &= \left[\left(Q \cdot \operatorname{diag}(a_i) \cdot Q^T\right) \cdot R\right] \cdot \nu(C) \\ &= \left\{\left[I_{m_s} - Q_f \cdot \operatorname{diag}_{i \le n_f}(1 - a_i) \cdot Q_f^T\right] R\right\} \cdot \nu(C) \\ &=: \tilde{R} \cdot \nu(C). \end{aligned} \tag{4.21}$$

Thereby a_i follows from equation (4.6). Thus, the fraction of every reaction in direction of a fast process is multiplied by a factor smaller than one before summation. Hence, the absolute value of this fraction is decreased, and thereby the round-off error is also reduced because the round-off error is related to the fractions related to large singular values. If the linearization of the nonlinear function $F(C)$ is a valid approximation, the introduced approximation error is small and the modification can be used to prevent failure of Newton's method.

4.3.5 Choice of the parameter λ_g

The choice of λ_g is crucial. It will be illustrated for the modification (4.8) of the linear differential equation (4.7). The linearization is regarded for nonlinear differential equations.
Two opposed aspects are involved in the choice of λ_g. First, equation (4.14) shows that the approximation error is decreasing for smaller λ_g and that the approximation error is equal to 0 for $\lambda_g \leq \min_i \lambda_i$. Thus, regarding this aspect λ_g should be chosen as small as possible.
On the other hand it is shown in the prior sections that the absolute value of the smallest negative eigenvalue is proportional to the occurring round-off error in the computer evaluation of the source term. Furthermore, the absolute value of the smallest negative eigenvalue of the modified system is bounded by $|\lambda_g|$. Regarding this aspect, $|\lambda_g|$ should be chosen as small as possible. Then the stiffness of the differential equation is also decreased.
Stiff differential equations are defined by restrictions on the step size h due to stability problems for explicit methods that do not occur for A-stable implicit solvers [28, 64, 126]. All stability regions S of explicit methods are a subset of some $K_r(0) = \{x : \|x\|_2 \leq r\}$, so $\lambda_i \cdot h \in S,\ 1 \leq i \leq m_s$, requires $|\lambda_i \cdot h| \leq r,\ 1 \leq i \leq m_s$, and restricts the step size h dependent on $\max_i |\lambda_i|$. Thus, restrictions on h due to stability problems occur if there are eigenvalues λ_i of the Jacobian matrix F_C of the source term $F(C)$ with $|\lambda_i| \gg 0$. A second restriction on h is given by the constraint

$$(h \cdot \tilde{\tau}_i) \cdot e^{\mu \cdot (t_{end} - t_i)} \overset{!}{\leq} h \cdot \mathrm{TOL}$$

due to the error propagation of the local error $h \cdot \tilde{\tau}_i$ with the end time t_{end} and the one-sided Lipschitz constant $\mu (= \max_i \mathrm{Re}(\lambda_i)$ for $F_C = F_C^T)$. In case of $\max_i \mathrm{Re}(\lambda_i) \ll \max_i |\lambda_i|$, the first restriction can be much more restrictive than the second one. Then, a problem is called stiff. Thus, if the absolute value of λ_g is smaller than the absolute value of the smallest eigenvalue of the system, the stiffness of the modified differential equation is decreased.
Nevertheless, the introduction of a large approximation error in order to save computing time is not an option. Thus, for a given step length h the bound λ_g is chosen such that the corresponding component decays in a small fraction of the step length to a small fraction of the tolerance. The corresponding requirement is

$$e^{\lambda_g \cdot h/\delta_2} \leq \delta_1 \cdot \mathrm{TOL}_{rel}, \qquad \delta_1 \ll 1 \ll \delta_2 .$$

The parameter δ_2 makes sure that the approximation error is small at all stages of the Runge-Kutta method. The parameter δ_1 gives a relation between the approximation error and the tolerance. It is chosen very small because the linearized system is regarded and small δ_1 provides a safety buffer. Furthermore, λ_g has also to be negative for large Δt. Therefore, an upper bound for λ_g is introduced. Then the following choice for λ_g is obtained:

$$\lambda_g = \min\left(-1, \frac{\delta_2}{h} \cdot \log(\delta_1 \cdot \mathrm{TOL}_{rel})\right)$$

The parameteres $\delta_1 = 10^{-3}$ and $\delta_2 = 50$ are used for the numerical example. However, different parameter sets are also possible.

4.3.6 Computational costs

The system of differential equations is modified in order to reduce the computational costs. Therefore, the computational costs of the modification is compared to the computational costs of one additional time step. The modification includes the computation of the Schur-decomposition or the eigendecomposition of the Jacobian matrix and the transformation of the stoichiometric matrix R to $\tilde{R}$. The computation of the Jacobian matrix is also necessary for additional steps due to reduced step sizes. Therefore, the computation of the Jacobian matrix can be omitted for the modification of the chemical source term and for any additional step. For a $(m_s \times m_s)$-matrix the computation of the Schur-decomposition with the Francis QR algorithm takes approximately $25 \cdot m_s^3$ flops [49, 50, 59]. However, the complete Schur-decomposition is not necessary. If the number of fast directions is smaller than m_s, the effort reduces. The number of fast directions is denoted by n_f, and the number of reactions is denoted by m_r. Furthermore, assume $m_r \approx m_s$. Then, the transformation of the stoichiometric matrix in (4.21) takes $m_s^2 + m_s \cdot n_f^2 + m_s^2 \cdot n_f + m_s^2 \cdot m_r \approx 3 \cdot m_s^3$ flops. In comparison every (rejected or accepted) time step of a Runge-Kutta method with s stages consists of solving one nonlinear equation system of the dimension $m_s \cdot s$. If Newton's method is used, the computational costs consist of the computation of the Jacobian matrix and solving at least one linear system of the size $m_s \cdot s$. In case of fully implicit Runge-Kutta methods the effort is $\frac{(s \cdot m_s)^3}{3}$ for Gaussian elimination with complete pivoting. In the numerical tests (see Section 4.4) a Runge-Kutta method with $s = 3$ is used. Therefore, every modification costs as much as three additional steps. Hence, the modification of the differential equation pays off if the number of time steps is reduced at least by three. In case of the modification (4.9), the eigendecomposition is obtained from the Schur-decomposition by backsubstitution in $\sum_{i=1}^{m_s} \frac{i^2}{2} < \frac{m_s^3}{4}$ flops [108] for $m_s \geq 3$. Additionally, the inverse of the eigenvector basis has to be computed. This takes $2 \cdot m_s^3$ flops. Furthermore, the transformation of the stoichiometric matrix in (4.20) takes $2 \cdot m_s^3 + m_r \cdot m_s^2$ flops. Thus, the computational costs of the modification of the chemical source term are less than the computational costs of four additional steps of the Runge-Kutta method.

Remark 4.6. *For diagonally implicit Runge-Kutta methods the effort of solving the linear system of the size* $m_s \cdot s$ *is* $\frac{s \cdot m_s^3}{3}$.

4.4 Numerical test case

The modified model is implemented using the algorithm RADAU5 by Hairer and Wanner [64]. Although, Hairer and Wanner provide a Fortran-code for this method, the Matlab version by Engstler [46] is used. The algorithm RADAU5 is a fully implicit Runge-Kutta method with three stages and order five. This method is chosen because the method suits well for the simulation of chemical reaction systems and it is well known and widely often recommended. An important aspect is that the modification of the problem consists of a modification of the chemical source term, and thus, it can be used for most solvers for ordinary differential equations. If Newton's method does not converge, the eigendecomposition of the Jacobian matrix of the chemical source term is computed. Thereby the fast processes can be determined

as column vectors of V_D belonging to eigenvalues smaller than a given parameter λ_g. The resulting factor (4.20) is kept until Newton's method fails again. The used relative tolerance is TOL_{rel}, and the used absolute tolerance is TOL_{abs}. The analytic solutions of the numerical example is not available, and thus, a reference solution that is computed with smaller tolerances is used to estimate the occurring error. For the computed solution C and the reference solution C_{ref} the error estimator is defined by

$$e = \max_i \left(\frac{|C_i - \left(C_{ref}\right)_i|}{\mathrm{TOL}_{rel} \cdot \left(C_{ref}\right)_i + \mathrm{TOL}_{abs}} \right). \tag{4.22}$$

If the computed error is between zero and one, the computed solution C is sufficiently reliable. The algorithm improves clearly its performance due to the described modification of the chemical source term for the following considered numerical example.

The numerical example bases upon the chemical system [66]

$$H_2 + O_2 \rightleftarrows H\cdot + HO_2, \qquad k_1 = 6.23 \cdot 10^{-5}, \qquad k_{-1} = 1.76 \cdot 10^{7}, \tag{4.23}$$
$$H_2 + OH\cdot \rightleftarrows H\cdot + H_2O, \qquad k_2 = 1.20 \cdot 10^{6}, \qquad k_{-2} = 2.64 \cdot 10^{3}, \tag{4.24}$$
$$O_2 + H\cdot \rightleftarrows OH\cdot + O\cdot, \qquad k_3 = 4.26 \cdot 10^{4}, \qquad k_{-3} = 1.14 \cdot 10^{7}, \tag{4.25}$$
$$H_2 + O\cdot \rightleftarrows H\cdot + OH\cdot, \qquad k_4 = 2.20 \cdot 10^{5}, \qquad k_{-4} = 2.51 \cdot 10^{5}, \tag{4.26}$$
$$OH\cdot \rightleftarrows (OHW), \qquad k_5 = 0.92, \qquad k_{-5} = 0.0092, \tag{4.27}$$
$$H\cdot \rightleftarrows (HW), \qquad k_6 = 0.08, \qquad k_{-6} = 0.0008, \tag{4.28}$$
$$O\cdot \rightleftarrows (OW), \qquad k_7 = 0.92, \qquad k_{-7} = 0.0092, \tag{4.29}$$
$$[k_i] = \frac{\mathrm{m}^3}{\mathrm{mol}\cdot\mathrm{s}} \text{ for } i \in \{\pm1, \pm2, \pm3, \pm4\}, \qquad [k_i] = \frac{1}{\mathrm{s}} \text{ for } i \in \{\pm5, \pm6, \pm7\}.$$

Thereby the expression $(\cdot W)$ means that the species is attached to the container wall. In the following the SI base units (metre, kilogram, second, ampere, kelvin, mole, candela) are used. The corresponding unit symbols are m, kg, s, A, K, mol and cd. The reaction rates of the reactions (4.23) to (4.26) are computed with the Arrhenius equation for a fixed temperature of 1000K. Thereby, the coefficients of the Arrhenius equation are taken from [92]. The reaction rates of the reactions (4.27) to (4.29) are listed in [66]. It is assumed that for the reactions (4.27) to (4.29) the backwards reaction rates are hundred times smaller than the forward reaction rates. The reactions (4.23) to (4.29) are elementary reactions, and therefore, the reaction rates of these reactions follow immediately.

4.4.1 Numerical Jacobian matrix

In the first scenario the Jacobian matrix of the chemical source term is not available. Therefore, the Jacobian matrix of the source term is computed numerically. The considered time span is $[0\mathrm{s}, 240\mathrm{s}]$, and

the tolerances are given by $\mathrm{TOL}_{rel} = 10^{-3}$ and $\mathrm{TOL}_{abs} = 10^{-6}$. Furthermore, the initial values are given by

$$C(0) = \begin{pmatrix} C_{H_2}(0) \\ C_{O_2}(0) \\ C_H(0) \\ C_{HO_2}(0) \\ C_{OH}(0) \\ C_{H_2O}(0) \\ C_O(0) \\ C_{OHW}(0) \\ C_{HW}(0) \\ C_{OW}(0) \end{pmatrix} = \begin{pmatrix} 5 \\ 2.5 \\ 0 \\ 0 \\ 0 \\ 2.5 \\ 2.5 \\ 0 \\ 0 \\ 0 \end{pmatrix} \text{ in } \frac{\text{mol}}{\text{m}^3}. \tag{4.30}$$

These initial values correspond to a total concentration of $12.5 \frac{\text{mol}}{\text{m}^3}$. With the ideal gas law this concentration and the given temperature match a pressure of $10^5 \frac{\text{kg}}{\text{m}\cdot\text{s}^2} \approx 1\text{bar}$. In case of non-convergence of Newton's method and $t > 0$, fast directions are computed with help of the eigendecomposition of the source term's Jacobian matrix. This happens once (if we modify the problem after the first time). Thus, the additional computing effort is just one eigendecomposition. For that matter, directions corresponding to eigenvalues smaller than $\lambda_g \approx -126$ are replaced by λ_g (current step size at the time of modification $t = 29$ is 5.46, for computation of λ_g see Subsection 4.3.5). Figure 5 shows that the modified problem improves performance. The total number of operations is listed in Table 1. Furthermore, Figure 6 shows that the occurring error is although not increased. Thereby, the reference solution is computed with a relative tolerance $TOL_{rel} = 10^{-5}$ and an absolute tolerance $TOL_{abs} = 10^{-9}$.

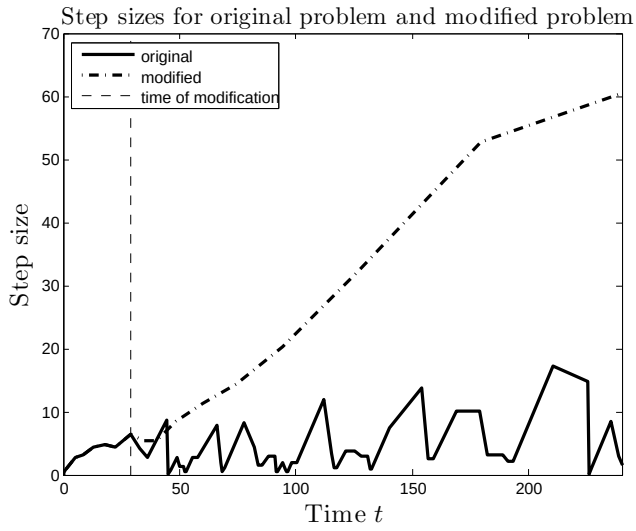

Figure 5: Time step size for RADAU5 and the modified RADAU5 for a numerically computed Jacobian matrix

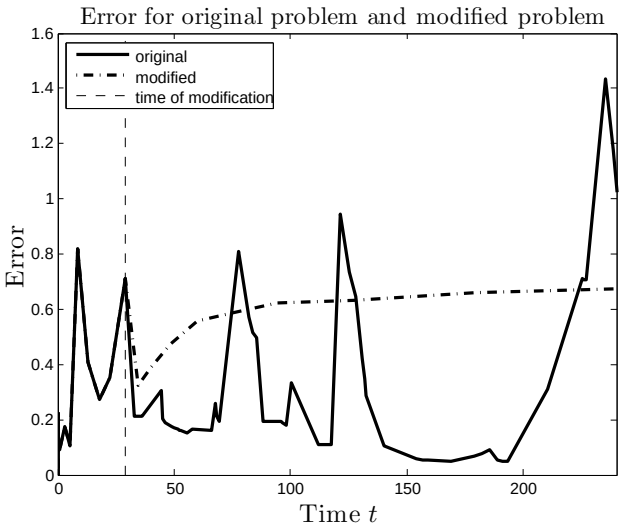

Figure 6: Error (4.22) for RADAU5 and the modified RADAU5 for a numerically computed Jacobian matrix

	Standard	Modification
function calls	2538	1091
jacobian	100	50
steps	191	71
accepted steps	100	49

Table 1: Costs for the simulation of Hydrogen Oxygen Combustion in case of a numerically computed Jacobian matrix

4.4.2 Analytical Jacobian matrix

In the second scenario the analytical expression of the Jacobian matrix F_C is available. In this case, the problem can be integrated using the standard model without any problems. Therefore, a very slow water-consuming reaction is added. Due to this slow reaction the steady state is reached very slowly, and the described problem, that means the non-convergence of Newton's method as a result of stiffness, occurs. The added reaction is defined by

$$r_{15} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & -1 & 0 & 0 & 0 & 0 \end{pmatrix}^T, \qquad v_{15}(C) = 10^{-8} \cdot C_6 \,.$$

Furthermore, the initial values are given by $12.2 \cdot C_0$, and the used tolerances are $\mathrm{TOL}_{rel} = 10^{-5}$ and $\mathrm{TOL}_{abs} = 10^{-10}$. The considered time interval is $[0\mathrm{s}, 10^9\mathrm{s}]$. Again one modification is necessary. According to Subsection 4.3.5 and the very large time step, the parameter λ_g is set equal to -1. Figure 7 shows that the modified problem improves performance in case of large step sizes. The total number of operations is listed in Table 2. Furthermore, Figure 8 shows that the occurring error is not increased. Thereby, the reference solution is computed with a relative tolerance $TOL_{rel} = 10^{-7}$ and an absolute tolerance $TOL_{abs} = 10^{-12}$. Note that the error estimator of the method RADAU5 is larger than 1 for very large t. Hence, the required tolerance is not met. However, the modified method fulfils the required tolerance.
All in all, the needed computing time is decreased clearly. If the reaction rate of the added reaction is decreased and the observed time span is increased, the improvement even gets larger.

	Standard	Modification
function calls	6338	1050
jacobian	426	125
steps	845	151
accepted steps	426	124

Table 2: Costs for the simulation of Hydrogen Oxygen Combustion in case of an analytical Jacobian matrix

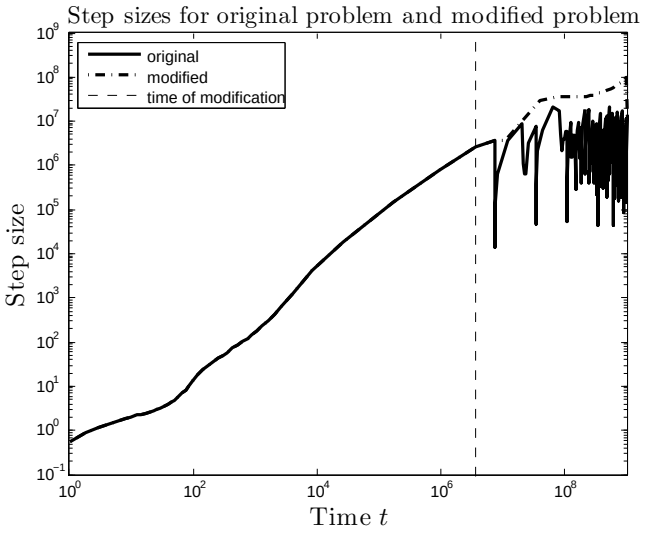

Figure 7: Logarithmic time step size for RADAU5 and the modified RADAU5 in case of an analytical Jacobian matrix

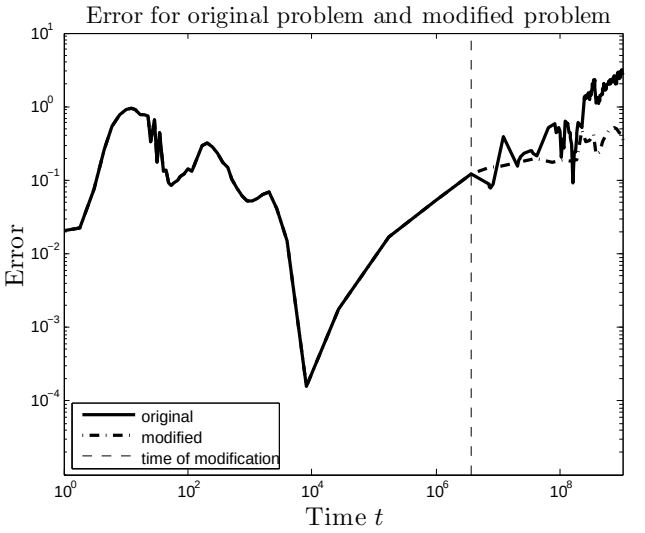

Figure 8: Error (4.22) for RADAU5 and the modified RADAU5 in case of an analytical Jacobian matrix

4.5 Conclusions and summary

A new method of modification for chemical reaction systems has been introduced in this chapter. Like the Intrinsic Low-Dimensional Manifold method [92], the REaction-DIffusion Manifold method [21], the Global Quasi-Linearization method [19], the Flamelet-Generated Manifold method [136], and many more, it uses the existence of different timescales. However in contrast to the named methods, the reaction velocities of fast processes are not set to infinity but reduced. Therefore, this method is not able to reduce the number of variables. But the method can be used for differential equations with only a few or a changing number of fast processes. Thus, the area of application is larger for this new method. The introduced method reduces these round-off errors in the evaluation of the chemical source term. Thus, step-size reductions are avoided, and the computing time can be decreased.

5 Parameter identification for chemical reaction systems

5.1 Introduction

In this chapter, the usage of different timescales for parameter identification of ODEs is discussed. Thereby we consider ODEs resulting from modelling chemical reaction systems. The mathematical model of a chemical reaction system is a partial differential equation. But in order to simplify the discussion of parameter identification for fast chemical reaction systems, a spatially homogeneous gas reaction system with constant temperature and constant volume is examined in this chapter. This is consistent with the assumption that any scientist would choose the simplest experimental set-up (stirred tank reactor with constant temperature) in the case of parameter identification of unknown reaction rate constants. Thus, the following ODE is considered:

$$C' = F(C, p^*), \qquad C(0, p^*) = C_0 \in \mathbb{R}^{m_s} \tag{5.1}$$

Thereby the parameter vector $p^* \in \mathbb{R}^{n_p}$ contains the unknown reaction rate constants. In the following, parameter sets are called p, but p is not necessarily equal to the true physical parameter p^*. If the temperature is not fixed, there are three parameters for each chemical reaction, according to the Arrhenius equation (2.6). The goal of parameter identification is to find a parameter set p that is in some sense the best fit to given measurements of the solution $C(t, p^*)$ of (5.1). Parameter identification for (5.1) has been a research topic since many years [14, 15, 45, 84, 95, 102, 117], and it usually requires many forward solutions of the chemical system [135] for different parameter sets p. However, simulations of a chemical reaction system can lead to large computational effort, and therefore, parameter identification for (5.1) can be an infeasible task. In order to reduce the computational effort of solving (5.1), many authors use reduction mechanisms, which are based on QSSA or the PEA (see Chapter 3 and [20, 51, 78, 92, 136, 132, 133, 134]). Thereby the state of the system is restricted to a manifold with a reduced dimension in comparison to the original number of state variables. Most reduction mechanisms use an eigendecomposition of the Jacobian matrix $F_C(C, p^*)$ of the source term $F(C, p^*)$ in order to construct the reduced manifold. Thereby the eigenvectors corresponding to negative eigenvalues with large absolute value are directions of fast decay, in which the chemical reaction system reaches equilibrium instantly. Afterwards, the state variables have to be determined using a reduced set of variables, and a cheap way to achieve this is needed. A common procedure is to create a look-up table [90, 91]. The look-up table is precomputed and can be used for all simulations with the true parameter p^*. Note that computing the corresponding look-up table is much more expensive than solving the differential equation once. However, the obtained manifold depends on the parameter p (see Section 5.2). From this follows that a look-up table cannot be valid for all parameter sets p in the parameter space. Therefore, it is impossible to precompute one reduced differential equation with one precomputed look-up table for all parameter sets in the parameter space. Moreover, it is too expensive to generate several parameter-dependent reduced models that cover the complete parameter space.

A new approach [86, 88] by Lukassen and Kiehl has been developed in order to use model order reduction for parameter identification. Thereby the given measurement data of the state of the system are

used as a reduced basis. A basic introduction of reduced basis methods is given in [60, 103, 107]. Then the obtained reduced model is a good approximation only close to the given data. Hence, the reduced system can lead to large errors for parameter sets that correspond to solutions far away from the given measurements of the state of the system. In order to prevent convergence of the parameter estimator to such a local minimizer the approximation error between the full and the reduced system is penalised. Our publications [86, 88] show that our approach performs very well if an efficient and reliable estimator of the approximation error is given. E.g., elliptic partial differential equations with a continuous and coercive bilinear form fulfil the required assumption. In this case our approach accelerates the parameter identification. However, the differential equation (5.1) is highly nonlinear, and the Lipschitz constant of the chemical source term is very large due to stiffness. Hence, an efficient and reliable error estimator is not available for reduced models of differential equation (5.1), and the approach presented in [86, 88] is not applicable to chemical reaction systems.

For the same reason, adaptive model reduction, which generates different look-up tables along the optimization path of p, does not work. Adaptive model reduction is used for partial differential equations [27, 39, 53, 109]. But in these cases a good estimator of the approximation error [25, 60, 61, 65, 103, 107, 140] is available in opposition to model reduction of (5.1). Without an estimator of the approximation error it is hardly possible to determine, whether the currently used reduced model provides a good approximation for the considered parameter p.

Nevertheless, the existence of directions of fast decay can be utilized for parameter identification of chemical reaction systems. In a first step, fast chemical processes are identified without using the unknown parameter p. For the identification of the fast processes QSSA for single species and PEA for single reactions are employed instead of a complex mechanism like eigendecomposition. Using QSSA for single species or PEA for single reactions is based on experience and intuition of the investigator instead of the known reaction rate constants. Furthermore, if the approximate order of reaction rate constants is known, QSSA and PEA can be verified, but an eigendecomposition might still result in large approximation errors for the reduced model (see Example 7). Selection criteria for PEA and QSSA are listed in Subsection 5.4.1. So automatic computation of fast chemical processes is abandoned due to the unknown parameters.

Furthermore, thermodynamic data are used in order to speed up parameter identification. The thermodynamic data result in a relation between the reaction rates of the forward and the backward reaction. Therefore, the dimension of the parameter space can be decreased if the reaction rate constants of a pair of forward and backward reaction are unknown. Besides an algebraic equation for the state of the system follows from thermodynamics for each reaction in partial equilibrium. This algebraic equation does not depend on the reaction rate constants. Hence, the algebraic equation is independent of the unknown parameters. Thus, the algebraic equation holds for all parameter sets in the parameter space. Moreover, the reaction rate constants of a reaction in partial equilibrium can be eliminated before estimating the other reaction rate constants. Hence, PEA results in a dimension reduction of the parameter space, which speeds up the optimization routine. Furthermore, if all involved reaction rate constants are given, QSSA results in an algebraic equation for the state of the system. Therefore, PEA and QSSA provide algebraic equations for the state of the system in spite of the existence of unknown reaction rate constants. If the low-dimensional manifold can be precomputed, the dimension of the

differential equation can be drastically reduced. Moreover, the dimension of the parameter space is decreased. Thus, the total costs of parameter identification are lowered by PEA and QSSA.

The structure of the chapter is the following. In Section 5.2, it is shown that the low-dimensional manifold is parameter-dependent. Moreover, the thermodynamic description of the equilibrium of a reaction is introduced in Section 5.3. Afterwards, in Section 5.4, selection criteria for reactions in partial equilibrium and species in quasi-steady state are listed. Furthermore, applicability of PEA and QSSA in case of unknown parameters p is discussed. Finally, numerical examples illustrate the proposed approach. In all sections of this chapter examples are used in order to clearify the matter. If the usage of a nonlinear chemical reaction system complicates the analysis (e.g., eigendecomposition of the Jacobian is not constant), linear fictitious reaction systems are employed.

5.2 Parameter-dependence of the low-dimensional manifold

In this section, a short example will clarify that the low-dimensional manifold, which is defined by the fast processes, can be parameter-dependent. In the considered example the equilibrium of the fast processes depends on a one-dimensional unknown parameter. Thereby the unknown parameter is a reaction rate constant in a fictitious chemical reaction system.

Example 6. *Consider*

$$\begin{aligned} A_1 &\to A_2, \qquad k_1 = 10^3, \\ A_2 &\to A_1, \qquad k_2 = p \gg 1, \\ A_3 &\to A_1, \qquad k_3 = 1\,. \end{aligned}$$

The Jacobian matrix has a negative eigenvalue $-(10^3+p)$ *with large absolute value, the eigenvalue* -1*, and the eigenvalue zero, which corresponds to the conservation of the concentrations* $C_1+C_2+C_3$*. The conserved quantity results in*

$$C_3(t) = \left[\sum_{i=1}^{3} C_i(0)\right] - C_1(t) - C_2(t)\,.$$

Therefore, $C_3(t)$ *can be eliminated. Then* $C_1(t)$ *and* $C_2(t)$ *are the unknown variables. Furthermore, the eigenvector that is connected to the eigenvalue* $-(10^3+p)$ *is* $v_1 = (1,-1,0)^T$*. Thus, equation* (3.4) *defines the manifold (for* C_1 *and* C_2*) by*

$$k_1 C_1 = p C_2\,.$$

Hence, the manifold is parameter-dependent. It is not possible to compute a single look-up table for different choices of the parameter. However, C_2 *depends linear on* C_1 *in this example. Therefore, computing the manifold is computational cheap.*

Thus, it has been shown that the low-dimensional manifold is parameter-dependent. Therefore, it is impossible to precompute the low-dimensional manifold for all parameter sets at once. Although,

precomputing several reduced models for all different parameter sets is computational expensive in time as well as storage. Hence, it is not feasible.

An obvious approach is the computation of a reduced model, while neglecting the chemical reactions that involve unknown reaction rate constants. Then an automatic reduction mechanism can be used, and the reduced model speeds up parameter identification. However, this approach might result in large approximation errors. Example 7 proves this point.

Example 7. *Consider*

$$\begin{aligned} A_1 &\to A_2, & k_1 &= 1, \\ A_2 &\to A_1, & k_2 &= 10^3, \\ A_3 &\to A_2, & k_3 &= 1, \\ A_2 &\to A_3, & p &= k_4 = 10^3 k_3 \,. \end{aligned}$$

So the reaction rate constant k_4 is the unknown parameter. The corresponding differential equation is

$$\begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix}' = \begin{pmatrix} -1 & 10^3 & 0 \\ 1 & -10^3 & 1 \\ 0 & 0 & -1 \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & -p & 0 \\ 0 & p & 0 \end{pmatrix} \begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix} =: M_1 \begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix} + M_2(p) \begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix}.$$

Regarding the parameter-independent matrix M_1, the eigenvector $v_1 = (1,-1,0)^T$ is associated with the eigenvalue -1001. According to reduction mechanisms using the Schur-decomposition, the following algebraic equation holds:

$$Q_f^T F(C,p^*) \stackrel{!}{=} 0 \text{ with } Q_f = v_1 \,.$$

Although, the matrix $M_1 + M_2(1000)$ has the eigenvector $\tilde{v}_1 = (-1,2,-1)^T$, which is associated with the eigenvalue -2001. Hence, the correct manifold is defined by

$$\tilde{Q}_f^T F(C,p^*) \stackrel{!}{=} 0 \text{ with } \tilde{Q}_f = \tilde{v}_1 \,. \tag{5.2}$$

On the correct manifold the concentration C_2 depends on the sum of C_1 and C_3. However, the fast subspace Q_f results in an overweighting of C_1 and an underweighting of C_3. Thus, neglecting the reaction that involves the unknown reaction rate constant causes a seriously perturbed manifold and a large approximation error. Note that the assumption $p \gg 1$ results in the QSSA for the species A_2. It follows

$$C_2' = C_1 - (10^3 + p)C_2 + C_3 \approx 0 \,. \tag{5.3}$$

Equation 5.3 depends on the unknown parameter p and it is equal to equation 5.2.

Example 7 shows that neglecting the unknown reaction rate constants might result in a large approximation error. Hence, the reduced model has to be computed for each considered parameter. The following steps are necessary for the dimension reduction of differential equations in parameter identification. Firstly, a parameter-dependent routine is necessary in order to determine the low-dimensional manifold. For example, the ILDM method requires an eigendecomposition of the Jacobian matrix $F_C(C,p)$. The eigendecomposition is used to generate a nonlinear equation system, which has to be solved in order to obtain the low-dimensional manifold. However, the Jacobian matrix depends on p. Thus, the reduced system depends on the unknown parameter p, and computing the eigendecomposition as well as solving the nonlinear equation system is necessary for each regarded parameter set. The computation of the simplified model is more expensive than solving the system (5.1) once. Therefore, using an automatic reduction mechanism does not reduce the computing time for parameter identification. Nevertheless, PEA and QSSA can be used for paramter identification. The corresponding approach is introduced in Section 5.4.

Remark 5.1. *The argumentation in this section is also valid for parameter identification of the parameters in the Arrhenius equation* (2.6).

5.3 Thermodynamic description of partial equilibrium

Example 6 shows that the determination of the low-dimensional manifold requires the unknown parameter set because the corresponding algebraic equations depend on the parameter set. However, thermodynamic data can be used in order to compute the partial equilibrium in the direction of fast chemical reactions [51, 110]. Thereby thermodynamic data are available for most chemical species (e.g., NIST Standard Reference Database Number 69 [83]). In this section, the most important results of [51] are recapitulated.

Determining the partial equilibrium of a certain reaction by thermodynamics is based on the second law of thermodynamics. The second law of thermodynamics states that the total entropy of an isolated system increases or stays constant over time. Therefore, the steady state of the system is described by constant entropy. In [51] a thermodynamic description of the partial equilibrium in the directions $q \in \mathbb{R}^{m_s}$ is derived for the reaction setting of (5.1) (a chemical reaction system with constant temperature and constant volume). Other reactor settings are also examined in [51]. For the thermodynamic description of the equilibrium of a reaction, the chemical potential μ_k of each species A_k with $q_k \neq 0$ is required. The chemical potential is the partial molar Gibbs free energy for a chemical reaction system with constant temperature and constant volume, and it depends on temperature and pressure. The constant temperature is denoted by T. The pressure depends on the concentration vector C as well as the temperature T, and it is denoted by $p_{press}(C,T)$. Hence, the chemical potential $\mu_k =: \mu_k(C,T)$ is a function of the concentration vector C and the temperature T. Furthermore, $\tilde{t} = \frac{T}{1000}$ is defined. Note that the volume V is constant. Thus, for an ideal gas it holds

$$p_{press}(C,T) = \|C\|_1 \cdot R_m \cdot T \qquad \text{with the ideal gas constant } R_m\,.$$

Therefore, the total concentration $C_{p_s}(T) \in \mathbb{R}$ that corresponds to the standard pressure $p_s = 1\text{bar}$ can be computed. Furthermore, the standard temperature is denoted by $T_s := 298.15\text{K}$ (= 25°Celsius). According to [51], the chemical potential is given by

$$\begin{aligned}\mu_k = \mu_k(C,T) = A_k^S \cdot (\tilde{t} - \tilde{t}\ln(\tilde{t})) - \frac{B_k^S}{2}\cdot\tilde{t}^2 - \frac{C_k^S}{6}\cdot\tilde{t}^3 - \frac{D_k^S}{12}\cdot\tilde{t}^4 - \frac{E_k^S}{2\cdot\tilde{t}} + F_k^S - G_k^S\cdot\tilde{t} \\ + H_{mk}(T_s) - \Delta_f H_k(T_s) + R_m\cdot\tilde{t}\cdot\ln\left(\frac{C_k}{C_{p_s}}\right).\end{aligned} \tag{5.4}$$

In equation (5.4) A_k^S to G_k^S are the species-dependent coefficients of the Shomate equation and can be taken from databases like NIST Standard Reference Database Number 69 [83]. The standard enthalpy $H_{mk}(T_s)$ and the enthalpy of formation of gas at standard conditions $\Delta_f H_k(T_s)$ can as well be taken from NIST Standard Reference Database Number 69 [83]. Due to equation (5.4) the chemical potential $\mu_k = \mu_k(C,T)$ can be written as the sum of a concentration-independent term $\mu_k(T)$ and a concentration-dependent term. It holds

$$\mu_k = \mu_k(C,T) = \mu_k(T) + R_m\cdot\tilde{t}\cdot\ln\left(\frac{C_k}{C_{p_s}}\right). \tag{5.5}$$

Note that $\mu_k(C,T) \neq \mu_k(T)$. According to [51], the chemical system is in equilibrium regarding a direction $q \in \mathbb{R}^{m_s}$ if

$$(\mu_k)_k \cdot q = \sum_{k=1}^{m_s} \mu_k(C,T)\cdot q_k \overset{!}{=} 0 \tag{5.6}$$

Equation (5.6) is illustrated by Example 8.

Example 8. *Thermodynamics is used in order to compute the partial equilibrium of a fast forward and backward reaction. We regard the chemical system*

$$\begin{aligned} HCl + H\cdot &\rightarrow_{k_1} H_2 + Cl\cdot \\ H_2 + Cl\cdot &\rightarrow_{k_2} HCl + H\cdot. \end{aligned} \tag{5.7}$$

Thereby the concentration vector is $C = \Big([HCl],\ \ [H\cdot],\ \ [H_2],\ \ [Cl\cdot]\Big)$. *Then the considered direction* q *is* $q = (1,1,-1,-1)^T$. *The temperature is the standard temperature* $T_s = 298.15\text{K} = 25°\text{C}$. *Furthermore, the standard pressure* C_{p_s} *is not required because the number of educts is equal to the number of products for the examined chemical reaction. In the following the equilibrium of this reaction is computed by* (3.2) *and by thermodynamics. In order to use* (3.2) *for computation of the equilibrium of* (5.7)*, the reaction*

rate constants are taken from the NIST Standard Reference Database Number 17 [93]. For the considered temperature T one obtains

$$k_1 = 3.75 \cdot 10^{-14} \frac{\mathrm{cm}^3}{\mathrm{molecules} \cdot \mathrm{s}}, \qquad k_2 = 1.18 \cdot 10^{-14} \frac{\mathrm{cm}^3}{\mathrm{molecules} \cdot \mathrm{s}}.$$

Therefore, the considered reaction is in partial equilibrium if

$$\frac{C_1 C_2}{C_3 C_4} \overset{!}{=} \frac{1.18}{3.75} \approx 0.315. \tag{5.8}$$

Note that the reaction rate constants are not exact. The database lists several different values, which result in slightly different approximations of the partial equilibrium. The data from [2] are used because the reference gives reaction rate coefficients for forward and backward reaction. This seems to be more reliable than using two different references for forward and backward reaction. Thus, equation (5.8) *is not exact, and thermodynamics will result in a slightly different relation.*

The partial equilibrium of the considered reaction can also be determined with thermodynamics. The coefficients of the Shomate equation are obtained from the NIST Standard Reference Database Number 69 [83]. Then the chemical potentials $\mu_k(T)$, $k \in \{1,2,3,4\}$, can be computed. They are listed in Table 3.

	HCl	$H\cdot$	H_2	Cl
μ(298.15K)	-148.04	183.80	-38.97	72.05

Table 3: Chemical potential of some species for $T = 298.15\mathrm{K}$

After that the following relation is derived:

$$\begin{aligned} \sum_{k=1}^{4} q_k \mu_k(C,T) &\overset{!}{=} 0, \\ \frac{2.6789}{R_m \cdot \tilde{t}} &\overset{!}{=} \ln\left(\frac{C_3 C_4}{C_1 C_2}\right), \\ \frac{C_1 C_2}{C_3 C_4} &\overset{!}{=} e^{-\frac{2.6789}{R_m \cdot \tilde{t}}} \approx 0.339. \end{aligned} \tag{5.9}$$

Hence, the result (5.8) *of the standard procedure for computation of the equilibrium state and* (5.9) *computed with thermodynamics agree well.*

Example 8 demonstrates the usage of thermodynamics for the computation of the partial equilibrium of a chemical reaction. Note that all involved chemical reactions are in equilibrium. In comparison QSSA does not depend on the equilibrium of forward and backward reaction. Therefore, thermodynamics is not applicable for the computation of the quasi-steady state of a species. The following example verifies this statement.

Example 9. *A fictitious chemical reaction system is examined. The corresponding differential equation is*

$$\begin{aligned} A_1 \to A_2, \quad & k_1 = 1, \quad & A_2 \to A_1, \quad & k_2 = 1000, \\ A_3 \to A_2, \quad & k_3 = 1, \quad & A_2 \to A_3, \quad & k_4 = 1000\,. \end{aligned}$$

The chemical reaction system results in a linear differential equation

$$C' = \begin{pmatrix} -k_1 & k_2 & 0 \\ k_1 & -k_2 - k_4 & k_3 \\ 0 & k_4 & -k_3 \end{pmatrix} \cdot \begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix} =: F_C C =: F(C)\,.$$

Then a direction of fast decay can be found by Schur-decomposition of the Jacobian matrix F_C. The direction of fast decay is $q = \begin{pmatrix} 1, & -2, & 1 \end{pmatrix}$. Thereby $qF(C) \approx 0$ is equivalent to $C_2' \approx 0$. The quasi-steady state assumption is fulfilled for the species A_2. Note that solving for C_2 induces

$$C_2 \approx \frac{k_1 \cdot C_1 + k_3 \cdot C_3}{k_2 + k_4} = \frac{C_1 + C_3}{2000}, \tag{5.10}$$

which is a very good approximation of C_2. Hence, the QSSA is a valid assumption, and the following results are not caused by an unjustified usage of QSSA.
The considered reaction system is an artificial system. Thus, the chemical potentials of the involved species are not available. However, it is possible to compute values for the chemical potentials, which match the given reaction system. The partial equilibrium of the forward and backward reactions is computed by chemical kinetics (3.2) *and by thermodynamics* (5.6)*, which results in*

$$\begin{aligned} 1000 = \frac{k_2}{k_1} = e^{\frac{\mu_{A_2}(T) - \mu_{A_1}(T)}{R_m \cdot \tilde{t}}} \quad & \Longrightarrow \quad \mu_{A_2}(T) - \mu_{A_1}(T) = R_m \cdot \tilde{t} \cdot \ln(1000)\,, \\ 1000 = \frac{k_4}{k_3} = e^{\frac{\mu_{A_2}(T) - \mu_{A_3}(T)}{R_m \cdot \tilde{t}}} \quad & \Longrightarrow \quad \mu_{A_2}(T) - \mu_{A_3}(T) = R_m \cdot \tilde{t} \cdot \ln(1000)\,. \end{aligned}$$

Moreover, the chemical potential of the species A_i is given by

$$\mu_{A_i}(C,T) := \mu_{A_i}(T) + R_m \cdot \tilde{t} \cdot \ln\left(\frac{C_i}{C_{p_s}}\right).$$

Therefore, the chemical potentials are determined up to a constant. Now the concentration C_2 is computed with thermodynamics. If the system reaches equilibrium in the direction of fast decay $q = (1, -2, 1)^T$, it holds

$$\sum_{i=1}^{3} \mu_i(C,T) \cdot q_i = R_m \cdot \tilde{t} \cdot \left(-2\ln(1000) + \ln\left(\frac{C_1}{C_{p_s}}\right) - 2\ln\left(\frac{C_2}{C_{p_s}}\right) + \ln\left(\frac{C_3}{C_{p_s}}\right)\right) \overset{!}{=} 0$$

Hence, equilibrium in direction q implies

$$(C_2)^2 = \frac{C_1 C_3}{1000^2}\,. \tag{5.11}$$

Equation (5.11) *holds for the steady state of the complete system ($C_1 = C_3$). However, if both pairs of forward and backward reactions are not in partial equilibrium ($C_1 \neq C_3$), it does not hold (compare to the quasi-steady state approximation* (5.10) *of C_2). Therefore, thermodynamics is suited for the computation of the partial equilibrium of a reaction, but thermodynamics is not applicable for computation of quasi-steady state of a species.*

Remark 5.2. *If a direction of fast decay is determined with an automatic reduction mechanism that is based on eigen- or Schur-decomposition, it is not possible to distinguish between PEA and QSSA. For instance in Example 9 the direction of fast decay is identified with Schur-decomposition of the Jacobian matrix. Hence, thermodynamics cannot be used for automatic reduction mechanisms that are based on eigen- or Schur-decomposition.*

Remark 5.3. *Note that the chemical potential of a species can be identified by the consideration of an arbitrary chemical reaction that involves the species as educt or product. Thus, the chemical potential is available for nearly all species. Hence, the ratio between the reaction rate coefficients of forward and backward reaction is given for nearly all reactions, especially if they are part of a fast chemical reaction system. Therefore, one parameter describes each pair of forward and backward reaction in the ODE* (5.1). *If the coefficients of the Arrhenius equation* (2.6) *are the unknown parameters, the relation of the reaction rate constants at different temperatures can be used in order to compute the Arrhenius parameters of the backward reaction as a function of the parameters of the forward reaction.*

5.4 Exploiting different timescales for parameter identification

In this section, a routine to exploit different timescales for parameter identification of unknown reaction rate constants is developed. Thereby the existence of unknown parameters is an obstacle. The existence of unknown reaction rate constants makes it difficult to determine the fast chemical processes. As already discussed the Jacobian matrix of the chemical source term is parameter dependent. Therefore, the eigen-decomposition of the Jacobian matrix depends on the unknown parameters, and it has to be computed for every parameter set. Hence, automatic reduction mechanisms like the ILDM method cannot be used for parameter identification. Even if the fast directions are given, the corresponding algebraic equations depend on the parameter p. Thus, the look-up table has to be computed for every used parameter p. This is more expensive than the computation of one solution $C(p)$ of the differential equation (5.1). Instead the PEA for single reactions and the QSSA for single species is used. The choice of fast chemical reactions or species in quasi-steady state is discussed in Subsection 5.4.1. Each species in quasi-steady state and each reaction in partial equilibrium results in an algebraic equation. If the corresponding reaction rate constants are known, the dimension of the differential equation is easily reduced by one. However, if at least one of the corresponding reaction rate constants is unknown, the obtained equation cannot be used to eliminate one species. An approach for this is presented in Subsections 5.4.2 and 5.4.3.

5.4.1 Determining PEA and QSSA

In this subsection, the choice of species in quasi-steady state and reactions in partial equilibrium is examined.

First, the identification of reactions in partial equilibrium is discussed. The chemical source term (3.1) consists of contributions of several reactions. Each reaction includes a forward and a backward reaction because we consider reversible elementary reactions. Then the reaction rate of a single reaction is a sum of a positive and a negative term. After suited sorting of the chemical reactions the backward and the forward reaction are characterized by $r_{2j-1}\nu_{2j-1}(C,T,p)$ and $r_{2j}\nu_{2j}(C,T,p) = -r_{2j-1}\nu_{2j}(C,T,p),\ j \leq m_r/2$. Then a frequently used method for the identification of reactions in partial equilibrium is to consider the fraction

$$\frac{\nu_{2j-1}(C,T,p) - \nu_{2j}(C,T,p)}{\nu_{2j-1}(C,T,p) + \nu_{2j}(C,T,p)}. \tag{5.12}$$

If the absolute of the fraction is small, the corresponding reaction can be considered as a partial equilibrium reaction. However, this method is not applicable if the reaction rate constant k_{2j-1} or k_{2j} is unknown. In this case the unknown parameters handicap the computation of an a-priori approximation of the magnitude of the fraction (5.12). In [97] an approach is introduced, which computes the timescale of each reaction. Thereby it is assumed that every reaction is at most first order in each species. If a chemical species is consumed by a chemical reaction that is second order in this species, the corresponding chemical reaction is neglected in the following equation (5.13). Using equation (2.5), for the jth reaction and every occurring species A_i the reaction rate can be transformed as follows

$$\begin{aligned}\big(\nu_{2j-1}(C,T,p) - \nu_{2j}(C,T,p)\big) &= k_{2j-1}(T)\prod_{l=1}^{m_s}(C_l)^{e_{l(2j-1)}} - k_{2j}(T)\prod_{l=1}^{m_s}(C_l)^{e_{l(2j)}} \\ &=: \pm\big(P_i^j(C,T,p) - L_i^j(C,T,p)C_i\big), \qquad L_i^j \geq 0\,.\end{aligned} \tag{5.13}$$

If the species A_i is a catalyst (it is a product and an educt of the jth reaction) or if the species A_i does not occur in the jth reaction, L_i^j is set equal zero. Note that L_i^j is a sum of positive summands. Hence, if the magnitude of the unknown parameter is given, the size of L_i^j can be approximated. The timescale τ^j of the jth reaction is

$$\tau^j = \max_{i\,:\,L_i^j \neq 0} \frac{1}{L_i^j(C,T,p)}. \tag{5.14}$$

Afterwards each timescale is compared to the time period of interest in order to determine reactions in partial equilibrium. Note that the timescale depends on the parameter p. However, if an approximation of the true parameter p^* exists, it is possible to obtain an approximation of each timescale τ_j. Hence, if the corresponding reaction rate constants are exactly given, chemical reactions in partial equilibrium can be identified by consideration of (5.12). However, if only an approximation of the chemical reaction

rate constants is given, partial equilibrium can be detected by comparison of the timescale (5.14) with the time period of interest (step size).

In order to determine the species in quasi-steady state several different selection criteria exist. For example, a species can be considered being in quasi-steady state if the sum of rates of consuming and production reactions is much higher than its net production rate. However, if at least one involved reaction rate constant is unknown, the exact rates of consuming and production reactions are unknown. Therefore, cancellation of consuming and producing reaction rates is hard to verify in parameter identification, and this criteria is not generally applicable. A rigorous choice of species in quasi-steady state is illustrated in [131]. A simple method to find the species in QSSA is to assume that the species with a short characteristic timescale are in quasi-steady state. Assume that for each species A_i there are no terms that are second order in the considered species. Otherwise, similar to equation (5.13), consuming higher-order reactions are neglected. Then the differential equation for C_i can be transformed to

$$C_i' = P_i(C,T,p) - L_i(C,T,p)C_i$$

with functions P_i and L_i independent of C_i. The timescale of the species A_i is

$$\tau_i(C,T,p) = \frac{1}{L_i(C,T,p)}\,.$$

Note that the timescale depends on the parameter p. However, similarly to the timescale analysis of chemical reactions (in partial equilibrium), an approximation of the timescales can be obtained. After computing the (approximate) characteristic timescales the timescales of the different species can be compared with the time period of interest in order to determine species in quasi-steady state.

So, the selection rules of this subsection can be used in order to choose reactions in partial equilibrium and species in quasi-steady state in parameter identification. Thereby chemical reactions in partial equilibrium and species in quasi-steady state can be detected if the magnitude of the unknown parameters is known.

Remark 5.4. *The results of Subsection 5.4.1 do not depend on the assumption that the temperature is fixed in the reactor. Hence, PEA and QSSA can also be determined for the identification of the parameters of the Arrhenius equation.*

5.4.2 Using thermodynamic description for PEA

In this subsection, the usage of PEA combined with thermodynamics for parameter identification is discussed. For each chemical reaction, the thermodynamic description results in the ratio between the reaction rate constants of forward and backward reaction. Thus, if only one of the two reaction rate constants is unknown, the missing parameter is immediately obtained and the dimension of the parameter space descreases by one. However, if both reaction rate constants are unknown, forward and backward reaction rate can be described by one parameter because the ratio between the reaction rate constants is given. Thereby the reduction of the dimension of the parameter space does not depend on the PEA.

In case of a reaction in partial equilibrium, it is not sufficient to reduce the dimension of the parameter space by one. The PEA states that the considered chemical reaction is in partial equilibrium during the considered time span. If the chemical reaction is in partial equilibrium, different parameter choices with the same ratio and large magnitude result in the same trajectory (see Example 10). Hence, the remaining parameter cannot be determined uniquely, and the optimization routine will not converge, because the problem is ill-conditioned. In order to prevent numerical difficulties, both parameters have to be eliminated. Thereby the unknown parameters of the chemical reaction in partial equilibrium are chosen such that they are larger than a threshold value, and such that the ratio between them is in agreement with the thermodynamic data. If an approach of Remark 3.1 is used, the dimension of the differential equation is not decreased. Although, the dimension of the parameter space is still decreased, and the stiffness related to the fast reactions in partial equilibrium is eliminated.

Example 10. *Consider a modification of Example 6. The reaction system is given by*

$$\begin{aligned} A_1 &\to A_2, \qquad p_1 = k_1 = 10^2, \\ A_2 &\to A_1, \qquad p_2 = k_2 = 10^2. \end{aligned}$$

Thereby the reaction rate constants are the unknown parameter set (p_1, p_2)*. The state of the system depends on the time* t *and the parameter* p*. It is denoted by* $C(t,p)$*. Moreover, assume that the initial value is* $C(0,p) = \begin{pmatrix} 1, & 0 \end{pmatrix}$ *and that the experimental data, which have to be fitted, is given at* $t_1 = 0.1$*. Furthermore, the measurement error is smaller than* 10^{-6}*, which is unrealistic small. Thermodynamics states that* $\frac{p_1}{p_2} = 1$*, such that the dimension of the parameter space can be reduced by one. We define* $p = p_1 = p_2$*. Then the parameter-dependent solution trajectory of the occurring differential equation is*

$$\begin{pmatrix} C_1(t,p) \\ C_2(t,p) \end{pmatrix} = \begin{pmatrix} \frac{1}{2} + \frac{e^{-2pt}}{2} \\ \frac{1}{2} - \frac{e^{-2pt}}{2} \end{pmatrix}.$$

Hence, it holds

$$\|C(t_1,p) - C(t_1,\tilde{p})\|_\infty = |0.5(e^{-0.2p} - e^{-0.2\tilde{p}})| \qquad \text{for } p, \tilde{p} \in \mathbb{R}.$$

Then the distance of $C(t_1, 100)$ *and* $C(t_1,p)$ *is smaller than the measurement error for* $p \geq 65.6$*. Thus, every parameter* $p \geq 65.6$ *fits the (perturbed) measurement, and is a possible value for the unknown parameter. Therefore, the parameters* $p_1 = p_2$ *cannot be determined. Although, it is possible to choose an arbitrary* $p = p_1 = p_2 \geq 65.6$*. Thereby the threshold depends on the size of the measurement error. Furthermore, it depends on the time step of the numerical integration method, the time* t_i*,* $i \geq 1$*, of the measurements, and the corresponding tolerances.*

According to Example 10, it is possible to estimate the reaction rate constants of all reactions in partial equilibrium before minimizing the objective function, which defines the best fit to the given measurement data. The dimension of the parameter space is decreased, and the parameter identification

is accelerated. Note that the elimination of the corresponding unknown parameters is necessary if the state of the system is projected onto a manifold that is defined by the fast chemical reactions. Otherwise the estimated parameters may result in a trajectory that does not fit the measurements.

Example 11. *Consider Example 10. Then the experimental measurements are given by* $C_1(t) \approx C_2(t) \approx$ *const because the reaction exhausts very fast and the equilibrium of the system is reached instantly. Furthermore, the system is projected onto a low-dimensional manifold that is computed with the PEA for the fast chemical reaction. The reaction contains two unknown reaction rate constants, such that thermodynamics is used to identify the (correct) one-dimensional manifold* $C_1 = C_2$ *and the relation* $p_1 = p_2$*. Hence, the slow subspace is spanned by*

$$Q_s = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}.$$

Introducing a slack variable x_f *and projecting the chemical system on the slow subspace results in*

$$\begin{aligned} \begin{pmatrix} C_1 \\ C_2 \end{pmatrix} &= \begin{pmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \end{pmatrix} \begin{pmatrix} -pC_1 + pC_2 \\ +pC_1 - pC_2 \end{pmatrix} + \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix} x_f, \\ 0 &= C_1 - C_2 \end{aligned} \tag{5.15}$$

The unknown parameter set p *does not appear in the reduced system. Moreover, the solution of* (5.15) *matches the given measurements for all parameter sets* p*. However,* p *has to be large enough (see Example 10) in order to result in the given trajectory. Therefore, consideration of the reduced differential equation, which is restricted to the one-dimensional manifold, is not suited for identification of the parameter* p*.*

The previous examples show that all parameters that are associated with chemical reactions in partial equilibrium can be determined separately. Thereby chemical reactions in partial equilibrium are determined by the selection rules in Subsection 5.4.1. If the magnitude of the unknown parameters is known, the selection rules are feasible. Hence, the dimension of the parameter space is decreased. Furthermore, chemical reactions in partial equilibrium result in algebraic equations (by thermodynamics or by using the estimation of the corresponding parameters), which can be used to reduce the dimension of the differential equation. The reduced differential equation can be used for parameter identification of the remaining parameters.

Remark 5.5. *It is possible, that a slow chemical reaction lays in the equilibrium reaction subspace. Then the reaction rate constant of the slow reaction cannot be determined because the corresponding slow reaction is also in partial equilibrium.*

Remark 5.6. *If the temperature is not fixed, similar results hold.*

5.4.3 Using QSSA

Thermodynamics combined with the QSSA must not be used for obtaining an algebraic equation, which reduces the dimension of the differential equation (see Example 9). Thus, if a species A_i is in quasi-steady

state, all involved reaction rate constants are necessary in order to eliminate C_i from the differential equation. In this case it is straightforward to derive an algebraic equation for C_i, which can be solved numerically. However, if there are unknown reaction rate constants involved, it is impossible to compute the concentration of the species in quasi-steady state as a p-independent function of the other species.

Then again a quasi-steady state equation gives a relation between the involved reaction rate constants, but the relation depends on the state of the system. Hence, the dimension of the parameter space cannot be decreased without any additional information [99]. Assume that the given measurements contain the concentration of all species at certain points in time. This assumption allows to use linear regression in order to reduce the dimension of the parameter space.

Example 12. *The dissociation of dinitrogen pentoxide (see Example 5, [52]) is considered in order to illustrate the linear regression. The initial values are*

$$\begin{aligned} &[N_2O_5](0) = 1.6 \cdot 10^{18} \frac{\text{molecules}}{\text{cm}^3}, \qquad [NO_2](0) = 10^{16} \frac{\text{molecules}}{\text{cm}^3}, \qquad [O_2](0) = 0 \\ &[NO_3](0) = 0, \qquad\qquad\qquad\quad [NO](0) = 0\,. \end{aligned}$$

Moreover, there are three measurements available

$$\begin{aligned} C^m(t_i) = &\left([N_2O_5]^m(t_i),\ \ [NO_2]^m(t_i),\ \ [O_2]^m(t_i),\ \ [NO_3]^m(t_i),\ \ [NO]^m(t_i)\right)^T, \\ &t_i \in \{10^2\text{s},\ 10^3\text{s},\ 10^4\text{s}\}\,. \end{aligned}$$

*Thereby the measurements contain 2.5% normal distributed noise and the measurements are listed in Appendix **A.1**. Note that the concentrations of all species are measured. In general, this is not fulfilled for an experiment. The given measurements $C^m(t_i)$, $t_i \in \{10^2\text{s},\ 10^3\text{s},\ 10^4\text{s}\}$, are used in order to estimate all reaction rate constants. Hence, the number of unknown parameters is four. As already shown in Example 5, the species NO_3 is in quasi-steady state, which results in*

$$\begin{aligned} [NO_3]' &= k_1[N_2O_5] - k_2[NO_2][NO_3] - k_3[NO_2][NO_3] - k_4[NO][NO_3] \\ &\stackrel{!}{=} 0\,. \end{aligned}$$

The minimum of the number of measurements and the number of species in quasi-steady state is one. Hence, the dimension of the parameter space can be reduced by one. Then the QSSA and the measurements induces

$$\begin{aligned} \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} &\stackrel{!}{\approx} \begin{pmatrix} [N_2O_5]^m(t_1) \\ [N_2O_5]^m(t_2) \\ [N_2O_5]^m(t_3) \end{pmatrix} k_1 - \begin{pmatrix} [NO_2]^m(t_1)[NO_3]^m(t_1) \\ [NO_2]^m(t_2)[NO_3]^m(t_2) \\ [NO_2]^m(t_3)[NO_3]^m(t_3) \end{pmatrix} (k_2 + k_3) - \begin{pmatrix} [NO]^m(t_1)[NO_3]^m(t_1) \\ [NO]^m(t_2)[NO_3]^m(t_2) \\ [NO]^m(t_3)[NO_3]^m(t_3) \end{pmatrix} k_4 \\ &:= A_1 k_1 - A_2(k_2 + k_3) - A_3 k_4\,. \end{aligned}$$

Due to measurement errors the linear system of equations has no solution. However, linear regression can be used in order to minimize the error regarding the norm $\|\cdot\|_2$. Then, one obtains

$$\begin{aligned} k_1 &\approx \left(A_1^T \cdot A_1\right)^{-1} \cdot A_1^T \cdot \left(A_2 \cdot (k_2 + k_3) + A_3 \cdot k_4\right) \\ &= 5.3015 \cdot 10^{10} \cdot (k_2 + k_3) + 1.3411 \cdot 10^6 \cdot k_4 = 0.10080\,. \end{aligned}$$

The obtained result is a very good approximation of $k_1 = 0.1$. Hence, the number of unknown parameters can be reduced by one.

The previous example shows that QSSA can be used to reduce the dimension of the parameter space. Nevertheless, there are several disadvantages that have to be considered. First, a measurement of the full state of the system is necessary for linear regression. However, the concentrations of only a few species are measured in most experiments. Second, linear regression might result in severe errors if the measurement error is too large. Furthermore, the algebraic equation obtained by QSSA can only be used to precompute the manifold if all involved reaction rate constants are known. It follows that in case of unknown parameters the QSSA cannot be used for a reduction of the dimension of the differential equation. Hence, PEA combined with thermodynamics is much more useful than QSSA for parameter identification in chemical reaction systems.

In order to create a look-up table the dimension of the reduced differential equation has to be very low (two or three). However, in general the number of reactions in partial equilibrium and parameter-independent QSSA equations is not large enough for the reduction of the dimension of the system to two or three. Therefore, in general it is not possible to use a low-dimensional differential equation for parameter identification. Nevertheless, PEA combined with thermodynamics results in a dimension reduction of the parameter space, which decreases the computational costs for the computation of the sensitivity matrix and avoids the identification of ill-conditioned parameters. Under certain conditions QSSA also achieves a reduction of the dimension of the parameter space. However, the measurement data do not fulfil the necessary conditions for most experimental designs.

5.5 Numerical test cases

In this section, different numerical examples are provided. First, a small, linear example is used in order to clearify the approach for the PEA. Afterwards the parameters of the Zeldovich mechanism are estimated with the QSSA and the parameters of the dissociation of dinitrogen pentoxide are estimated with the PEA. The given measurements are denoted by $C^m(t_i),\ 1 \le i \le n_m$, and the parameter-dependent state of the system is $C(t_i, p),\ 1 \le i \le n_m$. Then the objective function is

$$e(p) := \left(\sum_{i=1}^{n_m} \sum_{j=1}^{m_s} \left(\frac{|C_j(t_i, p) - C_j^m(t_i)|}{C_j^m(t_i)} \right)^2 \right)^{0.5}.$$

For the numerical examples a very simple parameter estimator is used. The parameter estimator consists of the matlab routine *ode15s* for solving the stiff differential equations and the matlab routines *lsqnonlin* or *fminsearch* for minimization of the objective function. Moreover, the 'experimental' data are obtained

by the integration of the original differential equation with perturbation by some noise. Due to the randomness of the noise the computation is repeated one hundered times in order to determine average computational costs and an average error. In the following the initial value of the unknown parameters has the same magnitude as the true parameters. Thus, species in quasi-steady state and reactions in partial equilibrium can be identified, according to Subsection 5.4.1. Note that the methods *lsqnonlin* and *fminsearch* find a local minimum. Therefore, convergence to a global minimum is not ensured. Multistart methods [94] can be used in order to find a global minimum. However, in case of a high-dimensional parameter space using a multistart method is very expensive.

Remark 5.7. *Reduction mechanisms for chemical reaction systems are based on fast processes as well as conserved quantities. The automatic detection of fast chemical reactions (for example, eigendecomposition) depends on the reaction rate constants. Hence, automatic reduction mechanisms are not applicable in case of unknown parameters. Although, conserved quantities are independent of reaction rate constants, temperature and pressure. If the stoichiometric matrix is given, the conserved quantities of the chemical reaction system can be determined (see Remark 3.2). Furthermore, most conserved quantities result in simple linear equations. Thus, conserved quantities can be used to reduce the size of the differential equation in the parameter identification of chemical systems.*

Remark 5.8. *For simplicity a fictitious small scale example is constructed. The necessary thermodynamic data are generated similar to Example 8. Therefore, the necessary thermodynamic data are given for the (fictitious) reaction system 5.5.1.*

5.5.1 Fictitious linear system

A (fictitious) chemical system similar to those in [17] is considered:

$$\begin{aligned} &A_1 \to A_2, && k_1 = 1000, && A_2 \to A_1, && k_2 = 1000,\\ &A_3 \to A_4, && k_3 = 2000, && A_4 \to A_3, && k_4 = 2000,\\ &A_1 \to A_3, && k_5 = 2, && A_3 \to A_1, && k_6 = 1\,. \end{aligned}$$

Matching chemical potentials are given by

$$\begin{aligned} \mu_{A_1}(T) &= 0,\\ \mu_{A_2}(T) &= 0,\\ \mu_{A_3}(T) &= R_m \cdot \tilde{t} \cdot \ln(0.5),\\ \mu_{A_4}(T) &= R_m \cdot \tilde{t} \cdot \ln(0.5)\,. \end{aligned}$$

Note that $\mu_{A_i}(T)$ is the concentration-independent term of the chemical potential (compare to equation (5.5)). All rate constants are unknown, and the 'experimental' data are given by a forward integration with initial data $(6, 1, 3, 1)$ added by 2.5% normal distributed noise. The measurement points are 1s, 2s, and 3s. The initial estimation of the parameter is $k = \begin{pmatrix} 500, & 1000, & 1500, & 2000, & 3, & 3 \end{pmatrix}$. Hence,

the first and the second reaction are in partial equilibrium. Note that for this example the dimension of the differential equation can be reduced by three (one conserved quantity and two reactions in partial equilibrium), and that the low-dimensional manifold can be precomputed (due to thermodynamic data). However, the full differential equation is used in order to evaluate the objective function. Moreover, the dimension of the parameter space reduces to one because the ratio of the reaction rate constants is known for each pair of forward and backward reaction, and two reactions are in partial equilibrium. Then the parameter identification with a six-dimensional parameter space, the parameter identification with a three-dimensional parameter space (just thermodynamics for forward/backward reaction) and the parameter identification with a one-dimensional parameter space (see Subsection 5.4.2) are compared. Thereby the routine *lsqnonlin* is used. According to Table 4, the computing time is smallest with the approach using partial equilibrium and thermodynamics. Furthermore, for the six-dimensional and the three-dimensional parameter space the error is much larger than the error of parameter identification with a one-dimensional parameter space because the optimization routine with the three-dimensional as well as the six-dimensional parameter space terminates in a local minimum. Hence, PEA is applied successfully for the reduction of the dimension of the parameter space.

	one-dimensional	three-dimensional	six-dimensional
computing time	0.740	1.430	3.170
$e(p)$	0.085	0.206	0.186

Table 4: Average computational costs and average error of parameter identification with different dimension of the parameter space

5.5.2 Zeldovich mechanism - QSSA

This example is based on the Zeldovich mechanism. The Zeldovich mechanism describes the formation of nitrogen oxides. The reaction mechanism is

$$O + N_2 \underset{k_2}{\overset{k_1}{\rightleftharpoons}} NO + N, \qquad N + O_2 \underset{k_4}{\overset{k_3}{\rightleftharpoons}} NO + O, .$$

The temperature is fixed at $T = 1500\text{K}$. Then the reaction rate constants [93] are

$$k_1 = 2.54 \cdot 10^{-21}, \qquad k_2 = 4.20 \cdot 10^{-11}, \qquad k_3 = 2.54 \cdot 10^{-12}, \qquad k_4 = 1.19 \cdot 10^{-17},$$
$$[k_i] = \frac{\text{cm}^3}{\text{molecules} \cdot \text{s}} \quad \text{for } 1 \leq i \leq 4\,.$$

Furthermore, the initial values correspond to air with some oxygen atoms (79% N_2, $0.999 \cdot 21\%$ O_2, $0.001 \cdot 21\%$ O) at 1atm and are given by

$$[N_2](0) = 3.87 \cdot 10^{18} \frac{\text{molecules}}{\text{cm}^3}, \qquad [O_2](0) = 1.03 \cdot 10^{18} \frac{\text{molecules}}{\text{cm}^3}, \qquad [O](0) = 1.03 \cdot 10^{15} \frac{\text{molecules}}{\text{cm}^3}$$

and zero for all other species. The considered time interval is [0s, 1000s]. Then the species N is in quasi-steady state. It holds

$$N' = \left(k_1[O][N_2] + k_4[NO][O]\right) - \left(k_2[N][NO] + k_3[N][O_2]\right) \overset{!}{\approx} 0,$$
$$[N] \overset{!}{\approx} \frac{k_1[O][N_2] + k_4[NO][O]}{k_2[NO] + k_3[O_2]}.$$

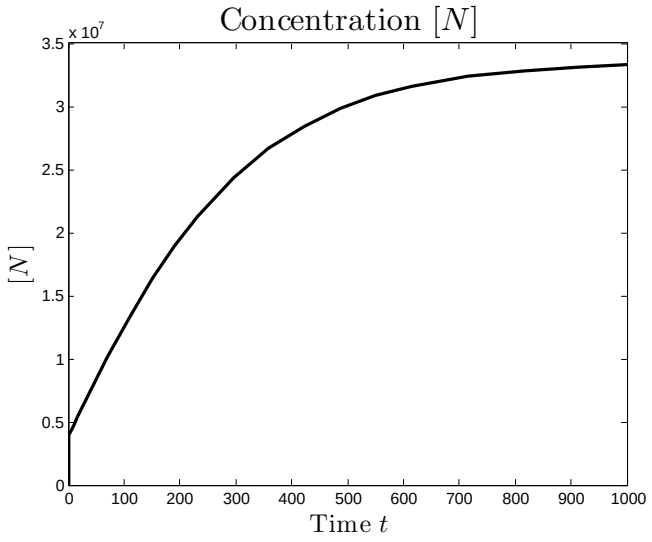

Figure 9: Temporal development of the concentration $[N]$

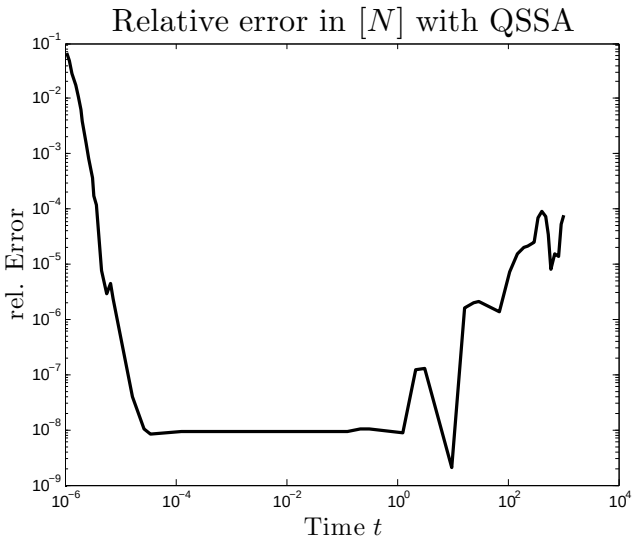

Figure 10: Relative error for the QSSA of $[N]$

In Figure 9 the temporal development of the species N is given. The temporal development of N shows that the considered time interval covers the longest timescale of the Zeldovich mechanism. Moreover, in Figure 10 the relative error

$$\frac{\left|[N] - \frac{k_1[O][N_2] + k_3[NO][O]}{k_2[NO] + k_3[O_2]}\right|}{[N]}$$

is plotted for $t > 10^{-6}$s. Figure 10 shows that the QSSA results in a very good approximation of $[N]$ after the fast transient phase.

Now, a parameter identification for the reaction rate constants k_i, $i \in \{1, 2, 3, 4\}$ is performed. The initial estimation of the parameter is $\tilde{k} = \left(10^{-20}, \quad 10^{-10}, \quad 10^{-10}, \quad 10^{-15}\right)$. In the following the quotient between the reaction rate constants of the forward and backward reactions is given by thermodynamics. Then, the reaction rate constants k_2 and k_4 are linear functions of k_1 and k_3. Three measurements at $t_i \in \{100\text{s}, 500\text{s}, 1000\text{s}\}$ are available. Thereby the measurements contain 2.5% normal distributed noise. Note that the concentrations of all species are measured. Then the original parameter space is four-dimensional. Thermodynamic data can be used in order to decrease the dimension by two and the QSSA with linear regression can be used in order to obtain a one-dimensional parameter space. Finally,

the parameter identification with a four-dimensional parameter space, the parameter identification with a two-dimensional parameter space (just thermodynamics) and the parameter identification with a one-dimensional parameter space (see Subsection 5.4.3) are compared. Thereby the routine *fminsearch* is used. According to Table 5, the computing time is smallest with the approach using QSSA and thermodynamics. However, the parameter identification with the full parameter space results in a slightly smaller average error than the other two approaches.

	one-dimensional	two-dimensional	four-dimensional
computing time	0.881	1.143	7.525
$e(p)$	0.244	0.260	0.197

Table 5: Average computational costs and average error of parameter identification of the Zeldovich mechanism with different dimension of the parameter space

5.5.3 Dissociation of dinitrogen pentoxide - PEA

In [52] the dissociation of dinitrogen pentoxide is examined. For the considered temperature $T_s = 298.15\text{K}$, the chemical reaction system is

$$\begin{aligned}
&N_2O_5 \leftrightarrow NO_2 + NO_3, && k_1 = 0.10, && k_{-1} = 1.90 \cdot 10^{-12},\\
&N_2O_4 \leftrightarrow NO_2 + NO_2, && k_2 = 3.69 \cdot 10^{6}, && k_{-2} = 1.00 \cdot 10^{-12},\\
&NO_2 + NO_3 \rightarrow NO_2 + O_2 + NO, && k_3 = 6.56 \cdot 10^{-16}, &&\\
&NO + NO_3 \rightarrow NO_2 + NO_2, && k_4 = 2.60 \cdot 10^{-11}. &&
\end{aligned}$$

Thereby the unit of the reaction rates of first order reactions is s^{-1}, and the unit of the reaction rates of second order reactions is $\frac{\text{cm}^3}{\text{s}\cdot\text{molecules}}$. The reaction rate constants are taken from the NIST Chemical Kinetics Database [93]. Note that the NIST Chemical Kinetics Database lists the Arrhenius coefficients of different references. Moreover, the results of different references differ clearly. Hence, several different reaction rate constants are listed in the NIST Chemical Kinetics Database. In this numerical example the reaction rate constant $k_2 = 3.69 \cdot 10^6$ is the average between the data of [6] and [24].
The initial value of N_2O_5 is $1.6 \cdot 10^{18} \frac{\text{molecules}}{\text{cm}^3}$, the initial value of N_2O_4 is $10^{16} \frac{\text{molecules}}{\text{cm}^3}$, and the initial value of NO_2 is $10^{16} \frac{\text{molecules}}{\text{cm}^3}$. The initial concentrations of the other species are zero. Furthermore, the state of the system is measured at $t_i \in \{10^2\text{s},\ 10^3\text{s},\ 10^4\text{s},\ 5 \cdot 10^4\text{s}\}$. Thereby 2.5% normal distributed noise is added. In the following we assume that the reaction rate constants k_2 and k_{-2} are unknown.

Thermodynamics can be used in order to compute k_2/k_{-2} (k_{-2} is given by a function of k_2).Thereby the dimension of the parameter space is reduced by one. The necessary chemical potentials are given in Table 6. It holds

$$\frac{k_2}{k_{-2}} = \frac{[NO_2][NO_2]}{[N_2O_4]} \overset{!}{=} C_{p_s}(T_s) \exp\Big(\frac{-81.671 + 2 \cdot 38.471}{R_m\ T_s/1000}\Big) = 3.6057 \cdot 10^{18} \frac{\text{molecules}}{\text{cm}^3}$$
$$\text{with } C_{p_s}(T_s) = \frac{p_s}{R_m \cdot T_s} = 2.429 \cdot 10^{19} \frac{\text{molecules}}{\text{cm}^3}.$$

	NO_2	N_2O_4
$\mu(T_s)$	-38.471	-81.671

Table 6: Chemical potential of necessary species in the dissociation of dinitrogen pentoxide

The parameter $k_2 = 3.6057 \cdot 10^{18} k_{-2}$ is not identifiable due to the partial equilibrium of the second reaction (see Example 10). Hence, any (large enough) value can be chosen for k_2. Table 7 shows the error $e(p)$ for the parameter identification with different choices for k_2. The prior choice of the parameter k_2 does not influence the error $e(p)$ if k_2 is large enough. Clearly, the estimated trajectory does not depend on the size of the parameter k_2. Therefore, the exact value of $k_2 = 3.6057 \cdot 10^{18} k_{-2}$ is not calculable. Similar results are obtained by sensitivity analysis. Thereby the sensitivity of the state of the system to the values of single parameters is approximated.

	$k_2 = 10^{-4}$	$k_2 = 10^{-2}$	$k_2 = 0.1$	$k_2 = 1$	$k_2 = 10^4$	$k_2 = 10^8$	$k_2 = 10^{11}$
$e(p)$	4.091938	1.064272	0.131214	0.131222	0.134825	0.130743	0.130053

Table 7: Average error for different fixed parameter estimations of k_2

5.6 Conclusions and summary

In this chapter, the occurrence of different timescales in chemical reaction systems has been used for parameter identification. For parameter identification the PEA can be combined with thermodynamics. Thereby the dimension of the parameter space is reduced, and algebraic equations that hold for all parameter sets in the modified parameter space are obtained. In comparison the QSSA does not result in parameter independent algebraic equations for the state of the system. Although, QSSA can be combined with linear regression in order to reduce the dimension of the parameter space. However, linear regression requires measurements of all involved species.

In general algebraic equations restrict the state of the system onto a low-dimensional manifold. This manifold can be used in order to generate a low-dimensional approximation of the original differential equation. Thereby the precomputed manifold is saved in a look-up table. This is only possible for a low-dimensional manifold due to the curse of dimensionality. However, in general the number of chemical reactions in partial equilibrium is not large enough to obtain a low-dimensional manifold. Hence, precomputing the low-dimensional manifold is impractical for parameter estimation, and the computational costs of each evaluation of the objective function are not reduced. But the number of necessary evaluations of the objective function correlates with the dimension of the parameter space, and the usage of the PEA and the QSSA results in a reduction of the dimension of the parameter space. Thus, the total computing time is reduced by the occurrence of fast chemical processes. Furthermore, the identification of ill-conditioned parameters is avoided.

Acknowledgement: This chapter was published (with few changes) as an article [87] in *Journal of Computational and Applied Mathematics*, vol. 344, Axel Ariaan Lukassen and Martin Kiehl, Operator splitting for chemical reaction systems with fast chemistry, pp. 495–511, Copyright Elsevier (2018).

6.1 Introduction

In Chapters 3 to 5, spatially homogeneous reactors, which are described by ODEs, are examined in order to simplify the analysis. Operator splitting methods solve a sequence of transport-only and chemistry-only differential equations. Thereby the chemistry-only equations correspond to homogeneous reactors. Hence, if a heterogeneous system is solved with an operator splitting method, the results of Chapters 3 to 5 can be applied. Furthermore, operator splitting is frequently used for large scale engineering problems and for chemical reaction systems with transport. Applications are pollution transport in the atmosphere [12, 68], combustion [139], transport in groundwater systems [1, 138], and many others. The most common splitting schemes are the Lie-Trotter splitting of order one and the Strang splitting [69, 124] of order two. However, in [121, 122, 141] it is shown that the Strang splitting suffers from order reduction in the stiff case. Therefore, practical estimation of the splitting error might fail, and the possible step size of the splitting method is prohibitive for the numerical computation of many problems. The Richardson extrapolation (in general [34, 111, 112], for splitting methods [47, 80, 137]) of the Lie-Trotter splitting is a second order scheme. In this chapter we show that the Richardson extrapolation does not suffer from order reduction if all stiffness is related to one splitting term. Furthermore, we examine stability of the extrapolated splitting scheme.

We consider the application of the splitting schemes to chemical reaction systems with transport. The corresponding differential equation is a partial differential equation, but after discretization in space we obtain an ordinary differential equation

$$\frac{\mathrm{d}C}{\mathrm{d}t} = F(C) + G(C), \qquad C(t_0) \text{ is given}. \tag{6.1}$$

Here $F(C)$ is the chemical source term, and $G(C)$ models the transport. In general $F(C)$ is a nonlinear polynomial of degree two in case of elementary reactions. The number of chemical species is m_s and the number of spatial grid points is m_g. Thus, the total dimension of equation (6.1) is $m_s \cdot m_g$. The unknown variables can be sorted such that the Jacobian matrix of $F(C)$ is a block diagonal matrix with m_g blocks of dimension m_s.

In the following we assume that the transport term only corresponds to slow processes. Therefore, solving a transport equation does not require implicit integration schemes. Furthermore, there exist specialised solvers for transport-only equations. These solvers can solve the differential equation $\frac{\mathrm{d}C}{\mathrm{d}t} = G(C)$ easily. Moreover, transport introduces the coupling between the different spatial grid points. Hence, if $G(C) = 0$, equation (6.1) decomposes to m_g m_s-dimensional differential equations, which can easily be solved by implicit methods. Each m_s-dimensional ODE corresponds to a homogeneous chemical reaction system. So, neglecting $F(C)$ or $G(C)$ results in easier differential equations. Thus, operator splitting methods [31, 70, 114] are an obvious approach for equation (6.1) because operator splitting methods

solve a sequence of transport-only and chemistry-only equations in order to compute the solution of (6.1).

An advantage of operator splitting is that the dimension of chemistry-only equations can be easily reduced by reduction mechanisms (see Chapter 3 and [78, 85, 92]). Typically, solving a reduced chemistry-only equation is computational cheap. Examples for reduction mechanisms combined with an operator splitting are given in [116, 118, 119]. Some reduction mechanisms do not incorporate transport into the reduced subspace. However, the chemistry-only equation corresponds to a homogeneous chemical reaction system for each node of the spatial mesh. Thus, neglecting the transport in the computation of the reduced subspace does not result in an additional error for the reduction mechanism in the chemistry-only equation, while the error of the operator splitting is easily controlled by the time step. Furthermore, the timescales of chemical reactions depend on the concentration and the temperature. Therefore, the reduced subspace depends on time and space. If the chemistry-only equation is considered, the reduction mechanism can be applied for a suited concentration and temperature interval, whereas the full model is used for grid points in the area of slow chemistry. Hence, operator splitting results in a time and space adaptive usage of reduction mechanisms, while being easy to implement.

The drawback of operator splitting methods is the occurrence of a splitting error. The splitting error depends on the step size Δt of the operator splitting method. Similar to the error control of numerical integration methods for ODEs, the splitting error can be controlled by the usage of splitting methods with different order [73, 74]. A possible combination is the first order Lie-Trotter splitting and the second order Strang splitting [74]. For error estimation it is necessary to know the order of the splitting method. However, as already stated, the Strang splitting suffers from order reduction. Hence, the error estimator and the error control fail. Yang and Pope [141] illustrated the order reduction with a mechanism for methane/air combustion, and Sportisse [121] used singular perturbation theory [128, 142] for the analysis of order reduction in the case of linear differential equations. For the analysis of a nonlinear differential equation the Lie operator formalism [40, 79] or singular perturbation theory [76] is necessary. The Lie operator formalism is used in [32, 33] for nonlinear stiff reaction-diffusion systems. Furthermore, for a stiff and a non-stiff operator the splitting error of Lie-Trotter splitting and Strang splitting is analyzed in [76].

The structure of the chapter is the following. In Section 6.2, we recall the Lie-Trotter splitting, the Strang splitting, and the extrapolated splitting scheme. Moreover, the classical order of the considered splitting schemes is presented in Section 6.3. In Section 6.4, the splitting order for the stiff differential equation (6.1) is examined. For stiff differential equations a stiffness parameter ϵ, which corresponds to the smallest timescale, exists. It holds $\epsilon \ll \Delta t$ for reasonable step sizes Δt. However, the classical order does not apply for application of a splitting scheme to stiff differential equations of type (6.1) with $\epsilon \ll \Delta t$. In Section 6.4, we derive the stiff order of the extrapolated Lie-Trotter splitting under the assumption $\epsilon \ll \Delta t \to 0$. Thereby we show that the Richardson extrapolation of the Lie-Trotter splitting as second order splitting method does not suffer from order reduction for the stiff differential equation (6.1). This statement is proved with the singular perturbation approach. In Section 6.5, the stability of the extrapolated scheme is shown. Finally, we verify the results from Section 6.4 with some numerical examples in Section 6.6.

In this section, we introduce the Lie-Trotter splitting [130], the Strang splitting [124], and the extrapolated Lie-Trotter splitting [47]. All splitting methods are applied to the differential equation (6.1). In the following, we assume that $t_0 = 0$. We denote the solution of the Lie-Trotter splitting by C_{LT}. It is defined on the mesh $\{t_i := i\Delta t,\ i \geq 0\}$. The Lie-Trotter splitting is initialized by

$$C_{LT}(t_0) = C_0\,.$$

For $i = 0, 1, \ldots$ the scheme is given by

$$\begin{aligned} \frac{\mathrm{d}C^*}{\mathrm{d}t} &= G(C^*), \qquad C^*(0) = C_{LT}(t_i), \qquad \text{on } [0, \Delta t], \\ \frac{\mathrm{d}C^{**}}{\mathrm{d}t} &= F(C^{**}), \qquad C^{**}(0) = C^*(\Delta t), \qquad \text{on } [0, \Delta t], \\ C_{LT}(t_{i+1}) &= C^{**}(\Delta t)\,. \end{aligned} \tag{6.2}$$

We write $C_{LT,\Delta t}$ in order to emphasise the splitting step size Δt. In general it is also possible to change the order of the substeps. This does not influence the convergence order in the non-stiff case. However, Sportisse [121] showed that the order of the substeps is important for the stiff case. Hence, we only regard the scheme (6.2), which integrates the (non-stiff) transport term $G(C)$ before the (stiff) chemistry term $F(C)$.

The scheme of the Strang splitting is given by

$$C_S(t_0) = C_0\,.$$

For $i = 0, 1, \ldots$

$$\begin{aligned} \frac{\mathrm{d}C^*}{\mathrm{d}t} &= F(C^*), \qquad C^*(0) = C_S(t_i), \qquad \text{on } [0, \frac{\Delta t}{2}], \\ \frac{\mathrm{d}C^{**}}{\mathrm{d}t} &= G(C^{**}), \qquad C^{**}(0) = C^*(\frac{\Delta t}{2}), \qquad \text{on } [0, \Delta t], \\ \frac{\mathrm{d}C^{***}}{\mathrm{d}t} &= F(C^{***}), \qquad C^{***}(\frac{\Delta t}{2}) = C^{**}(\Delta t), \qquad \text{on } [\frac{\Delta t}{2}, \Delta t], \\ C_S(t_{i+1}) &= C^{***}(\Delta t)\,. \end{aligned} \tag{6.3}$$

Following [121], the integration routine is ended with the integration of the stiff part of the ODE. This reduces the splitting error in the stiff case.

Finally, we introduce the Richardson extrapolation of the Lie-Trotter splitting. We use the Lie-Trotter splitting (6.2) with two different step sizes Δt and $\Delta t/2$. Thereby both schemes are restarted at $t_i := t_0 + i\Delta t$ for all i, and the corresponding initial value at t_i is the extrapolated value $C_{RE}(t_i)$. The

order of the Lie-Trotter splitting is one. Therefore, the Richardson extrapolation C_{RE} of the Lie-Trotter splitting is given by

$$C_{RE}(t_{i+1}) = 2 \cdot C_{LT,\Delta t/2}(t_i + \Delta t) - C_{LT,\Delta t}(t_i + \Delta t). \tag{6.4}$$

The derivation of the splitting scheme can be looked up in [47]. The coefficients of the extrapolated Lie-Trotter splitting are 2 and -1. Thus, stability of the scheme is non-trivial. In Section 6.5, stability of the extrapolated Lie-Trotter splitting is examined. Successive applications of Richardson extrapolation can generate splitting schemes with any required order. However, the stability of the extrapolated schemes is not guaranteed. We only consider the scheme (6.4). Results for other extrapolated schemes can be obtained in a similar way. Extrapolation methods (without splitting) for stiff ODEs are considered in [35, 64].

6.3 Convergence order for (non-stiff) problems

In this section, we review the standard asymptotic analysis of splitting methods for non-stiff differential equations (e.g., [121]). For simplicity we consider a linear differential equation

$$C' = A \cdot C + B \cdot C, \qquad C(0) = C_0. \tag{6.5}$$

The solution $C(\Delta t)$ is

$$C(\Delta t) = e^{(A+B)\cdot \Delta t} \cdot C_0.$$

We compare the solution $C(\Delta t)$ to the numerical approximation of the splitting methods (6.2), (6.3), and (6.4). The approximation after one step of the Lie-Trotter splitting is

$$C_{LT}(\Delta t) = e^{A\cdot \Delta t} \cdot e^{B\cdot \Delta t} \cdot C_0.$$

Furthermore, the approximation after one step of the Strang splitting is

$$C_S(\Delta t) = e^{A\cdot \frac{\Delta t}{2}} \cdot e^{B\cdot \Delta t} \cdot e^{A\cdot \frac{\Delta t}{2}} \cdot C_0.$$

Finally, the approximation of the extrapolated scheme [47] is

$$C_{RE}(\Delta t) = \left(2 \cdot e^{A\cdot \Delta t/2} \cdot e^{B\cdot \Delta t/2} \cdot e^{A\cdot \Delta t/2} \cdot e^{B\cdot \Delta t/2} - e^{A\cdot \Delta t} \cdot e^{B\cdot \Delta t}\right) \cdot C_0.$$

In order to obtain the convergence order of all splitting methods, we examine the local error of the splitting schemes. The local errors are

$$e_{LT}(\Delta t) = C_{LT}(\Delta t) - C(\Delta t), \quad e_S(\Delta t) = C_S(\Delta t) - C(\Delta t), \quad e_{RE}(\Delta t) = C_{RE}(\Delta t) - C(\Delta t).$$

Taylor expansion for Δt close to 0 results in

$$e_{LT}(\Delta t) = \frac{A \cdot B - B \cdot A}{2} \cdot \Delta t^2 \cdot C_0 + \mathcal{O}(\Delta t^3) = \mathcal{O}(\Delta t^2). \tag{6.6}$$

Similar to that, it holds

$$e_S(\Delta t) = \frac{2ABA + 2AB^2 + 2B^2A - 4BAB - BA^2 - A^2B}{24} \cdot \Delta t^3 \cdot C_0 + \mathcal{O}(\Delta t^4) = \mathcal{O}(\Delta t^3), \tag{6.7}$$

$$e_{RE}(\Delta t) = \frac{2ABA + 2BAB - A^2B - AB^2 - B^2A - BA^2}{24} \cdot \Delta t^3 \cdot C_0 + \mathcal{O}(\Delta t^4) = \mathcal{O}(\Delta t^3). \tag{6.8}$$

Equations (6.6), (6.7), and (6.8) are derived in den Appendix **A.2**. The error constants of all considered schemes are zero for commuting matrices A and B. In the following we assume that the operators do not commute. Therefore, in the linear case the Lie-Trotter splitting is a first-order scheme, whereas the Strang splitting as well as the extrapolated Lie-Trotter splitting are second-order schemes.

The coefficients in the Taylor expansion depend on the matrices A and B. If the linear differential equation (6.5) describes a chemical reaction system with transport, parts of the matrix A are scaled with $1/\epsilon$. Thereby ϵ is the fastest timescale of the chemistry-only equation. An implicit integration method has no advantages over an explicit integration method if the step size Δt is smaller than ϵ. Hence, for smooth solutions, realistic step sizes are $\Delta t \gg \epsilon$. Thus, for stiff differential equations we do not consider the asymptotic case $\Delta t \to 0$, and the results of this Section do not hold [121, 122, 141]. More information about order reduction can be found in [64, 106].

Note that splitting methods also suffer from order reduction if the boundary conditions are non-trivial [42, 43, 44]. In this chapter we only consider order reduction that is caused by stiffness of the chemical reaction system. Order reduction due to non-trivial boundary conditions is not examined in this PhD thesis. However, correction strategies for non-trivial boundary conditions are elaborated in [42, 43, 44].

The different splitting schemes have the same order in case of a (non-stiff) nonlinear differential equation

$$C' = F(C) + G(C), \qquad C(0) = C_0. \tag{6.9}$$

The Lie operator formalism [40] is used for the proof. For the Strang splitting a detailed description is given in [79].

In this section, we show that the convergence rate of the extrapolated splitting scheme in the stiff case is two. The definition of the 'order' of an integration method refers to $\Delta t \to 0$. However, for stiff differential equations and reasonable time steps Δt a stiffness parameter ϵ with $\epsilon \ll \Delta t$ exists. Therefore, the convergence rate in the stiff case does not match the classical order (see Remark 6.4). Lie-Trotter splitting and Strang splitting only provide convergence of order one [76] for stiff ODEs. The considered differential equations (6.1) are chemical reaction systems with transport, which are discretized in space. We assume that the transport term $G(C)$ is not stiff. Hence, all the stiffness is related to the chemical source term $F(C)$. Furthermore, for the analysis in this section we assume that the examined differential equation is given by the standard form of a singular perturbation problem (compare to equation (3.14))

$$\begin{pmatrix} z^1 \\ \epsilon \cdot z^2 \end{pmatrix}' = \begin{pmatrix} f_0(z^1, z^2) \\ f_1(z^1, z^2) \end{pmatrix} + \begin{pmatrix} g_0(z^1, z^2) \\ \epsilon \cdot g_1(z^1, z^2) \end{pmatrix}, \qquad \begin{pmatrix} z^1(t_0) \\ z^2(t_0) \end{pmatrix} \text{ is given,} \qquad 0 < \epsilon \ll 1\,. \tag{6.10}$$

In (6.10) the functions f_0, f_1, g_0, g_1, and their derivatives are $\mathcal{O}(1)$ for $\epsilon \to 0$, and the parameter $0 < \epsilon \ll 1$ describes the stiffness of the differential equation. Furthermore, we assume that a logarithmic norm of f_{1,z^2} is smaller than -1. Thus, f_{1,z^2} is nonsingular and $f_1(z^1, z^2) = 0$ can be solved with respect to z^2. Assume that $C = (z^1, z^2)^T$. Moreover, it holds $C_{LT} = (z^1_{LT}, z^2_{LT})^T$ and $C_{RE} = (z^1_{RE}, z^2_{RE})^T$. In general the differential equation (6.1) does not occur in the form (6.10). However, if $g(C)$ is not stiff and all timescales of $f(C)$ are in $\mathcal{O}(1)$ or in $\mathcal{O}(\epsilon)$, the differential equation can be converted to the form (6.10) (see [18, 19, 48] with an additional transport term). Note that the transformation of the problem is not necessary for the application of the extrapolated Lie-Trotter splitting. Hence, the given assumption does not restrict the applicability of the obtained results. Previous work about the stiff case covers extrapolation of integration methods of ODEs [7, 35, 63, 64] and the convergence order of Lie-Trotter splitting as well as Strang splitting [32, 33, 76, 121, 122, 141]. However, extrapolated splitting schemes have not been examined in the stiff case. For further analysis we hypothesize that the stiffness parameter ϵ is much smaller than the time step Δt. Under this assumption Kozlov et al. [76] analyzed the local error of the Lie-Trotter splitting (6.2). The following estimate of the splitting error for the extrapolated Lie-Trotter splitting follows the proof given in [76] based on singular perturbation techniques.

Theorem 6.1. *We assume that* $0 < \epsilon \ll 1$ *and* $\epsilon \ll \Delta t$. *Furthermore, we assume that the initial values of* (6.10) *can be expressed as a power series in* ϵ *(compare to equation* (6.18)*). Then, the extrapolated splitting scheme* (6.4) *applied to the differential equation* (6.10) *has the local error*

$$z^1(t_0 + \Delta t) - z^1_{RE}(t_0 + \Delta t) = \mathcal{O}(\Delta t^3 + \epsilon \Delta t + \epsilon^2) \tag{6.11}$$

$$z^2(t_0 + \Delta t) - z^2_{RE}(t_0 + \Delta t) = \mathcal{O}(\Delta t^3 + \epsilon)\,. \tag{6.12}$$

Therefore, it is a second-order scheme in time for sufficiently small ϵ.

Proof. In [76], the local error of the Lie-Trotter splitting is examined for system (6.10) with given initial values $(z_*^1, z_*^2)^T$. Thereby the step size is denoted by h. Under the assumption $\epsilon \ll h$, the transients are damped fastly, and only the smooth parts $z_s^1(t)$, $z_s^2(t)$ of the solution remain.

The power series in ϵ of the smooth parts of the solution are

$$\begin{aligned} z_s^1(t) &= z_0^1(t) + \epsilon \cdot z_1^1(t) + \epsilon^2 \cdot z_2^1(t) + \dots, \\ z_s^2(t) &= z_0^2(t) + \epsilon \cdot z_1^2(t) + \epsilon^2 \cdot z_2^2(t) + \dots. \end{aligned} \tag{6.13}$$

If not stated differently, functions are evaluated at (z_0^1, z_0^2) in the following. Inserting the power series (6.13) into equation (6.10) and sorting according to powers of ϵ results in differential-algebraic equations of increasing index:

$$\left.\begin{aligned} &\left.\begin{aligned} z_0^{1\prime} &= f_0 + g_0 \\ 0 &= f_1 \end{aligned}\right\} \text{index 1} \\ &\left.\begin{aligned} z_1^{1\prime} &= (f_0 + g_0)_{z^1} \cdot z_1^1 + (f_0 + g_0)_{z^2} \cdot z_1^2 \\ z_0^{2\prime} &= g_1 + f_{1,z^1} \cdot z_1^1 + f_{1,z^2} \cdot z_1^2 \end{aligned}\right\} \text{index 2} \\ &\quad \vdots \end{aligned}\right. \tag{6.14}$$

Constraints for the z_i^2, $i \geq 0$, follow immediately from (6.14)

$$f_1 = 0, \tag{6.15}$$

$$\frac{\mathrm{d}f_1}{\mathrm{d}t} = f_{1,z^1} \cdot (f_0 + g_0) + f_{1,z^2} \cdot \left(g_1 + f_{1,z^1} \cdot z_1^1 + f_{1,z^2} \cdot z_1^2\right) = 0, \tag{6.16}$$

$$\vdots$$

According to [76], the function $z_0^1(t)$ fulfils

$$\begin{aligned} z_0^1(t+h) = z_0^1(t) + h\,(f_0 + g_0) + \\ \frac{h^2}{2}\Big((f_0 + g_0)_{z^1} \cdot (f_0 + g_0) + (f_0 + g_0)_{z^2} \cdot (-f_{1,z^2})^{-1} \cdot f_{1,z^1} \cdot (f_0 + g_0)\Big) + \mathcal{O}(h^3). \end{aligned} \tag{6.17}$$

Thereby

$$\begin{aligned} 0 = f_1(z_0^1, z_0^2) \\ \implies 0 = f_{1,z^1} z_0^{1\prime} + f_{1,z^2} z_0^{2\prime} \\ \implies z_0^{2\prime} = \left(-f_{1,z^2}\right)^{-1} f_{1,z^1} z_0^{1\prime} \end{aligned}$$

is used. In the following we will use the notation

$$H_1(h) = \frac{h^2}{2}\left((f_0+g_0)_{z^1}\cdot(f_0+g_0)+(f_0+g_0)_{z^2}\cdot(-f_{1,z^2})^{-1}\cdot f_{1,z^1}\cdot(f_0+g_0)\right).$$

According to the assumptions, the initial values of (6.10) can be expressed as a power series in ϵ:

$$\begin{aligned} z_*^1 &= z_*^{1,0} + \epsilon z^{1,1}{}_* + \epsilon^2 z^{1,2}{}_* + \ldots, \\ z_*^2 &= z_*^{2,0} + \Delta z^{2,0} + \epsilon\left(z_*^{2,1} + \Delta z^{2,1}\right) + \epsilon^2\left(z_*^{2,2} + \Delta z^{2,2}\right)\ldots. \end{aligned} \tag{6.18}$$

In (6.18) $z_*^{2,j}$, $j \geq 0$, denote the variables satisfying the algebraic constraints

$$f_1 = 0, \tag{6.19}$$

$$f_{1,z^1}\cdot f_0 + f_{1,z^2}\cdot\left(f_{1,z^1}\cdot z_1^1 + f_{1,z^2}\cdot z_1^2\right) = 0 \tag{6.20}$$

$$\vdots$$

of the chemistry-only equation. These constraints are derived similar to (6.15) and (6.16). Hence, we have $\Delta z^{2,j} = 0$ for all $j \geq 0$, after a chemistry-only step. With these settings, [76] derives the following error bound for (6.10):

$$\begin{aligned} z^1(t_0+h) - z_{LT}^1(t_0+h) &= -h g_{0,z^2}\Delta z^{2,0} + \mathcal{O}(h^3+\epsilon h+\epsilon^2) \\ &\quad + \frac{h^2}{2}\left(g_{0,z^1}f_0 - f_{0,z^1}g_0 + g_{0,z^2}\left(-f_{1,z^2}\right)^{-1} f_{1,z^1}(g_0+f_0) - f_{0,z^2}\left(-f_{1,z^2}\right)^{-1} f_{1,z^1}g_0 - g_{0,z^2}g_1\right) \\ z^2(t_0+h) - z_{LT}^2(t_0+h) &= \epsilon\left(-f_{1,z^2}\right)^{-1}\left(g_1 - \left(-f_{1,z^2}\right)^{-1} f_{1,z^1}g_0\right) + \mathcal{O}(h^2+\epsilon h+\epsilon^2). \end{aligned} \tag{6.21}$$

In the following we will use the notation

$$H_2(h) = \frac{h^2}{2}\left(g_{0,z^1}f_0 - f_{0,z^1}g_0 + g_{0,z^2}\left(-f_{1,z^2}\right)^{-1} f_{1,z^1}(g_0+f_0) - f_{0,z^2}\left(-f_{1,z^2}\right)^{-1} f_{1,z^1}g_0 - g_{0,z^2}g_1\right).$$

With this preliminary work we consider the extrapolated Lie-Trotter splitting. According to equation (6.4), we have to compute the error for $C_{LT,\Delta t/2}(t_0+\Delta t)$ as well as the error for $C_{LT,\Delta t}(t_0+\Delta t)$. The error for $C_{LT,\Delta t}(t_0+\Delta t)$ follows directly from (6.21) with $h = \Delta t$.

The solution $C_{LT,\Delta t/2}(t_0+\Delta t)$ is computed with two substeps. For the first substep with length $\Delta t/2$ we can use (6.21). Afterwards another step with step size $\Delta t/2$ has to be performed. The initial value of the second substep is not the value of the solution $C(t_0+\Delta t/2)$ of (6.10), but the result of the first substep. Therefore, the occurring error assembles from two different sources. Equation (6.17) and Taylor expansion result in the defect due to the difference in the initial values of the second substep at time $t_0+\Delta t/2$. Besides, (6.21) gives the additional splitting error for the second substep. For the second substep we have $\Delta z^{2,j} = 0$ for all j because the previous step ends with the chemistry-only part.

The trajectory of (6.10) with initial values $(z_*^1, z_*^2)^T$ fulfils

$$z_0^1(t_0+\Delta t) = z_0^1(t_0+\Delta t/2) + \frac{\Delta t}{2}(f_0+g_0) + H_1(\Delta t/2) + \mathcal{O}(\Delta t^3). \tag{6.22}$$

Note that all functions in (6.22) are evaluated in $z_0^1(t_0+\Delta t/2), z_0^2(t_0+\Delta t/2)$.
The initial values $z_{LT}^1(t_0+\Delta t/2)$, $z_{LT}^2(t_0+\Delta t/2)$ of the second substep are compared with the true solution of the differential equation at $t_0+\Delta t/2$:

$$\begin{aligned}
z_{LT}^1(t_0+\Delta t/2) &= z^1(t_0+\Delta t/2) + \frac{\Delta t}{2} g_{0,z^2}\Delta z^{2,0} - H_2(\Delta t/2) + \mathcal{O}(\Delta t^3 + \epsilon\Delta t + \epsilon^2),\\
z_{LT}^2(t_0+\Delta t/2) &= z^2(t_0+\Delta t/2) - \epsilon\left(-f_{1,z^2}\right)^{-1}\left(g_1 - \left(-f_{1,z^2}\right)^{-1} f_{1,z^1} g_0\right) + \mathcal{O}(\Delta t^2 + \epsilon\Delta t + \epsilon^2).
\end{aligned}$$

Thereby all functions are evaluated in $z_0^1(t_0)$ and $z_0^2(t_0)$. For the smooth solution $\tilde{z}^1$ of the differential equation (6.10) with initial values $z_{LT}^1(t_0+\Delta t/2)$, $z_{LT}^2(t_0+\Delta t/2)$ a similar power series expansion as (6.13) is available. The first term is

$$\tilde{z}_0^1(t_0+\Delta t/2) = z_0^1(t_0+\Delta t/2) + \frac{\Delta t}{2} g_{0,z^2}\Delta z^{2,0} - H_2(\Delta t/2) + \mathcal{O}(\Delta t^3 + \epsilon\Delta t + \epsilon^2).$$

Again all functions are evaluated in $z_0^1(t_0)$ and $z_0^2(t_0)$. If $\tilde{z}^1(t+\Delta t/2)$ is inserted into equation (6.17), the occurring functions $\frac{\Delta t}{2}(f_0+g_0)$ and $H_1(\Delta t/2)$ are evaluated in $\tilde{z}_0^1(t_0+\Delta t/2)$. Hence, for initial values $(z_{LT}^1(t_0+\Delta t/2), z_{LT}^2(t_0+\Delta t/2))$ at time $t_0+\Delta t/2$ the solution $\tilde{z}^1(t_0+\Delta t)$ of (6.10) fulfils

$$\tilde{z}_0^1(t_0+\Delta t) = z_0^1(t_0+\Delta t/2) + \frac{\Delta t}{2} g_{0,z^2}\Delta z^{2,0} - H_2(\Delta t/2) + \frac{\Delta t^2}{4}(f_0+g_0)_{z^1} g_{0,z^2}\Delta z^{2,0} \tag{6.23}$$

$$+\frac{\Delta t}{2}(f_0+g_0) + H_1(\Delta t/2) + \mathcal{O}(\Delta t^3 + \epsilon\Delta t + \epsilon^2). \tag{6.24}$$

Taylor expansion is used in order to evaluate line (6.23) in $(z_0^1(t_0), z_0^2(t_0))$ and line (6.24) in $(z_0^1(t_0+\Delta t/2), z_0^2(t_0+\Delta t/2))$. Similar to that, we obtain

$$\tilde{z}_1^1(t_0+\Delta t) = z_1^1(t_0+\Delta t) + \mathcal{O}(\Delta t). \tag{6.25}$$

Using (6.22) and (6.23) we obtain the difference that is caused by the error in the initial values at time $t_0+\Delta t/2$ of the second substep, which is

$$\begin{aligned}
z^1(t_0+\Delta t) - \tilde{z}^1(t_0+\Delta t) &= z^1(t_0+\Delta t) - z^1(t_0+\Delta t/2) - \left(\tilde{z}^1(t_0+\Delta t) - z^1(t_0+\Delta t/2)\right)\\
&= -\frac{\Delta t}{2} g_{0,z^2}\Delta z^{2,0} + H_2(\Delta t/2) - \frac{\Delta t^2}{4}(f_0+g_0)_{z^1} g_{0,z^2}\Delta z^{2,0} + \mathcal{O}(\Delta t^3 + \epsilon\Delta t + \epsilon^2).
\end{aligned} \tag{6.26}$$

In equation (6.26) all functions are evaluated in $z_0^1(t_0)$ and $z_0^2(t_0)$.
In addition a splitting error is generated. If we use the Lie-Trotter splitting with initial values $(z_{LT}^1(t_0 + \Delta t/2), z_{LT}^2(t_0 + \Delta t/2))$ at time $\Delta t/2$, we obtain

$$z_{LT,\Delta t/2}^1(t_0 + \Delta t) = \tilde{z}^1(t_0 + \Delta t) - H_2(\Delta t/2) + \mathcal{O}(\Delta t^3 + \epsilon \Delta t + \epsilon^2). \tag{6.27}$$

Equation (6.27) follows from (6.21) with $\Delta z^{2,0} = 0$. Moreover, Taylor expansions are used in order to evaluate all functions at $z_0^1(t_0), z_0^2(t_0)$. Hence, the splitting error for two substeps of the Lie-Trotter splitting is given by

$$\begin{aligned} z^1(t_0 + \Delta t) - z_{LT,\Delta t/2}^1(t_0 + \Delta t) &= z^1(t_0 + \Delta t) - \tilde{z}^1(t_0 + \Delta t) + \tilde{z}^1(t_0 + \Delta t) - z_{LT,\Delta t/2}^1(t_0 + \Delta t) \\ &= -\frac{\Delta t}{2} g_{0,z^2} \Delta z^{2,0} + 2H_2(\Delta t/2) - \frac{\Delta t^2}{4} (f_0 + g_0)_{z^1} g_{0,z^2} \Delta z^{2,0} + \mathcal{O}(\Delta t^3 + \epsilon \Delta t + \epsilon^2). \end{aligned} \tag{6.28}$$

It follows from (6.4), (6.21) and (6.28) that

$$\begin{aligned} z^1(t_0 + \Delta t) - z_{RE}^1(t_0 + \Delta t) &= z^1(t_0 + \Delta t) - \left(2z_{LT,\Delta t/2}^1(t_0 + \Delta t) - z_{LT,\Delta t}^1(t_0 + \Delta t)\right) \\ &= 2\left(z^1(t_0 + \Delta t) - z_{LT,\Delta t/2}^1(t_0 + \Delta t)\right) - \left(z^1(t_0 + \Delta t) - z_{LT,\Delta t}^1(t_0 + \Delta t)\right) \\ &= -\frac{\Delta t^2}{2} (f_0 + g_0)_{z^1} g_{0,z^2} \Delta z^{2,0} + \mathcal{O}(\Delta t^3 + \epsilon \Delta t + \epsilon^2). \end{aligned} \tag{6.29}$$

The error bound for z_{RE}^2 is derived analogously to the error bound for z_{RE}^1. Similar to equation (6.17), we derive

$$z_0^2(t + h) = z_0^2(t) + h\left[(-f_{1,z^2}^{-1} f_{1,z^1}(f_0 + g_0)\right] + H_3(h) + \mathcal{O}(h^3). \tag{6.30}$$

Note that the function $H_3(h) = \mathcal{O}(h^2)$ is comparable to $H_1(h)$. Hence, it holds

$$z_0^2(t_0 + \Delta t) = z_0^2(t_0 + \Delta t/2) + \Delta t/2\left[(-f_{1,z^2}^{-1} f_{1,z^1}(f_0 + g_0)\right] + H_3(\Delta t/2) + \mathcal{O}(\Delta t^3). \tag{6.31}$$

In this equation all functions are evaluated in $(z_0^1(t_0 + \Delta t/2), z_0^2(t_0 + \Delta t/2))$. The initial value of the second substep is the result of the Lie-Trotter splitting with step size $\Delta t/2$. The corresponding inital value of the second substep is

$$\begin{aligned} z_{LT}^1(t_0 + \frac{\Delta t}{2}) &= z^1(t_0 + \frac{\Delta t}{2}) + \frac{\Delta t}{2} g_{0,z^2} \Delta z^{2,0} - H_2(\frac{\Delta t}{2}) + \mathcal{O}(\Delta t^3 + \epsilon \Delta t + \epsilon^2), \\ z_{LT}^2(t_0 + \frac{\Delta t}{2}) &= z^2(t_0 + \frac{\Delta t}{2}) - \epsilon \left(-f_{1,z^2}\right)^{-1} \left(g_1 - \left(-f_{1,z^2}\right)^{-1} f_{1,z^1} g_0\right) - H_4(\frac{\Delta t}{2}) + \mathcal{O}(\Delta t^3 + \epsilon \Delta t + \epsilon^2). \end{aligned}$$

The function $H_4(h) = \mathcal{O}(h^2)$ is comparable to $H_2(h)$. For the initial value $(z^1_{LT}(t_0+\Delta t/2), z^2_{LT}(t_0+\Delta t/2))$ at time $t_0 + \Delta t/2$, the solution $\tilde{z}^2(t_0 + \Delta t)$ of (6.10) fulfils

$$\tilde{z}^2_0(t_0+\Delta t) = z^2_0(t_0 + \Delta t/2) - H_4(\frac{\Delta t}{2}) \tag{6.32}$$

$$+ \Delta t/2\big[(-f^{-1}_{1,z^2} f_{1,z^1}(f_0 + g_0)\big] + H_3(\Delta t/2) + \mathcal{O}(\Delta t^3 + \epsilon)\,. \tag{6.33}$$

Taylor expansion is used in order to evaluate line (6.33) in $(z^1_0(t_0 + \Delta t/2), z^2_0(t_0 + \Delta t/2))$. From (6.31), (6.32), and (6.33) we obtain

$$\begin{aligned} z^2(t_0 + \Delta t) - \tilde{z}^2(t_0 + \Delta t) &= z^2(t_0 + \Delta t) - z^2(t_0 + \Delta t/2) - \big(\tilde{z}^2(t_0 + \Delta t) - z^2(t_0 + \Delta t/2)\big) \\ &= H_4(\frac{\Delta t}{2}) + \mathcal{O}(\Delta t^3 + \epsilon)\,. \end{aligned} \tag{6.34}$$

However, in the second substep of the Lie-Trotter splitting, an additional splitting error is introduced. The splitting error of the second substep is

$$z^2_{LT,\Delta t/2}(t_0 + \Delta t) - \tilde{z}^2(t_0 + \Delta t) = -H_4(\frac{\Delta t}{2}) + \mathcal{O}(\Delta t^3 + \epsilon)\,. \tag{6.35}$$

Equation (6.35) follows directly from equation (6.21). Thus, we obtain

$$z^2(t_0 + \Delta t) - \tilde{z}^2(t_0 + \Delta t) + \tilde{z}^2(t_0 + \Delta t) - z^2_{LT,\Delta t/2}(t_0 + \Delta t) = 2H_4(\Delta t/2) + \mathcal{O}(\Delta t^3 + \epsilon)\,. \tag{6.36}$$

Thus, the solution z^2_{RE} of the extrapolated Lie-Trotter splitting fulfils

$$\begin{aligned} z^2(t_0 + \Delta t) - z^2_{RE}(t_0 + \Delta t) &= z^2(t_0 + \Delta t) - \big(2z^2_{LT,\Delta t/2}(t_0 + \Delta t) - z^2_{LT,\Delta t}(t_0 + \Delta t)\big) \\ &= 2\big(z^2(t_0 + \Delta t) - z^2_{LT,\Delta t/2}(t_0 + \Delta t)\big) - \big(z^2(t_0 + \Delta t) - z^2_{LT,\Delta t}(t_0 + \Delta t)\big) \\ &= 4H_4(\Delta t/2) - H_4(\Delta t) + \mathcal{O}(\Delta t^3 + \epsilon) = \mathcal{O}(\Delta t^3 + \epsilon)\,. \end{aligned} \tag{6.37}$$

The exact solution of (6.10) fulfils equation (6.19) (due to the initial values the equation might be violated during the fast transient phase). Hence, the error bound (6.37) implies $\Delta z^{2,0} = \mathcal{O}(\Delta t^3 + \epsilon)$ (after a time step Δt). Thus, we obtain

$$z^1(t_0 + \Delta t) - z^1_{RE}(t_0 + \Delta t) = \mathcal{O}(\Delta t^3 + \epsilon\Delta t + \epsilon^2)\,. \tag{6.38}$$

□

Theorem 6.2. *Under the assumptions of Theorem 6.1 the extrapolated splitting scheme* (6.4) *applied to the differential equation* (6.10) *has the global error*

$$z^1(t) - z^1_{RE}(t) = \mathcal{O}(\Delta t^2 + \epsilon) \tag{6.39}$$
$$z^2(t) - z^2_{RE}(t) = \mathcal{O}(\Delta t^2 + \epsilon)\,. \tag{6.40}$$

Proof. The smooth part of the solution of (6.10) fulfils equations (6.15) and (6.16). We assumed that f_{1,z^2} is nonsingular. Hence, z_0^2 is derived by equation (6.15) and the implicit function theorem. Furthermore, (6.16) results in

$$z_1^2 = -\left(f_{1,z^2}^{-1}\right)^2 \cdot f_{1,z^1} \cdot (f_0 + g_0) - f_{1,z^2}^{-1} \cdot \left(g_1 + f_{1,z^1} \cdot z_1^1\right). \tag{6.41}$$

Thus, z^2 is determined by z^1 up to $\mathcal{O}(\epsilon^2)$. The Lie-Trotter splitting terminates with a chemistry-only step. Therefore, equations (6.19) and (6.20) hold for the numerical solution of the Lie-Trotter splitting. Equation (6.19) is equal to equation (6.15). However, equation (6.20) differs from equation (6.16). Instead of (6.41), the solution of the Lie-Trotter splitting fulfils

$$z^2_{LT,1} = -\left(f_{1,z^2}^{-1}\right)^2 \cdot f_{1,z^1} \cdot f_0 - f_{1,z^2}^{-1} \cdot f_{1,z^1} \cdot z^1_{LT,1}\,. \tag{6.42}$$

Thereby the functions f_0, f_1, g_0, g_1, and their derivatives are $\mathcal{O}(1)$. As a result, equation (6.42) corresponds to equation (6.41) with an error $\mathcal{O}(1)$. Hence, the error in $z^2_{LT,1}$ has the same order as the error in $z^1_{LT,1}$ with an additional part $\mathcal{O}(1)$. Equation (6.42) is a linear equation, and Richardson extrapolation does not cancel out this error. Thus, we obtain that the error in $z^2_{RE,1}$ has the same order as the error in $z^1_{RE,1}$ with an additional part $\mathcal{O}(1)$.
First, we consider the global error in z^1. For this purpose we use that z_0^2 is given by a function of z_0^1, and z_1^2 is given by a function of z_1^1. Higher orders of ϵ are ignored. Therefore, it holds

$$z^2 = \tilde{z}^2(z^1) + \mathcal{O}(\epsilon^2)\,.$$

If the function f is sufficiently smooth, we obtain the differential equation

$$\left(z^1\right)' = f_0\left(z^1, \tilde{z}^2(z^1)\right) + g_0\left(z^1, \tilde{z}^2(z^1)\right) + \mathcal{O}(\epsilon^2)\,. \tag{6.43}$$

Due to the difference between equations (6.41) and (6.42), the extrapolated Lie-Trotter splitting results in an approximation of the solution of the defective differential equation

$$\left(\bar{z}^1\right)' = f_0\left(\bar{z}^1, \tilde{z}^2(\bar{z}^1)\right) + g_0\left(\bar{z}^1, \tilde{z}^2(\bar{z}^1)\right) + \mathcal{O}(\epsilon)\,. \tag{6.44}$$

The global error in z^1_{RE} can be computed by the sum of the propagated local errors (6.11) plus the approximation error $\mathcal{O}(\epsilon)$ of equation (6.44). It follows (6.39):

$$z^1(t) - z^1_{RE}(t) = \mathcal{O}(\Delta t^2 + \epsilon).$$

Thus, the global error in $z^1_{RE,1}$ is $\mathcal{O}(1)$. Then, the global error of $z^2_{RE,1}$ is also $\mathcal{O}(1)$, and the global error (6.40) of z^2_{RE} fulfils

$$z^2(t) - z^2_{RE}(t) = \mathcal{O}(\Delta t^2 + \epsilon).$$

□

Remark 6.3. *Under the assumptions of Theorem 3.3 the stiff differential equation* (6.10) *is approximated by the differential-algebraic equation (DAE)*

$$\begin{pmatrix} z^1 \\ 0 \end{pmatrix}' = \begin{pmatrix} f_0(z^1, z^2) \\ f_1(z^1, z^2) \end{pmatrix} + \begin{pmatrix} g_0(z^1, z^2) \\ 0 \end{pmatrix}, \qquad \begin{pmatrix} z^1(0) \\ z^2(0) \end{pmatrix} = \begin{pmatrix} z^1_0 \\ f_1^{-1}(z^1_0) \end{pmatrix}.$$

The approximation error is in $\mathcal{O}(\epsilon)$. Hence, if (6.10) *is approximated by a DAE, the occurring approximation error is comparable with the Δt-independent part of the splitting error.*

Remark 6.4. *The definition of the 'order' of an integration method refers to $\Delta t \to 0$. However, the previous proof is only valid in the case $\epsilon \ll \Delta t$. Therefore, the term 'order' is mathematically incorrect. However, in the case $\Delta t \leq \epsilon$ explicit methods can be used. Thus, we are interested in the setting $\epsilon \ll \Delta t$.*

Remark 6.5. *If $f(C)$ as well as $g(C)$ are stiff in equation* (6.1)*, order reduction occurs for the extrapolated splitting method [47].*

Remark 6.6. *It is also possible to extrapolate the Strang splitting in order to obtain a higher-order scheme for stiff differential equations. However, Strang splitting suffers from an order reduction. Hence, if we use the Richardson extrapolation for a first order scheme, we obtain an extrapolated Strang method with order two for all time steps Δt. If we use the Richardson extrapolation for a second order scheme, we obtain order one for the stiff case $\Delta t > \epsilon$ and order four for the non-stiff case $\Delta t < \epsilon$.*

6.5 Stability of the extrapolated Lie-Trotter splitting

In this section, we will discuss the stability of the extrapolated splitting method (6.4). First we have to define an appropriate test equation. Usually, the test equation

$$y' = \lambda y, \qquad \lambda \in \mathbb{C}, \tag{6.45}$$

is used to examine the stiff stability of numerical methods for ODEs [64, 143]. The numerical scheme is applied to the test equation (6.45), and for $z := \Delta t\lambda$ a recursion

$$\begin{aligned} y_0 &= y(t_0), \\ y_i &= R(\Delta t\lambda)y_{i-1} = R(z)y_{i-1} \approx y(t_0 + i\Delta t) \end{aligned}$$

is obtained. The considered numerical method is stable if the stability function fulfils $|R(z)| \leq 1$ for all z in the left half-plane $\{z \in \mathbb{C} \ : \ \Re(z) \leq 0\}$. Obviously, the test equation (6.45) is not suitable for splitting methods because the splitting of the source term is not considered. In the following we list some approaches, which are used in order to examine stability for splitting methods.

In [144] the authors consider extrapolated splitting schemes combined with the θ-method. Let the stability function of the θ-method be denoted by $R_\theta(\Delta t\lambda)$. In [144] the considered stability function of the extrapolated splitting scheme combined with the θ-method is

$$\tilde{R}_\theta(\Delta t\lambda) = 2R_\theta^4(0.5\Delta t\lambda) - R_\theta^2(\Delta t\lambda).$$

This stability function corresponds to the extrapolated Lie-Trotter splitting applied to the test equation

$$y' = \lambda y + \lambda y\,. \tag{6.46}$$

However, linear operators commute for one-dimensional equations. Thus, no splitting error occurs, and the test equation (6.46) is not suitable for stability analysis of method (6.4).

In [113] the stability of different operator splitting methods is investigated. The used test equation is a multidimensional linear ODE, and stability criteria are deduced from submultiplicativity of the matrix norm. In case of extrapolated splitting schemes the exponentials are not only multiplied but also summed up. Thus, submultiplicativity combined with stability of all operators is not sufficient in order to verify stability of the extrapolated splitting scheme.

Remark 6.7. *According to [125], for any given vector norm $\|\cdot\|_*$ in $\mathbb{R}^n$ the logarithmic matrix norm $\mu_*[\cdot]$ is defined by*

$$\mu_*[A] = \lim_{h \to +0} \frac{\|I + hA\|_* - 1}{h}\,.$$

The matrix norms and the logarithmic matrix norms related to the 1-norm and the ∞-norm are given by

$$\begin{aligned} \|A\|_1 &= \max_{1\leq j\leq n} \sum_{i=1}^{n} |a_{ij}|, \qquad \mu_1[A] = \max_{1\leq j\leq n} \left(a_{jj} + \sum_{i=1, i\neq j}^{n} |a_{ij}| \right), \\ \|A\|_\infty &= \max_{1\leq i\leq n} \sum_{j=1}^{n} |a_{ij}|, \qquad \mu_\infty[A] = \max_{1\leq i\leq n} \left(a_{ii} + \sum_{j=1, j\neq i}^{n} |a_{ij}| \right). \end{aligned}$$

According to [127], the logarithmic matrix norm fulfils $\|e^{A\Delta t}\|_* \leq e^{\mu_*[A]\Delta t}$. *Thus, for the differential equation* $y' = Ay$ *and two different initial values* y_0 *and* $\tilde{y}_0$, *the logarithmic matrix norm related to the norm* $\|\cdot\|_*$ *results in the estimation*

$$\|y(t) - \tilde{y}(t)\|_* \leq e^{(\mu_*[A]\cdot t)} \|y_0 - \tilde{y}_0\|_* .$$

If $\tilde{y} = 0$, *an upper bound for* $\|y(t)\|_*$ *is derived.*

In [41] a splitting formula is applied to (6.5). Thus, the stability function $R(\Delta tA, \Delta tB)$ depends on the time step Δt, the matrix A and the matrix B. The splitting scheme is defined to be $\mathscr{A}$-stable if

$$\begin{aligned} &\|R(\Delta tA, \Delta tB)\|_* \leq 1 \qquad \text{for all } \Delta t > 0 \\ &\text{on } \{\mu_*[A] \leq 0 \text{ and } \mu_*[B] \leq 0\} \text{ for } * = 1 \text{ or } * = \infty . \end{aligned}$$

However, in order to prove their results the authors assume that the stability function fulfils $R(\Delta tA, \Delta tB) = R_1(\Delta tA)R_2(\Delta tB)$. Obviously, this is not true for extrapolated splitting methods. In this chapter we use a modified definition of stability. We investigate a linear test equation

$$y' = A \cdot y + B \cdot y \tag{6.47}$$

with 2×2-matrices A and B. In general it holds $AB \neq BA$. Therefore, a splitting error occurs and the test equation (6.47) is adequate for stability analysis.

Theorem 6.8. *Consider the two-dimensional linear test equation*

$$y' = A \cdot y + B \cdot y \qquad \text{with } A, B \in \mathbb{R}^{2\times 2} .$$

The extrapolated Lie-Trotter splitting results in a recursion

$$\begin{aligned} y_0 &= y(t_0), \\ y_i &= R(\Delta tA, \Delta tB) y_{i-1} \approx y(t_0 + i\Delta t) . \end{aligned}$$

Assume that the matrices A and B satisfy the following conditions:

A1 *the matrices A and B are diagonalizable,*

A2 *the matrix A is a real diagonal matrix,*

A3 $\mu_*[A] \leq 0$ *and* $\mu_*[B] \leq 0$ *for* $* \in \{1, \infty\}$.

Then, the stability function fulfils

$$\|R(\Delta tA, \Delta tB)\|_\infty \leq 1 \qquad \text{for all } \Delta t > 0 .$$

Proof. We distinguish two different cases. First we assume that the matrix B has real eigenvalues. The assumptions **A1** and **A2** result in matrices

$$\begin{aligned} A &= \begin{pmatrix} a_{11} & 0 \\ 0 & a_{22} \end{pmatrix} \in \mathbb{R}^{2\times 2}, \\ B &= \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} u_{11} & u_{12} \\ u_{21} & u_{22} \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} u_{22} & -u_{12} \\ -u_{21} & u_{11} \end{pmatrix} \\ &= \begin{pmatrix} \lambda_1 u_{11}u_{22} - \lambda_2 u_{12}u_{21} & (\lambda_2 - \lambda_1)u_{11}u_{12} \\ (\lambda_1 - \lambda_2)u_{22}u_{21} & \lambda_2 u_{11}u_{22} - \lambda_1 u_{12}u_{21} \end{pmatrix} \in \mathbb{R}^{2\times 2} \\ &\text{with } u_{11}u_{22} - u_{12}u_{21} = 1\,. \end{aligned}$$

The extrapolated Lie-Trotter splitting (6.4) with step size Δt is given by

$$y_i = \left(2e^{\frac{\Delta tA}{2}} e^{\frac{\Delta tB}{2}} e^{\frac{\Delta tA}{2}} e^{\frac{\Delta tB}{2}} - e^{\Delta tA} e^{\Delta tB}\right) y_{i-1}\,.$$

We use submultiplicativity of $\|\cdot\|_\infty$ in order to obtain

$$\begin{aligned} \|R(\Delta tA, \Delta tB)\|_\infty &= \|2e^{\frac{\Delta tA}{2}} e^{\frac{\Delta tB}{2}} e^{\frac{\Delta tA}{2}} e^{\frac{\Delta tB}{2}} - e^{\Delta tA} e^{\Delta tB}\|_\infty \\ &\le \|e^{\frac{\Delta tA}{2}}\|_\infty \cdot \|2e^{\frac{\Delta tB}{2}} e^{\frac{\Delta tA}{2}} - e^{\frac{\Delta tA}{2}} e^{\frac{\Delta tB}{2}}\|_\infty \cdot \|e^{\frac{\Delta tB}{2}}\|_\infty \\ &\le \|2e^{\frac{\Delta tB}{2}} e^{\frac{\Delta tA}{2}} - e^{\frac{\Delta tA}{2}} e^{\frac{\Delta tB}{2}}\|_\infty\,. \end{aligned}$$

Due to assumption **A3** and Remark 6.7 it holds

$$\|e^{\frac{\Delta tA}{2}}\|_\infty \le 1 \text{ and } \|e^{\frac{\Delta tB}{2}}\|_\infty \le 1\,.$$

Thus, we have to show

$$\|2e^{\frac{\Delta tB}{2}} e^{\frac{\Delta tA}{2}} - e^{\frac{\Delta tA}{2}} e^{\frac{\Delta tB}{2}}\|_\infty \le 1\,.$$

We define the matrix $Z := 2e^{\frac{\Delta tB}{2}} e^{\frac{\Delta tA}{2}} - e^{\frac{\Delta tA}{2}} e^{\frac{\Delta tB}{2}}$. Standard computation results in

$$\begin{aligned} Z_{11} &= e^{\lambda_1 \Delta t/2} e^{a_{11}\Delta t/2} u_{11}u_{22} - e^{\lambda_2 \Delta t/2} e^{a_{11}\Delta t/2} u_{12}u_{21}, \\ Z_{12} &= (2e^{a_{22}\Delta t/2} - e^{a_{11}\Delta t/2})(e^{\lambda_2 \Delta t/2} - e^{\lambda_1 \Delta t/2}) u_{11}u_{12}, \\ Z_{21} &= (2e^{a_{11}\Delta t/2} - e^{a_{22}\Delta t/2})(e^{\lambda_1 \Delta t/2} - e^{\lambda_2 \Delta t/2}) u_{22}u_{21}, \\ Z_{22} &= e^{\lambda_2 \Delta t/2} e^{a_{22}\Delta t/2} u_{11}u_{22} - e^{\lambda_1 \Delta t/2} e^{a_{22}\Delta t/2} u_{12}u_{21}\,. \end{aligned} \tag{6.48}$$

Without loss of generality, we assume $a_{22} \leq a_{11} \leq 0$, which results in $e^{a_{22}\Delta t/2} \leq e^{a_{11}\Delta t/2}$ and $|2e^{a_{22}\Delta t/2} - e^{a_{11}\Delta t/2}| \leq e^{a_{11}\Delta t/2} \leq 1$. Thereby $a_{11} > 0$ would imply $\mu_*[A] > 0$. Furthermore, from assumption **A3** we deduce $\|e^{\frac{\Delta tA}{2}}\|_\infty,\ \|e^{\frac{\Delta tB}{2}}\|_\infty \leq 1$. Therefore, we obtain

$$|Z_{11}| + |Z_{12}| \leq 1 \tag{6.49}$$

by the following computation:

$$\begin{aligned}
|Z_{11}| + |Z_{12}| &= e^{a_{11}\Delta t/2}|e^{\lambda_1\Delta t/2}u_{11}u_{22} - e^{\lambda_2\Delta t/2}u_{12}u_{21}| + |(2e^{a_{22}\Delta t/2} - e^{a_{11}\Delta t/2})(e^{\lambda_2\Delta t/2} - e^{\lambda_1\Delta t/2})u_{11}u_{12}| \\
&\leq e^{a_{11}\Delta t/2}|e^{\lambda_1\Delta t/2}u_{11}u_{22} - e^{\lambda_2\Delta t/2}u_{12}u_{21}| + |2e^{a_{22}\Delta t/2} - e^{a_{11}\Delta t/2}| \cdot |(e^{\lambda_2\Delta t/2} - e^{\lambda_1\Delta t/2})u_{11}u_{12}| \\
&\leq e^{a_{11}\Delta t/2}|e^{\lambda_1\Delta t/2}u_{11}u_{22} - e^{\lambda_2\Delta t/2}u_{12}u_{21}| + e^{a_{11}\Delta t/2} \cdot |(e^{\lambda_2\Delta t/2} - e^{\lambda_1\Delta t/2})u_{11}u_{12}| \\
&= e^{a_{11}\Delta t/2} \cdot \left(|e^{\lambda_1\Delta t/2}u_{11}u_{22} - e^{\lambda_2\Delta t/2}u_{12}u_{21}| + |(e^{\lambda_2\Delta t/2} - e^{\lambda_1\Delta t/2})u_{11}u_{12}| \right) \\
&\leq |e^{\lambda_1\Delta t/2}u_{11}u_{22} - e^{\lambda_2\Delta t/2}u_{12}u_{21}| + |(e^{\lambda_2\Delta t/2} - e^{\lambda_1\Delta t/2})u_{11}u_{12}| \\
&= |\left(e^{\Delta tB/2}\right)_{11}| + |\left(e^{\Delta tB/2}\right)_{12}| \\
&\leq 1\,.
\end{aligned}$$

Due to assumption **A3** it holds

$$0 \geq \lambda_1 u_{11}u_{22} - \lambda_2 u_{12}u_{21} + |(\lambda_1 - \lambda_2)u_{22}u_{21}|, \qquad 0 \geq \lambda_2 u_{11}u_{22} - \lambda_1 u_{12}u_{21} + |(\lambda_1 - \lambda_2)u_{22}u_{21}|\,.$$

Adding both these inequalities yields

$$|u_{22}u_{21}| \leq \left| \frac{\lambda_1 + \lambda_2}{2(\lambda_1 - \lambda_2)} \right|. \tag{6.50}$$

Therefore, it holds

$$\begin{aligned}
\left|(e^{\lambda_1\Delta t/2} - e^{\lambda_2\Delta t/2})u_{22}u_{21}\right| &\leq \left| (e^{\lambda_1\Delta t/2} - e^{\lambda_2\Delta t/2}) \frac{\lambda_1\Delta t/2 + \lambda_2\Delta t/2}{2(\lambda_1\Delta t/2 - \lambda_2\Delta t/2)} \right| \\
&:= h(\lambda_1\Delta t/2, \lambda_2\Delta t/2)\,.
\end{aligned} \tag{6.51}$$

We examine the function

$$\tilde{h}(x,y) = (e^x - e^y)\,\frac{x+y}{2\,(x-y)}, \qquad x, y \leq 0\,.$$

It holds

$$-0.5 \leq \tilde{h}(x,y) \leq 0.5 \qquad \text{for } x, y \leq 0\,. \tag{6.52}$$

The statement (6.52) is shown in three steps. First the boundary of the domain $\{(x,y) \,:\, x,y \leq 0\}$ is studied. We assume that $y = 0$ and $x \neq 0$. It follows

$$\tilde{h}(x,0) = \frac{(e^x - 1)}{2} \in (-0.5,\ 0), \qquad \text{for } x < 0\,.$$

Therefore, the absolute of the function $\tilde{h}(x,y)$ is smaller than 0.5 on the boundary of the considered domain. Next we compute the extrema. The gradient of $\tilde{h}(x,y)$ is zero for an extremum. Thus, an extremum fulfils

$$\begin{aligned}\frac{1}{2(x-y)^2}\left(e^x(x^2-y^2-2y)+2ye^y\right) &\stackrel{!}{=} 0,\\ \frac{1}{2(x-y)^2}\left(e^y(y^2-x^2-2x)+2xe^x\right) &\stackrel{!}{=} 0\,.\end{aligned}$$

If the function is evaluated in $x = y$, L'Hôpital's rule results in

$$\lim_{x\to y} \tilde{h}(x,y) = \lim_{x\to y} \frac{e^x(x+y)+(e^x-e^y)}{2} = ye^y \in [-\frac{1}{e},\ 0] \qquad \text{for } y \leq 0\,.$$

Therefore, we can neglect the case $x = y$. An extremum (except $x = y$) fulfils

$$\begin{aligned}\left(e^x(x^2-y^2-2y)+2ye^y\right)+\left(e^y(y^2-x^2-2x)+2xe^x\right) &\stackrel{!}{=} 0,\\ (e^y-e^x)(y^2+2y-x^2-2x) &\stackrel{!}{=} 0,\\ (y^2+2y-x^2-2x) &\stackrel{!}{=} 0,\\ x &\stackrel{!}{=} -y-2\,.\end{aligned}$$

Furthermore, we have

$$\tilde{h}(-y-2,y) \in [-0.5,\ 0] \qquad \text{for } -2 \leq y \leq 0\,.$$

Hence, the absolute values of the extrema are smaller than 0.5. Moreover, due to L'Hôpital's rule, it holds

$$\lim_{x\to-\infty} \tilde{h}(x,y) = -\frac{e^y}{2} \in [-0.5, 0] \qquad \text{for fixed } y \leq 0\,.$$

We have shown that the absolute of the function $\tilde{h}(x,y)$ is smaller than 0.5 on the boundary of the domain and that the absolute of the extrema is smaller than 0.5. In conclusion the equation (6.52) is fulfilled and we obtain

$$|(e^{\lambda_1\Delta t/2} - e^{\lambda_2\Delta t/2})u_{22}u_{21}| \leq h(\lambda_1\Delta t/2, \lambda_2\Delta t/2) \leq 0.5\,. \tag{6.53}$$

Furthermore, $\|e^{B\Delta t/2}\|_\infty \leq 1$ results in

$$|e^{\lambda_2\Delta t/2}u_{11}u_{22} - e^{\lambda_1\Delta t/2}u_{12}u_{21}| + |(e^{\lambda_1\Delta t/2} - e^{\lambda_2\Delta t/2})u_{22}u_{21}| \leq 1 \tag{6.54}$$

With equation (6.53) and (6.54) we obtain

$$|Z_{21}| + |Z_{22}| \leq 1 \tag{6.55}$$

by the following computation:

$$\begin{aligned}
|Z_{21}| + |Z_{22}| &= |(2e^{a_{11}\Delta t/2} - e^{a_{22}\Delta t/2})(e^{\lambda_1\Delta t/2} - e^{\lambda_2\Delta t/2})u_{22}u_{21}| + e^{a_{22}\Delta t/2}|e^{\lambda_2\Delta t/2}u_{11}u_{22} - e^{\lambda_1\Delta t/2}u_{12}u_{21}| \\
&\leq |(2e^{a_{11}\Delta t/2} - e^{a_{22}\Delta t/2})(e^{\lambda_1\Delta t/2} - e^{\lambda_2\Delta t/2})u_{22}u_{21}| + e^{a_{22}\Delta t/2}\Big(1 - |(e^{\lambda_1\Delta t/2} - e^{\lambda_2\Delta t/2})u_{22}u_{21}|\Big) \\
&\leq 2\Big(e^{a_{11}\Delta t/2} - e^{a_{22}\Delta t/2}\Big)\cdot|(e^{\lambda_1\Delta t/2} - e^{\lambda_2\Delta t/2})u_{22}u_{21}| + e^{a_{22}\Delta t/2} \\
&\leq 2\Big(e^{a_{11}\Delta t/2} - e^{a_{22}\Delta t/2}\Big)\cdot\frac{1}{2} + e^{a_{22}\Delta t/2} \\
&= e^{a_{11}\Delta t/2} \leq 1\,.
\end{aligned}$$

In case of real eigenvalues the statement follows with (6.49) and (6.55).

Now we assume that the matrix B has complex eigenvalues. The matrix B is a real matrix. Hence, the complex eigenvalues and the corresponding eigenvectors appear in pairs. Therefore, the matrix B fulfils

$$\begin{aligned}
B &= \frac{1}{v_{11}\bar{v}_{21} - \bar{v}_{11}v_{21}}\begin{pmatrix} v_{11} & \bar{v}_{11} \\ v_{21} & \bar{v}_{21}\end{pmatrix}\begin{pmatrix}\lambda & 0 \\ 0 & \bar{\lambda}\end{pmatrix}\begin{pmatrix}\bar{v}_{21} & -\bar{v}_{11} \\ -v_{21} & v_{11}\end{pmatrix} \\
&= \frac{1}{v_{11}\bar{v}_{21} - \bar{v}_{11}v_{21}}\begin{pmatrix}\lambda v_{11}\bar{v}_{21} - \bar{\lambda}\bar{v}_{11}v_{21} & (\bar{\lambda}-\lambda)|v_{11}|^2 \\ (\lambda-\bar{\lambda})|v_{21}|^2 & \bar{\lambda}v_{11}\bar{v}_{21} - \lambda\bar{v}_{11}v_{21}\end{pmatrix}.
\end{aligned}$$

The eigenvalue λ is complex. Thus, the eigenvalue has the form $\lambda = b + ic$ with $b \in \mathbb{R}$ and $c \in \mathbb{R}$. Again we compute Z and obtain

$$\begin{aligned}
Z_{11} &= \frac{1}{v_{11}\bar{v}_{21} - \bar{v}_{11}v_{21}}e^{a_{11}\Delta t/2}\Big[e^{\lambda\Delta t/2}v_{11}\bar{v}_{21} - e^{\bar{\lambda}\Delta t/2}\bar{v}_{11}v_{21}\Big], \\
Z_{12} &= \frac{1}{v_{11}\bar{v}_{21} - \bar{v}_{11}v_{21}}\Big[(2e^{a_{22}\Delta t/2} - e^{a_{11}\Delta t/2})(e^{\bar{\lambda}\Delta t/2} - e^{\lambda\Delta t/2})|v_{11}|^2\Big], \\
Z_{21} &= \frac{1}{v_{11}\bar{v}_{21} - \bar{v}_{11}v_{21}}\Big[(2e^{a_{11}\Delta t/2} - e^{a_{22}\Delta t/2})(e^{\lambda\Delta t/2} - e^{\bar{\lambda}\Delta t/2})|v_{21}|^2\Big], \\
Z_{22} &= \frac{1}{v_{11}\bar{v}_{21} - \bar{v}_{11}v_{21}}e^{a_{22}\Delta t/2}\Big[e^{\bar{\lambda}\Delta t/2}v_{11}\bar{v}_{21} - e^{\lambda\Delta t/2}\bar{v}_{11}v_{21}\Big]\,.
\end{aligned}$$

Without loss of generality we assume $a_{22} \leq a_{11} \leq 0$, which results in

$$|Z_{11}| + |Z_{12}| \leq 1\,. \tag{6.56}$$

The eigenvalues of B are $\lambda = b + ic$ and $\bar{\lambda} = b - ic$. $B_{11} + B_{22} + 2|B_{21}| \leq 0$ implies

$$(\lambda + \bar{\lambda}) + 2 \cdot \left| \frac{(\lambda - \bar{\lambda})|v_{21}|^2}{v_{11}\bar{v}_{21} - \bar{v}_{11}v_{21}} \right| \leq 0\,.$$

Hence, it holds

$$\left| \frac{2|v_{21}|^2}{v_{11}\bar{v}_{21} - \bar{v}_{11}v_{21}} \right| \leq \left| \frac{b}{c} \right|.$$

It follows

$$\begin{aligned}
\left| \frac{(e^{\lambda\Delta t/2} - e^{\bar{\lambda}\Delta t/2})|v_{21}|^2}{v_{11}\bar{v}_{21} - \bar{v}_{11}v_{21}} \right| &= \left| \frac{e^{b\Delta t/2}\big(\cos(c\Delta t/2) + i\sin(c\Delta t/2) - \cos(-c\Delta t/2) - i\sin(c\Delta t/2)\big)|v_{21}|^2}{v_{11}\bar{v}_{21} - \bar{v}_{11}v_{21}} \right| \\
&= \left| \frac{2e^{b\Delta t/2} i\sin(c\Delta t/2)|v_{21}|^2}{v_{11}\bar{v}_{21} - \bar{v}_{11}v_{21}} \right| \\
&\leq \left| e^{b\Delta t/2} i\sin(c\Delta t/2) \right| \cdot \left| \frac{2|v_{21}|^2}{v_{11}\bar{v}_{21} - \bar{v}_{11}v_{21}} \right| \\
&\leq \left| e^{b\Delta t/2} i\sin(c\Delta t/2) \right| \cdot \left| \frac{b}{c} \right| \\
&= \left| e^{b\Delta t/2} b\Delta t/2 \right| \cdot \left| \frac{\sin(c\Delta t/2)}{c\Delta t/2} \right| \\
&\leq \frac{1}{e} \cdot \left| \frac{\sin(c\Delta t/2)}{c\Delta t/2} \right| \leq \frac{1}{e} \leq 0.5\,.
\end{aligned}$$

Thereby we used that $b \leq 0$. Similar to equation (6.55), it follows that

$$\begin{aligned}
|Z_{21}| + |Z_{22}| &= \left| \frac{(2e^{a_{11}\Delta t/2} - e^{a_{22}\Delta t/2})(e^{\lambda\Delta t/2} - e^{\bar{\lambda}\Delta t/2})|v_{21}|^2}{v_{11}\bar{v}_{21} - \bar{v}_{11}v_{21}} \right| + \left| e^{a_{22}\Delta t/2} \frac{e^{\bar{\lambda}\Delta t/2} v_{11}\bar{v}_{21} - e^{\lambda\Delta t/2}\bar{v}_{11}v_{21}}{v_{11}\bar{v}_{21} - \bar{v}_{11}v_{21}} \right| \\
&\leq (2e^{a_{11}\Delta t/2} - e^{a_{22}\Delta t/2}) \left| \frac{(e^{\lambda\Delta t/2} - e^{\bar{\lambda}\Delta t/2})|v_{21}|^2}{v_{11}\bar{v}_{21} - \bar{v}_{11}v_{21}} \right| + e^{a_{22}\Delta t/2} \left| \frac{e^{\bar{\lambda}\Delta t/2} v_{11}\bar{v}_{21} - e^{\lambda\Delta t/2}\bar{v}_{11}v_{21}}{v_{11}\bar{v}_{21} - \bar{v}_{11}v_{21}} \right| \\
&\leq (2e^{a_{11}\Delta t/2} - e^{a_{22}\Delta t/2}) \left| \frac{(e^{\lambda\Delta t/2} - e^{\bar{\lambda}\Delta t/2})|v_{21}|^2}{v_{11}\bar{v}_{21} - \bar{v}_{11}v_{21}} \right| + e^{a_{22}\Delta t/2} \left(1 - \left| \frac{(e^{\lambda\Delta t/2} - e^{\bar{\lambda}\Delta t/2})|v_{21}|^2}{v_{11}\bar{v}_{21} - \bar{v}_{11}v_{21}} \right| \right) \\
&= 2(e^{a_{11}\Delta t/2} - e^{a_{22}\Delta t/2}) \left| \frac{(e^{\lambda\Delta t/2} - e^{\bar{\lambda}\Delta t/2})|v_{21}|^2}{v_{11}\bar{v}_{21} - \bar{v}_{11}v_{21}} \right| + e^{a_{22}\Delta t/2} \\
&\leq 2(e^{a_{11}\Delta t/2} - e^{a_{22}\Delta t/2})\frac{1}{2} + e^{a_{22}\Delta t/2} = e^{a_{11}\Delta t/2} \leq 1\,.
\end{aligned}$$

Therefore, we have

$$\|R(\Delta tA, \Delta tB)\|_\infty \leq 1\,.$$

□

Remark 6.9. *The operator A is the chemical source term. In general the eigenvalues of the chemical source term are real, non-positive values. Hence, the assumption that the diagonalisation of A is a real matrix is reasonable.*

Remark 6.10. *Assume that the matrix A from equation* (6.47) *is not a diagonal matrix, but it is diagonalizable. Thus, there exists a transformation matrix* V_D *such that* $\bar{A} = V_D A V_D^{-1}$ *is a diagonal matrix. Furthermore, we define* $\bar{B} = V_D B V_D^{-1}$*. The basis transformation of the test equation* (6.47) *yields the system*

$$\bar{y}' = \bar{A}\bar{y} + \bar{B}\bar{y}\,. \tag{6.57}$$

We can apply Theorem 6.8 to equation (6.57)*. Therefore, sufficient conditions exist, which guarantee* $\|R(\Delta t\bar{A}, \Delta t\bar{B})\| \leq 1$*. For the system* (6.47) *we obtain that*

$$\begin{aligned} y_i &= R(\Delta tA, \Delta tB)^i y(t_0) = V_D \bar{y}_i = V_D R(\Delta t\bar{A}, \Delta t\bar{B})^i V_D^{-1} y(t_0), \\ \|y_i\| &\leq \|V_D\| \cdot \|R(\Delta t\bar{A}, \Delta t\bar{B})\|^i \cdot \|V_D^{-1}\| \cdot \|y(t_0)\|\,. \end{aligned}$$

Hence, errors will not grow without any limit. The system is stable.

6.6 Numerical examples

In this section, the results of Theorem 6.1 and Theorem 6.2 are illustrated. First, a linear example is considered. Afterwards, three nonlinear chemical reaction systems are examined. Thereby, the convergence order of the extrapolated Lie-Trotter scheme is compared to the convergence order of the Lie-Trotter splitting and the convergence order of the Strang splitting.

6.6.1 Linear example

In this subsection, we examine a linear chemical reaction system with advection on the spatial domain $\Omega = [0,1]$. The differential equation is given by

$$\begin{aligned} \partial_t C(x,t) &= F\left(C(x,t)\right) - a \cdot \nabla C(x,t) \qquad \text{for } x \in \Omega \subset \mathbb{R},\ t \in [0,1], \\ C(x,0) &= \left(1 - \sqrt{x}, 0, 0\right)^T\,. \end{aligned}$$

The components $0 \leq a_i \leq 1$ of the vector a are the velocities of the different chemical species and the function $F(C)$ is the chemical source term. We consider the time interval $[0,1]$ and we are interested in

the state of the system at $(x,t) = (1, 0.5)$. Thus, the boundary condition at the inflow is of no importance. The occurring chemical reactions are

$$\begin{aligned} A_1 \to A_2, \quad & k_1 = 1, \quad & A_2 \to A_1, \quad & k_{-1} = 10000, \\ A_3 \to A_2, \quad & k_2 = 1, \quad & A_2 \to A_3, \quad & k_{-2} = 10000\,. \end{aligned}$$

Therefore, the chemical source term is given by

$$F(C) = \begin{pmatrix} -k_1 & k_{-1} & 0 \\ k_1 & -k_{-1} - k_{-2} & k_2 \\ 0 & k_{-2} & -k_2 \end{pmatrix} \cdot \begin{pmatrix} C_1 \\ C_2 \\ C_3 \end{pmatrix} =: A \cdot C$$

The eigenvalues of the Jacobian matrix of the linear chemical source term are given by -20001, -1, and 0. The smallest timescale (stiffness parameter) ϵ of the chemistry-only equation can be estimated by the reciprocal of the largest absolute value of the eigenvalues. Therefore, it holds $\epsilon \approx 5 \cdot 10^{-5}$. In order to clarify the numerical results, we also consider the case $k_{-1} = k_{-2} = 1000$. In this case the eigenvalues of the Jacobian are -2001, -1, and 0, which yields $\epsilon \approx 5 \cdot 10^{-4}$. Furthermore, the velocity is given by $a = (0, 1, 0)^T$. Hence, we assume that species A_1 and A_3 are attached to the walls. Note, if the velocity vector a is a multiple of $\vec{1}$, transport and chemistry commute and no splitting error occurs.

For this example, the logarithmic matrix norms (related to $\|\cdot\|_1$ and $\|\cdot\|_\infty$) of the chemical source term are $\mu_1[A] = 0$ and $\mu_\infty[A] = k_{-1} - k_1 = 9999$. Hence, the assumption **A3** of Theorem 6.8 is not fulfilled. After the transformation to a diagonal matrix $\tilde{A}$, we obtain $\mu_1[\tilde{A}] = \mu_\infty[\tilde{A}] = 0$ (compare to Remark 6.10). However, in this example the stability of the extrapolated Lie-Trotter splitting follows from the conservation of the 1-norm for the chemistry-only step and the transport-only step (with inflow equal outflow). The conservation for the substeps implies the conservation for the Lie-Trotter splitting. For small step sizes the relative error of the Lie-Trotter splitting is small enough such that the solution of the extrapolated scheme will be positive. Thus, the extrapolated scheme also conserves the 1-norm.

The order of the different splitting methods is examined with a step size sequence $\Delta t \in \{0.5^i,\ 1 \leq i \leq 15\}$. An ODE of the form (6.1) is obtained by a discretization with $1/\Delta t$ equidistant grid points. Moreover, the chemistry-only and the transport-only equations are solved exactly. For the considered example, the exact solution of the corresponding transport equation is obtained by shifting the concentration values of the species A_2. The exact solution of the linear chemistry equation is computed with an eigenvector transformation. Afterwards the splitting error at $(x,t) = (1, 0.5)$ can be evaluated. The splitting error depends on the chosen splitting method $(\cdot)$ and the time step Δt. We define the error by

$$e_{(\cdot)}(\Delta t) := \frac{\|C_{(\cdot)}(x = 1, t = 0.5, \Delta t) - C_{ref}\|_2}{\|C_{ref}\|_2}\,.$$

Thereby C_{ref} is the solution at $(x,t) = (1, 0.5)$, which is obtained with the finest splitting time step. For the different splitting approaches, the splitting error $e_{(\cdot)}(\Delta t)$ is plotted in Figures 11 and 12. In agreement with [76] the Lie-Trotter splitting has order one for all Δt. The Strang splitting has order one

for $\Delta t > \epsilon$ and order two for $\Delta t < \epsilon$. Thus, order reduction occurs for the Strang splitting. Furthermore, the extrapolated splitting scheme has order two for all Δt. However, an additional error of size $\mathcal{O}(\epsilon)$ is introduced for $\Delta t > \epsilon$. The additional error of size $\mathcal{O}(\epsilon)$ is predicted in Theorem 6.2. An important feature of Figures 11 and 12 is, that the extrapolated splitting method has the smallest splitting error for large step sizes.

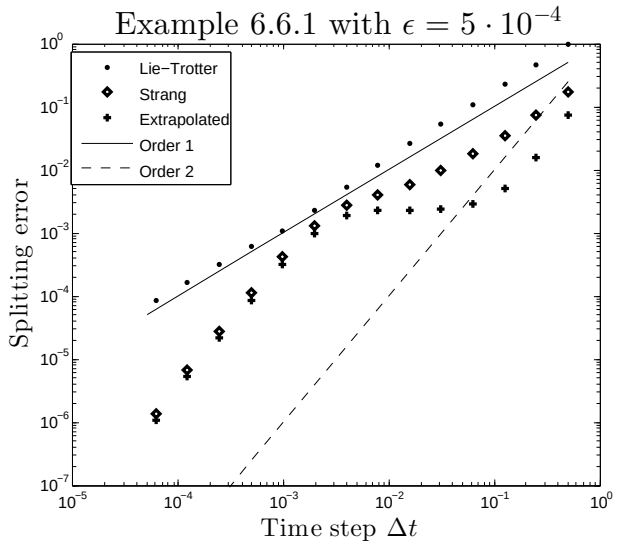

Figure 11: Splitting error $e_{(\cdot)}(\Delta t)$ plotted against the splitting time step Δt for the stiffness parameter $\epsilon \approx 5 \cdot 10^{-4}$

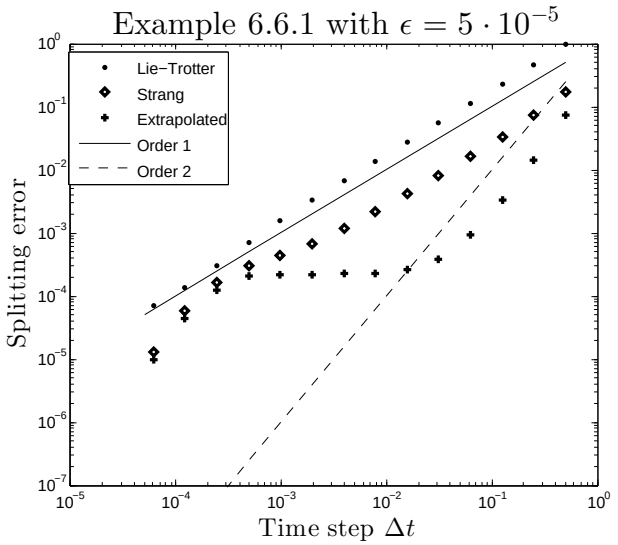

Figure 12: Splitting error $e_{(\cdot)}(\Delta t)$ plotted against the splitting time step Δt for the stiffness parameter $\epsilon \approx 5 \cdot 10^{-5}$

Remark 6.11. *If the stiffness parameter ϵ of the system is smaller than the required accuracy, the extrapolated splitting approach has order two. In this case the required step size of the extrapolated splitting approach is much larger than the required step sizes of the Lie-Trotter splitting and the Strang splitting.*

6.6.2 Slow dimerisation

The first nonlinear example bases upon a slow dimerisation reaction of a fast species [22]. The considered differential equation is

$$\begin{aligned} \partial_t C(x,t) &= F(C(x,t)) - a \cdot \nabla C(x,t) \qquad \text{for } x \in \Omega = [0,6],\ t \in [0,6], \\ C(x,0) &= 100 \cdot \left(0, 1 - \sqrt{x/6}, 0\right)^T . \end{aligned}$$

The velocity vector is given by $a = (1,0,0)^T$. Hence, the species A_1 is moved by advection, while the species A_2 and A_3 are attached to the wall. The chemical reaction system of the slow dimerisation [22] is given by

$$\begin{aligned} A_1 &\to A_2, & k_1 &= 200, \\ A_2 &\to A_1, & k_{-1} &= 1, \\ A_1 + A_1 &\to A_3, & k_2 &= 1\,. \end{aligned}$$

For this example, the general examination is very similar to Subsection 6.6.1. However, the chemistry-only differential equation is not solved exactly. Instead we use the matlab routine *ode15s* with tolerances 10^{-8}. Furthermore, the splitting error of the different splitting schemes is evaluated at $(x,t) = (6,3)$. The obtained results are plotted in Figure 13. Similar to the linear example in Subsection 6.6.1, we observe that the Lie-Trotter splitting is a first order scheme. Moreover, the Strang splitting suffers from order reduction. For step sizes larger than 0.05, the order is only one. However, for smaller step sizes, the usual order two is obtained. Finally, we note that the order of the extrapolated Lie-Trotter splitting is two for all step sizes. However, for step sizes larger than 0.05, an additional error with size ~ 0.005 is introduced. Hence, this example confirms the theoretical results from Section 6.4.

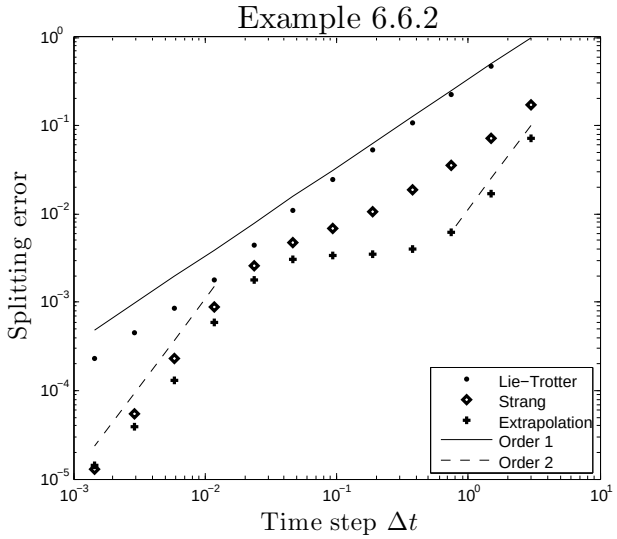

Figure 13: Splitting error $e_{(\cdot)}(\Delta t)$ plotted against the splitting time step Δt for the slow dimerisation

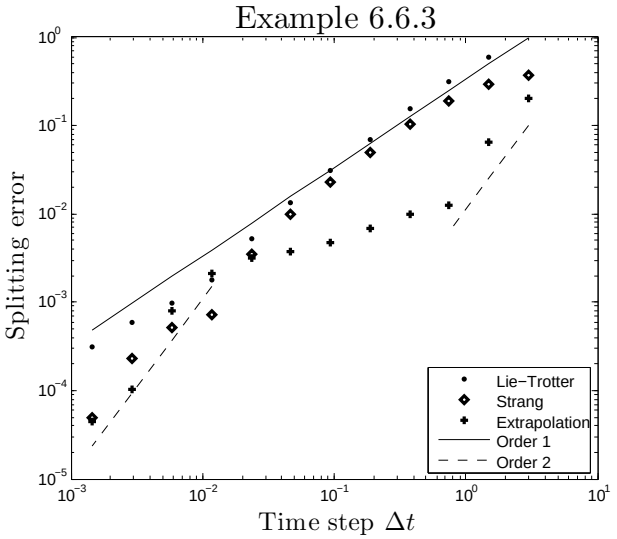

Figure 14: Splitting error $e_{(\cdot)}(\Delta t)$ plotted against the splitting time step Δt for the fast dimerisation

6.6.3 Fast dimerisation

A fast dimerisation [22] is the second nonlinear example. The corresponding chemical reaction system is given by

$$\begin{aligned} A_1 &\to A_2, && k_1 = 1, \\ A_2 + A_2 &\leftrightarrow A_3, && k_2 = 1000, \qquad k_3 = 100, \\ A_3 &\to A_4, && k_4 = 1\,. \end{aligned}$$

In comparison to Subsection 6.6.2 we change the advection velocity and the initial values. The considered advection velocity is $a = (0,1,0,0)^T$ and the considered initial values are

$$C(x,0) = 100 \cdot \left(1 - \sqrt{x/6}, 0, 0, 0\right)^T .$$

The obtained results are plotted in Figure 14. Again the Lie-Trotter splitting is a first order scheme. Furthermore, the Strang splitting suffers from order reduction. For large splitting step sizes the order is only one. However, for step sizes smaller than 0.01, the usual order two is obtained. In comparison, the extrapolated Lie-Trotter splitting has order two for all step sizes. However, for step sizes larger than 0.01, an additional error in $\mathcal{O}(\epsilon)$ is visible. Thus, the result of this example accords with the theoretical results from Section 6.4.

6.6.4 Extremely stiff reaction system

In this subsection, the extremely stiff chemical reaction system (4.23)–(4.29) is considered. The fastest timescale ϵ of this chemical reaction system is smaller than 10^{-6}. Thus, in case of a closed chemical reactor, the state of the system is slowly evolving after an extremely fast transient phase. The time evolution of some species is plotted in Figure 15 and Figure 16 for the initial values (4.30). These initial values correspond to a total concentration of $12.5\frac{\text{mol}}{\text{m}^3} \approx 1\text{bar}$. The fast transient phase is plotted in Figure 15. The timescales of the fast chemical processes are smaller than $3 \cdot 10^{-5}$ seconds. The slow evolution of the system is given in Figure 16. The steady state of the system is not reached after 240 seconds. In this subsection, the chemical reaction system (4.23) – (4.29) is coupled with a transport term. For simplicity, a continuous ideally stirred-tank reactor (CISTR) is analyzed. Thereby an additional inflow is given. Hence, the chemical reactions take place in the perfect mixed reactor and the transport term is represented by an additional source term. This setting results in an ODE. Therefore, operator splitting is not necessary for the examination of the considered problem. However, all characteristics of a transport-chemistry model are given. Thus, the example is sufficient in order to verify Theorem 6.1 and Theorem 6.2 for an authentic chemical reaction system. Moreover, a reference solution can be computed easily. In the following, the chemical reactor has the initial values (4.30) and the considered timespan is $[0, 240]$ seconds.

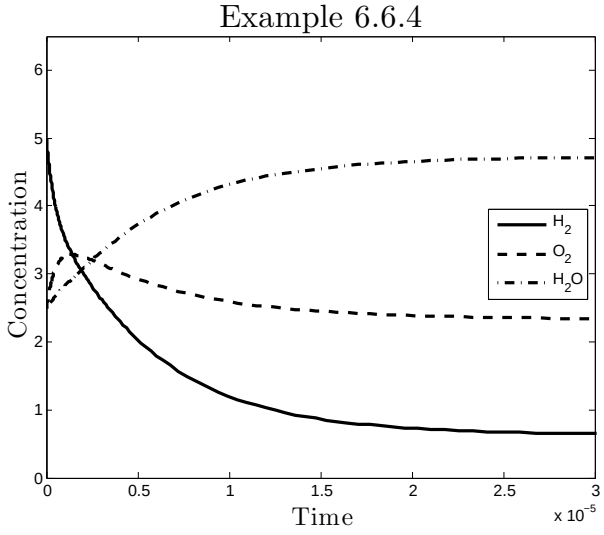

Figure 15: Fast transient phase of the species H_2, O_2 and H_2O

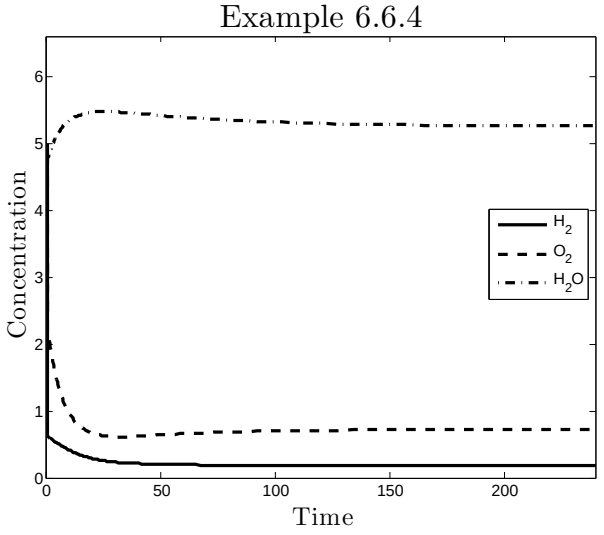

Figure 16: Time evolution of the species H_2, O_2 and H_2O

The inflow depends on the current state of the chemical reactor and results in an additional source term for the concentration $[H_2]$ of hydrogen molecules. The additional source term is given by

$$G([H_2]) = \frac{5 \text{ mol m}^{-3} - [H_2]}{120\text{s}}. \tag{6.58}$$

A possible explanation of (6.58) is that the ideally stirred-tank reactor is connected to a large reservoir of H_2 by a semi-permeable wall. Thereby, the concentration of hydrogen molecules is 5 mol m^{-3} in the large reservoir.

The obtained differential equation has the form (6.1) and is solved by operator splitting. The chemistry-only equation is solved by the algorithm RADAU5 [64] with a very high accuracy. The transport-only equation is a linear differential equation for $[H_2]$, which is solved exactly.

The order of the different splitting methods is examined with a step size sequence $\Delta t \in \{240 \cdot 0.5^i,\ 3 \leq i \leq 14\}$. The splitting error of a splitting method $(\cdot)$ is defined by

$$e_{(\cdot)}(\Delta t) := \left\| C_{(\cdot)}(t = 240, \Delta t) - C_{ref} \right\|_2.$$

The reference solution C_{ref} is the solution that is obtained with the full ODE (6.1). The splitting error for Lie-Trotter splitting, Strang splitting, and extrapolated Lie-Trotter splitting is given in Figure 17. The order of the Lie-Trotter splitting is one. The Strang splitting suffers from order reduction so that the order is only one instead of two. The extrapolated Lie-Trotter splitting has order two for step sizes larger than 10^{-1} seconds. Therefore, the results verify Theorem 6.1 and Theorem 6.2. The stiffness parameter ϵ is extremely small. Therefore, the additional error of size $\mathcal{O}(\epsilon)$ does not matter for reasonable step sizes and the error of the extrapolated Lie-Trotter splitting is in $\mathcal{O}(\Delta t^2)$.

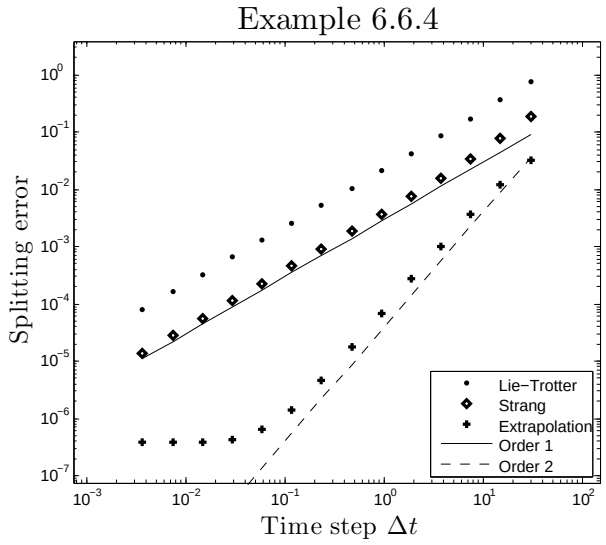

Figure 17: Splitting error $e_{(\cdot)}(\Delta t)$ plotted against splitting time step Δt for inflow (6.58)

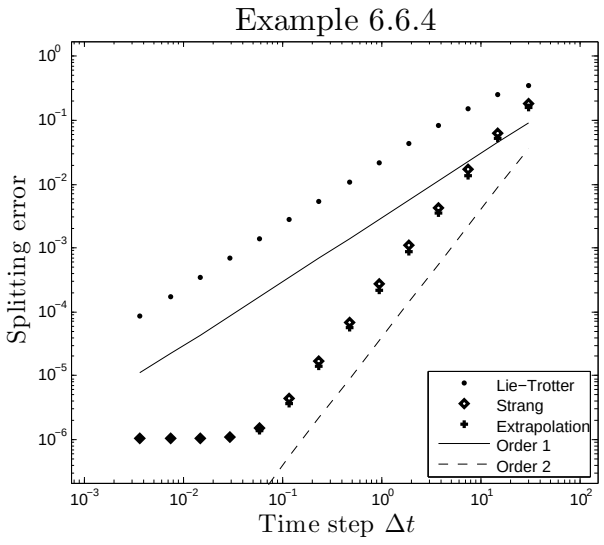

Figure 18: Splitting error $e_{(\cdot)}(\Delta t)$ plotted against splitting time step Δt for inflow (6.59)

Note that the order of the Strang splitting method depends on the considered inflow $G(C)$. The rate of the inflow (6.58) is proportional to $5 - [H_2]$. Furthermore, the timescale of the species H_2 is very small. Hence, the concentration $[H_2]$ is given by the partial equilibrium of the fast chemical reactions. As a result, the inflowing hydrogen is rapidly transformed into other species and the partial equilibrium is slowly shifted. Thus, the inflow $5 - [H_2]$ is almost constant for small time intervals. Thereby the concentration of H_2 is increased by the inflow and it is decreased by the chemical reactions. However, in a transport-only equation the rate of the inflow changes fast, because the linear transport equation implies exponential convergence of $[H_2]$ to 5. Afterwards, in the chemistry-only step the concentration $[H_2]$ converges to the partial equilibrium of the fast processes, which is smaller than 5. The equilibration has a very small timescale (see Figure 15), which is much smaller than the smallest splitting time step ($240 \cdot 0.5^{15} \approx 7 \cdot 10^{-3}$). Therefore, in case of Lie-Trotter or Strang splitting method the total amount of inflowing H_2 depends on the length of the splitting time step. Hence, the computed inflow depends on the chosen step size and the existence of fast chemical processes is crucial for the system with inflow (6.58). For comparison consider the inflow

$$G(C) = \frac{C_0 - C}{120}. \tag{6.59}$$

The inflow (6.59) also depends on the current concentration. Although, the total amount of inflowing material depends on the pressure inside the chemical reactor. Moreover, the chemical reactions do not change the pressure, because the number of educts is equal to the number of products for (4.23) – (4.29). Thus, the existence of chemical reactions does not influence the total amount of inflowing material. In particular, the full scheme, in which fast chemical reactions transform inflowing material directly produces similar results as a conventional splitting scheme, which transforms all inflowing material after the transport step. Therefore, the existence of very fast chemical processes does not influence the order of the Strang splitting (the splitting error of the scheme with inflow (6.59) is plotted in Figure 18). Hence, stiffness, which is introduced by timescale separation of chemical reactions, does not always result in order reduction for the Strang splitting scheme.

6.7 Conclusions and summary

In this chapter we have examined the splitting error of the extrapolated Lie-Trotter splitting for the simulation of chemistry with transport. Usually, the timescales of chemical reaction systems cover several different magnitudes. Therefore, chemical reaction systems result in stiff differential equations. Stiffness causes problems for explicit integration schemes, and implicit integration schemes are necessary. However, for large-scale applications, the computational cost of the implicit integration is prohibitive. Hence, splitting methods are used. Splitting methods result in chemistry-only and transport-only differential equations, which can be solved cheaper than the original problem. Nevertheless, splitting methods introduce a splitting error, which depends on the used time step. The most popular splitting methods are the Lie-Trotter splitting and the Strang splitting. The Lie-Trotter splitting has order one, and the Strang splitting has order two. Unfortunately, the Strang splitting suffers from order reduction. If the used time step Δt is larger than the smallest timescale ϵ of the chemical system, the order of the Strang splitting reduces to one. Therefore, the Lie-Trotter splitting as well as the Strang splitting have order one for

reasonable time steps. Thus, we have proposed the Richardson extrapolation of the Lie-Trotter splitting as a second order scheme, which does not suffer from order reduction in the stiff case. We have proven, that the global splitting error of the extrapolated scheme is in $\mathcal{O}(\Delta t^2) + \mathcal{O}(\epsilon)$ for $\epsilon \ll \Delta t$. Furthermore, we have examined the stability of the extrapolated scheme. Opposed to approaches of other authors, the stability of the single operators combined with submultiplicativity of the considered norm cannot be used in this case. Hence, we have investigated a two-dimensional test equation and prove stability for this test equation. The considered numerical examples have shown that the extrapolated scheme results in larger time steps for a required accuracy larger than ϵ.

7 Summary and outlook

7.1 Summary

In this work the numerical simulation of chemical reaction systems has been examined. Firstly, the differential equations describing chemical reaction systems have been introduced. Thereby spatially homogeneous reactors result in ordinary differential equations whereas spatial inhomogeneous reactors are modelled by partial differential equations. Usually, the resulting differential equations are stiff because the chemical reactions cover many different timescales with a range from 10^{-9} to 10^{2} seconds.

In Chapter 3, the stiffness has been used in order to simplify the considered differential equation. Due to fast chemical reactions there exist processes, whose timescales are smaller than the used step size. Hence, these fast processes reach equilibrium in a fraction of an integration time step, and can be approximated by the equilibrium. If automatic reduction mechanisms are applicable, great savings in computing time are possible. Although, most reduction mechanisms rely on the existence of a low-dimensional attracting manifold because a look-up table has to be precomputed in order to save computing time by the dimension reduction.

This requirement is often not fulfilled. Hence, a new approach [85] by Lukassen and Kiehl, which does not depend on the existence of a low-dimensional attracting manifold, has been introduced in Chapter 4. The proposed new method reduces the range of the occurring timescales, and thereby reduces the stiffness of the differential equation. The reduction of stiffness results in a decrease of the occurring round-off errors in the evaluation of the chemical source term. Hence, the proposed approach avoids the failure of Newton's method and leads to a reduction in computing time. Furthermore, the dimension of the differential equation is not changed.

Besides the reduction of the stiffness and the dimension of a chemical reaction system, the existence of fast processes can be used for parameter identification. This has been ellaborated in Chapter 5. The algebraic equations given by PEA and QSSA contain informations about the unknown parameters. Thus, the dimension of the parameter space can be reduced. Then the computing time of parameter identification is decreased because it correlates to the number of unknown parameters.

In Chapters 3 to 5, spatially homogeneous reactors, which are described by ODEs, have been examined in order to simplify the derivation of the different approaches. The results of Chapters 3 to 5 can be applied to heterogeneous reactors if splitting methods, which solve a sequence of transport-only and chemistry-only differential equations, are used. Recommended splitting methods are the Lie-Trotter splitting and the Strang splitting. However, order reduction occurs in case of a stiff differential equation. In Chapter 6 it has been shown that the extrapolated Lie-Trotter splitting is a second order integration method for stiff transport-chemistry systems.

In this PhD thesis the simulation of chemical reaction systems has been examined thoroughly. However, simplification have often been necessary or the discussion has been limited to one of several possible approaches. Hence, open research fields or possible extensions of the given methods are elaborated in the following section.

Several opportunities for further research are introduced in this section. There are four main fields for improvements of the presented results.

First of all, the considered differential equation 2.13 contains several simplifications. The most important simplification is the assumption of a continuum. In this PhD thesis the concentrations of chemical species are variables in the positive real numbers. However, molecules come in whole numbers. If the size of the considered reactor is large enough, the differential equations (2.4) and (2.13) work very well. Although, if the considered reactor is very small (e.g., a living cell), discreteness and stochasticity have to be considered. Suitable approaches are listed in [23, 55]. A possible research topic is the stochastic simulation of a chemical reaction system with large timescale separation.

Another aspect is the special structure of the ODE that results from discretizing a chemical reaction system in space. This structure can be utilised by an approximation of the Jacobian matrix of the source term by a block diagonally matrix. Thereby each block is defined by the chemical source term in a spatial gridpoint, and the coupling between the blocks, which is given by slow transport, is neglected. Using a block diagonally approximation of the Jacobian matrix in an inexact Newton method decreases the computational effort. However, an additional error, which has to be analysed in future work, is introduced.

Furthermore, in this work the velocity of fast processes is set infinity (Chapter 3) or it is reduced (Chapter 4). Both approaches are justified by PEA and QSSA. Although, a cheap, reliable, and effective a-posteriori error estimator for the introduced approximation error is not available. Hence, providing a cheap a-posteriori error estimator for methods that are based on the separation of timescales is an open question for further work. If a cheap a-posteriori error estimator for reduction mechanisms like ILDM is on hand, the adaptive usage of reduction mechanisms in parameter estimation of unknown reaction rate constants is possible. Thus, an extension of this PhD thesis is the derivation of such an error estimator. Note that parameter estimation requires the computation of the sensitivity matrix. Switching between different reduced models, different step sizes, or/and different splitting schemes leads to very unreliable results, if the sensitivity matrix is computed by difference approximation [72]. Similar to step size freezing [72], the algorithm has to be adapted in order to increase the accuracy of the sensitivity matrix.

Moreover, the analysis of operator splitting methods can be expanded. In this work operator splitting is considered for a stiff chemical source term and a non-stiff transport term. Moreover, a major assumption is that the corresponding ODE can be transformed to (6.10). Hence, a major assumption is that the timescales are separated. However, diffusion can also result in stiffness of the advection-diffusion equation, and timescale separation is not guaranteed. In [32, 33] the Lie-operator is used for nonlinear stiff reaction-diffusion systems that have a fast diffusion process. Although, extrapolated schemes as well as advection-diffusion-reaction systems are not considered in [32, 33]. Thus, an additional analysis of splitting methods in case of a stiff chemical source term and a stiff diffusion term is a possibility for future research. Furthermore, in Chapter 6 the analysis of the splitting approach is based on the exact solution of the subsystems. However, the subsystems are solved with a numerical integration method. Therefore, the subsystems are not solved exactly, and the analysis of the splitting approach should incorporate a discretization error. Thereby the order of the used integration method defines the dependence of the discretization error on the used step size. Hence, the order of the splitting approach depends on the order

of the used integration method, but there is no difference between different integration methods with the same order. Although, if the solution of the subsystems is approximated with an integration method, the stability of the method depends on the discretization method. Thus, the stability analysis is only valid for the used discretization method. If the discretization method is changed, the stability analysis has to be executed again. Another aspect is the used splitting method. Lie-Trotter splitting and Strang splitting are the most common splitting methods, and many authors [69, 121, 122, 124, 141] examine them thoroughly. Moreover, in Chapter 6 the extrapolated Lie-Trotter splitting is analysed. However, there are many other splitting approaches, which can be used for advection-diffusion-reaction systems and have (dis)advantages over the described extrapolated Lie-Trotter splitting. E. g., for the class of Douglas splitting methods [4, 67] the solution of each substep is consistent with the exact solution. Thus, steady state solutions of (6.1) are stationary points of the Douglas splitting methods. The (extrapolated) Lie-Trotter splitting as well as the Strang splitting do not own this property. Therefore, future work can include an extension to other splitting methods, and unlimited time can be spent on the analysis of splitting methods in combination with different integration methods. Furthermore, efficiency of numerical integration methods depends on adaptive step size control. Thereby adaptive step size control requires an error estimator for the splitting error. Two splitting schemes with different order or one splitting scheme with two different step sizes can be used for estimation of the splitting error. Both approaches are compromised by order reduction. Hence, the influence of order reduction on adaptive step size control is an important research topic for future work.

As a result this PhD thesis provides solutions for existing problems, but it also gives rise to new reasearch topics and further work.

References

[1] R. H. Abrams and K. Loague. A compartmentalized solute transport model for redox zones in contaminated aquifers: 1. Theory and development. *Water Resources Research*, 36(8):2001–2013, 2000.

[2] G. Y. Adusei and A. Fontijn. Experimental studies of Cl-atom reactions at high temperatures: $Cl+H_2 \rightarrow HCl+H$ from 291 to 1283 K. In *Symposium (International) on Combustion*, volume 25, pages 801–808. Elsevier, 1994.

[3] R. Allez and J.-P. Bouchaud. Eigenvector dynamics: general theory and some applications. *Physical Review E*, 86(4), 2012.

[4] A. Arrarás, W. Hundsdorfer, L. Portero, et al. Modified Douglas splitting methods for reaction–diffusion equations. *BIT Numerical Mathematics*, pages 1–25, 2015.

[5] S. Arrhenius. Über die Reaktionsgeschwindigkeit bei der Inversion von Rohrzucker durch Säuren. *Zeitschrift für physikalische Chemie*, 4(1):226–248, 1889.

[6] R. Atkinson, D. Baulch, R. Cox, R. Hampson Jr, J. Kerr, M. Rossi, and J. Troe. Evaluated kinetic and photochemical data for atmospheric chemistry: supplement vi. iupac subcommittee on gas kinetic data evaluation for atmospheric chemistry. *Journal of Physical and Chemical Reference Data*, 26(6):1329–1499, 1997.

[7] W. Auzinger, R. Frank, and F. Macsek. Asymptotic error expansions for stiff equations: the implicit Euler scheme. *SIAM Journal on Numerical Analysis*, 27(1):67–104, 1990.

[8] M. Baerns, H. Hofmann, and A. Renken. *Chemische Reaktionstechnik*. Wiley-VCH Weinheim, Germany, 2004.

[9] J. Bailey. Lumping analysis of reactions in continuous mixtures. *The Chemical Engineering Journal*, 3:52–61, 1972.

[10] F. L. Bauer and C. T. Fike. Norms and exclusion theorems. *Numerische Mathematik*, 2(1):137–141, 1960.

[11] S. W. Benson. The induction period in chain reactions. *The Journal of Chemical Physics*, 20(10):1605–1612, 1952.

[12] I. Bey, D. J. Jacob, R. M. Yantosca, J. A. Logan, B. D. Field, A. M. Fiore, Q. Li, H. Y. Liu, L. J. Mickley, and M. G. Schultz. Global modeling of tropospheric chemistry with assimilated meteorology: Model description and evaluation. *Journal of Geophysical Research: Atmospheres*, 106(D19):23073–23095, 2001.

[13] T. Blasenbrey. *Entwicklung und Implementierung automatisch reduzierter Reaktionsmechanismen für die Verbrennung von Kohlenwasserstoffen*. PhD thesis, Universität Stuttgart, 2000.

[14] H. Bock. A Multiple Shooting Method for Parameter Identification in Nonlinear Differential Equations. In *GAMM Conference, Brussels*, 1978.

[15] H. G. Bock. *Numerical Treatment of Inverse Problems in Chemical Reaction Kinetics*. Springer, 1981.

[16] H. Bongers, J. Van Oijen, and L. De Goey. Intrinsic low-dimensional manifold method extended with diffusion. *Proceedings of the Combustion Institute*, 29(1):1371–1378, 2002.

[17] J. Bowen, A. Acrivos, and A. Oppenheim. Singular perturbation refinement to quasi-steady state approximation in chemical kinetics. *Chemical Engineering Science*, 18(3):177–188, 1963.

[18] V. Bykov, I. Goldfarb, and V. Gol'dshtein. Singularly perturbed vector fields. In *Journal of Physics: Conference Series*, volume 55, page 28. IOP Publishing, 2006.

[19] V. Bykov, V. Gol'dshtein, and U. Maas. Simple global reduction technique based on decomposition approach. *Combustion Theory and Modelling*, 12(2):389–405, 2008.

[20] V. Bykov and U. Maas. Extension of the ILDM method to the domain of slow chemistry. *Proceedings of the Combustion Institute*, 31(1):465–472, 2007.

[21] V. Bykov and U. Maas. The extension of the ILDM concept to reaction–diffusion manifolds. *Combustion Theory and Modelling*, 11(6):839–862, 2007.

[22] Y. Cao, D. Gillespie, and L. Petzold. Multiscale stochastic simulation algorithm with stochastic partial equilibrium assumption for chemically reacting systems. *Journal of Computational Physics*, 206(2):395–411, 2005.

[23] Y. Cao, D. T. Gillespie, and L. R. Petzold. The slow-scale stochastic simulation algorithm. *The Journal of Chemical Physics*, 122(1):014116, 2005.

[24] T. Carrington and N. Davidson. Shock waves in chemical kinetics: The rate of dissociation of n204. *The Journal of Physical Chemistry*, 57(4):418–427, 1953.

[25] S. Chaturantabut and D. C. Sorensen. A state space error estimate for POD-DEIM nonlinear model reduction. *SIAM Journal on Numerical Analysis*, 50(1):46–63, 2012.

[26] A. W. Cook, J. J. Riley, and G. Kosály. A laminar flamelet approach to subgrid-scale chemistry in turbulent flows. *Combustion and Flame*, 109(3):332–341, 1997.

[27] T. Cui, Y. M. Marzouk, and K. E. Willcox. Data-driven model reduction for the Bayesian solution of inverse problems. *International Journal for Numerical Methods in Engineering*, 102(5):966–990, 2015.

[28] C. Curtiss and J. O. Hirschfelder. Integration of stiff equations. *Proceedings of the National Academy of Sciences*, 38(3):235–243, 1952.

[29] S. J. Danby and T. Echekki. Proper orthogonal decomposition analysis of autoignition simulation data of nonhomogeneous hydrogen–air mixtures. *Combustion and Flame*, 144(1):126–138, 2006.

[30] C. Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

[31] C. De Dieuleveult, J. Erhel, and M. Kern. A global strategy for solving reactive transport equations. *Journal of Computational Physics*, 228(17):6395–6410, 2009.

[32] S. Descombes, M. Duarte, T. Dumont, F. Laurent, V. Louvet, and M. Massot. Analysis of operator splitting in the nonasymptotic regime for nonlinear reaction-diffusion equations. Application to the dynamics of premixed flames. *SIAM Journal on Numerical Analysis*, 52(3):1311–1334, 2014.

[33] S. Descombes and M. Massot. Operator splitting for nonlinear reaction-diffusion systems with an entropic structure: singular perturbation and order reduction. *Numerische Mathematik*, 97(4):667–698, 2004.

[34] P. Deuflhard. Recent progress in extrapolation methods for ordinary differential equations. *SIAM Review*, 27(4):505–535, 1985.

[35] P. Deuflhard, E. Hairer, and J. Zugck. One-step and extrapolation methods for differential-algebraic systems. *Numerische Mathematik*, 51(5):501–516, 1987.

[36] P. Deuflhard, J. Heroth, and U. Maas. *Towards dynamic dimension reduction in reactive flow problems*. ZIB, 1996.

[37] P. Deuflhard and A. Hohmann. *Numerische Mathematik 1: Eine algorithmisch orientierte Einführung*. Walter de Gruyter, Berlin, 2008.

[38] R. Djouad and B. Sportisse. Some reduction techniques for simplifying atmospheric chemical kinetics. In *Air Pollution Modelling and Simulation*, pages 235–244. Springer, 2002.

[39] V. Druskin and M. Zaslavsky. On combining model reduction and Gauss–Newton algorithms for inverse partial differential equation problems. *Inverse Problems*, 23(4):1599, 2007.

[40] I. S. Duff and G. A. Watson. *The State of the Art in Numerical Analysis*. Number 63. Oxford University Press, 1997.

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

[42] L. Einkemmer, M. Moccaldi, and A. Ostermann. Efficient boundary corrected strang splitting. *Applied Mathematics and Computation*, 332:76–89, 2018.

[43] L. Einkemmer and A. Ostermann. A comparison of boundary correction methods for strang splitting. *arXiv preprint arXiv:1609.05505*, 2016.

[44] L. Einkemmer and A. Ostermann. Overcoming order reduction in diffusion-reaction splitting. part 2: oblique boundary conditions. *SIAM Journal on Scientific Computing*, 38(6):A3741–A3757, 2016.

[45] P. Englezos and N. Kalogerakis. *Applied Parameter Estimation for Chemical Engineers*. CRC Press, 2000.

[46] C. Engstler. *MATLAB code RADAU5*. Mathematisches Institut, Universität Tübingen, 1999.

[47] I. Faragó, Á. Havasi, and Z. Zlatev. Richardson-extrapolated sequential splitting and its application. *Journal of Computational and Applied Mathematics*, 226(2):218–227, 2009.

[48] N. Fenichel. Geometric singular perturbation theory for ordinary differential equations. *Journal of Differential Equations*, 31(1):53–98, 1979.

[49] J. G. Francis. The QR transformation a unitary analogue to the LR transformation – Part 1. *The Computer Journal*, 4(3):265–271, 1961.

[50] J. G. Francis. The QR transformation – part 2. *The Computer Journal*, 4(4):332–345, 1962.

[51] S. Franz. *Modellierung und Simulation reaktionskinetischer Prozesse mit Hilfe dynamisch angepasster quasi-stationärer Zustände*. Logos-Verlag, 2003.

[52] A. A. Frost, R. G. Pearson, F. G. Helfferich, and U. Schindewolf. *Kinetik und Mechanismen homogener chemischer Reaktion*. Verlag Chemie, 1973.

[53] D. Garmatter, B. Haasdonk, and B. Harrach. A reduced basis Landweber method for nonlinear inverse problems. *Inverse Problems*, 32(3):035001, 2016.

[54] Z. P. Gerdtzen, P. Daoutidis, and W.-S. Hu. Non-linear reduction for kinetic models of metabolic reaction networks. *Metabolic Engineering*, 6(2):140–154, 2004.

[55] D. T. Gillespie. Stochastic simulation of chemical kinetics. *Annual Review of Physical Chemistry*, 58:35–55, 2007.

[56] V. Giovangigli and M. Massot. Entropic structure of multicomponent reactive flows with partial equilibrium reduced chemistry. *Mathematical Methods in the Applied Sciences*, 27(7):739–768, 2004.

[57] J. Goddard. Consequences of the partial-equilibrium approximation for chemical reaction and transport. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 431, pages 271–284. The Royal Society, 1990.

[58] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The John's Hopkins Univ Press, Baltimore, 1983.

[59] G. H. Golub and C. F. Van Loan. *Matrix Computations*, volume 3. The John's Hopkins Univ Press, Baltimore, 2012.

[60] B. Haasdonk. Reduced basis methods for parametrized PDEs–A tutorial introduction for stationary and instationary problems. *Reduced Order Modelling. Luminy Book series*, 2014.

[61] B. Haasdonk and M. Ohlberger. Efficient reduced models and a posteriori error estimation for parametrized dynamical systems by offline/online decomposition. *Mathematical and Computer Modelling of Dynamical Systems*, 17(2):145–161, 2011.

[62] J. W. Haefner. *Modeling Biological Systems: Principles and Applications*. Springer Science & Business Media, 2005.

[63] E. Hairer and C. Lubich. Asymptotic expansions of the global error of fixed-stepsize methods. *Numerische Mathematik*, 45(3):345–360, 1984.

[64] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II. Stiff and Differential-Algebraic Problems*. Springer, Berlin, 1996.

[65] C. Homescu, L. R. Petzold, and R. Serban. Error Estimation for Reduced-Order Models of Dynamical Systems. *SIAM Journal on Numerical Analysis*, 43(4):1693–1714, 2005.

[66] F. Hoppensteadt, P. Alfeld, and R. Aiken. Numerical Treatment of Rapid Chemical Kinetics by Perturbation and Projection Methods. In *Modelling of Chemical Reaction Systems*, pages 31–37. Springer, 1981.

[67] W. Hundsdorfer. A note on stability of the Douglas splitting method. *Mathematics of Computation*, pages 183–190, 1998.

[68] M. Z. Jacobson. *Fundamentals of Atmospheric Modeling*. Cambridge university press, 2005.

[69] T. Jahnke and C. Lubich. Error bounds for exponential operator splittings. *BIT Numerical Mathematics*, 40(4):735–744, 2000.

[70] J. F. Kanney, C. T. Miller, and C. T. Kelley. Convergence of iterative split-operator approaches for approximating nonlinear reactive transport problems. *Advances in Water Resources*, 26(3):247–261, 2003.

[71] H. G. Kaper and T. J. Kaper. Asymptotic analysis of two reduction methods for systems of chemical reactions. *Physica D: Nonlinear Phenomena*, 165(1):66–93, 2002.

[72] M. Kiehl. Sensitivity analysis of ODEs and DAEs—theory and implementation guide. *Optimization Methods and Software*, 10(6):803–821, 1999.

[73] O. Koch, C. Neuhauser, and M. Thalhammer. Embedded exponential operator splitting methods for the time integration of nonlinear evolution equations. *Applied Numerical Mathematics*, 63:14–24, 2013.

[74] O. Koch and M. Thalhammer. Embedded split-step formulae for the time integration of nonlinear evolution equations. *preprint*, 2010.

[75] K. König and U. Maas. On-demand generation of reduced mechanisms based on hierarchically extended intrinsic low-dimensional manifolds in generalized coordinates. *Proceedings of the Combustion Institute*, 32(1):553–560, 2009.

[76] R. Kozlov, A. Kværnø, and B. Owren. The behaviour of the local error in splitting methods applied to stiff problems. *Journal of Computational Physics*, 195(2):576–593, 2004.

[77] S. Lam. Singular perturbation for stiff equations using numerical methods. In *Recent Advances in the Aerospace Sciences*, pages 3–19. Springer, 1985.

[78] S. Lam and D. Goussis. The CSP method for simplifying kinetics. *International Journal of Chemical Kinetics*, 26(4):461–486, 1994.

[79] D. Lanser and J. G. Verwer. Analysis of operator splitting for advection–diffusion–reaction problems from air pollution modelling. *Journal of Computational and Applied Mathematics*, 111(1):201–216, 1999.

[80] J. D. Lawson and J. L. Morris. The extrapolation of first order methods for parabolic partial differential equations. I. *SIAM Journal on Numerical Analysis*, 15(6):1212–1224, 1978.

[81] G. Li and H. Rabitz. A general analysis of exact lumping in chemical kinetics. *Chemical Engineering Science*, 44(6):1413–1430, 1989.

[82] G. Li, A. S. Tomlin, H. Rabitz, and J. Tóth. A general analysis of approximate nonlinear lumping in chemical kinetics. I. Unconstrained lumping. *The Journal of Chemical Physics*, 101(2):1172–1187, 1994.

[83] P. Linstrom and W. Mallard. NIST chemistry webbook, NIST standard reference database number 69, National Institute of Standards and Technology, Gaithersburg MD, 20899, 2010.

[84] T. Lohmann, H. G. Bock, and J. P. Schloeder. Numerical methods for parameter estimation and optimal experiment design in chemical reaction systems. *Industrial & Engineering Chemistry Research*, 31(1):54–57, 1992.

[85] A. Lukassen and M. Kiehl. Reduction of round-off errors in chemical kinetics. *Combustion Theory and Modelling*, 21(2):183–204, 2017.

[86] A. A. Lukassen and M. Kiehl. Parameter estimation with model order reduction and global measurements. *PAMM*, 17(1):773–774, 2017.

[87] A. A. Lukassen and M. Kiehl. Operator splitting for chemical reaction systems with fast chemistry. *Journal of Computational and Applied Mathematics*, 344:495–511, 2018.

[88] A. A. Lukassen and M. Kiehl. Parameter estimation with model order reduction for elliptic differential equations. *Inverse Problems in Science and Engineering*, 26(4):479–497, 2018.

[89] U. Maas. Coupling of chemical reaction with flow and molecular transport. *Applications of Mathematics*, 40(3):249–266, 1995.

[90] U. Maas. Efficient calculation of intrinsic low-dimensional manifolds for the simplification of chemical kinetics. *Computing and Visualization in Science*, 1(2):69–81, 1998.

[91] U. Maas and S. B. Pope. Implementation of simplified chemical kinetics based on intrinsic low-dimensional manifolds. In *Symposium (International) on Combustion*, volume 24, pages 103–112. Elsevier, 1992.

[92] U. Maas and S. B. Pope. Simplifying chemical kinetics: intrinsic low-dimensional manifolds in composition space. *Combustion and Flame*, 88(3):239–264, 1992.

[93] J. Manion, R. Huie, R. Levin, D. Burgess Jr, V. Orkin, W. Tsang, W. McGivern, J. Hudgens, V. Knyazev, D. Atkinson, E. Chai, A. Tereza, C. Lin, T. Allison, W. Mallard, F. Westley, J. Herron, R. Hampson, and D. Frizzell. NIST chemical kinetics database, NIST standard reference database 17, version 7.0 (web version), release 1.6.8, data version 2015.12, National Institute of Standards and Technology, Gaithersburg, Maryland, 20899-8320, 2015.

[94] R. Martí. Multi-start methods. In F. Glover and G. Kochenberger, editors, *Handbook of Metaheuristics*, pages 355–368. Springer US, 2003.

[95] P. Mendes and D. Kell. Non-linear optimization of biochemical pathways: applications to metabolic engineering and parameter estimation. *Bioinformatics*, 14(10):869–883, 1998.

[96] P. K. Moore and L. R. Petzold. A stepsize control strategy for stiff systems of ordinary differential equations. *Applied Numerical Mathematics*, 15(4):449–463, 1994.

[97] D. R. Mott. *New Quasi-Steady-State and Partial-Equilibrium Methods for Integrating Chemically Reacting Systems*. PhD thesis, Citeseer, 1999.

[98] H. Niemann. *Niedrigdimensionale Modellierung Dynamischer Systeme am Beispiel reduzierter Reaktionsmechanismen*. PhD thesis, Ruprecht-Karls-Universität Heidelberg, 2003.

[99] I. E. Nikerel, W. A. van Winden, P. J. Verheijen, and J. J. Heijnen. Model reduction and a priori kinetic parameter identifiability analysis using metabolome time series for metabolic reaction networks with linlog kinetics. *Metabolic Engineering*, 11(1):20–30, 2009.

[100] K. Nipp. Invariant manifolds of singularly perturbed ordinary differential equations. *Zeitschrift für Angewandte Mathematik und Physik (ZAMP)*, 36(2):309–320, 1985.

[101] K. Nipp. Numerical integration of stiff ODE's of singular perturbation type. *Zeitschrift für Angewandte Mathematik und Physik (ZAMP)*, 42(1):53–79, 1991.

[102] U. Nowak and P. Deuflhard. Numerical identification of selected rate constants in large chemical reaction systems. *Applied Numerical Mathematics*, 1(1):59–75, 1985.

[103] A. T. Patera and G. Rozza. *Reduced Basis Approximation and A Posteriori Error Estimation for Parametrized Partial Differential Equations*. MIT Cambridge, 2006.

[104] R. Penrose. A generalized inverse for matrices. In *Mathematical proceedings of the Cambridge philosophical society*, volume 51, pages 406–413. Cambridge Univ Press, 1955.

[105] M. D. Porter and G. B. Skinner. The steady-state approximation in free-radical calculations. A numerical example. *Journal of Chemical Education*, 53(6):366, 1976.

[106] A. Prothero and A. Robinson. On the stability and accuracy of one-step methods for solving stiff systems of ordinary differential equations. *Mathematics of Computation*, 28(125):145–162, 1974.

[107] A. Quarteroni, A. Manzoni, and F. Negri. *Reduced Basis Methods for Partial Differential Equations: An Introduction*, volume 92. Springer, 2015.

[108] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical Mathematics*, volume 37. Springer Science & Business Media, 2010.

[109] S. Ravindran. Adaptive reduced-order controllers for a thermal flow system using proper orthogonal decomposition. *SIAM Journal on Scientific Computing*, 23(6):1924–1942, 2002.

[110] M. Rein. The partial-equilibrium approximation in reacting flows. *Physics of Fluids A: Fluid Dynamics (1989-1993)*, 4(5):873–886, 1992.

[111] L. F. Richardson. The approximate arithmetical solution by finite differences of physical problems involving differential equations, with an application to the stresses in a masonry dam. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 210:307–357, 1911.

[112] L. F. Richardson and J. A. Gaunt. The deferred approach to the limit. Part I. Single lattice. Part II. Interpenetrating lattices. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 226:299–361, 1927.

[113] D. L. Ropp and J. N. Shadid. Stability of operator splitting methods for systems with indefinite operators: Advection–diffusion–reaction systems. *Journal of Computational Physics*, 228(9):3508–3516, 2009.

[114] M. W. Saaltink, C. Ayora, and J. Carrera. A mathematical formulation for reactive transport that eliminates mineral concentrations. *Water Resources Research*, 34(7):1649–1656, 1998.

[115] B. A. Schramm. *Automatische Reduktion chemischer Reaktionsmechanismen am Beispiel der Oxidation von höheren Kohlenwasserstoffen und deren Verwendung in reaktiven Strömungen*. PhD thesis, Ruprecht-Karls-Universität Heidelberg, 2002.

[116] D. A. Schwer, P. Lu, W. H. Green, and V. Semiao. A consistent-splitting approach to computing stiff steady-state reacting flows with adaptive chemistry. *Combustion Theory and Modelling*, 7(2):383–399, 2003.

[117] A. Siade. *Model Reduction and Parameter Estimation in Groundwater Modeling*. PhD thesis, 2012.

[118] M. Singer, S. Pope, and H. Najm. Operator-splitting with ISAT to model reacting flow with detailed chemistry. *Combustion Theory and Modelling*, 10(2):199–217, 2006.

[119] M. A. Singer, S. B. Pope, and H. N. Najm. Modeling unsteady reacting flow with operator splitting and ISAT. *Combustion and Flame*, 147(1):150–162, 2006.

[120] G. Söderlind. The logarithmic norm. History and modern theory. *BIT Numerical Mathematics*, 46(3):631–652, 2006.

[121] B. Sportisse. An analysis of operator splitting techniques in the stiff case. *Journal of Computational Physics*, 161(1):140–168, 2000.

[122] B. Sportisse, G. Bencteux, and P. Plion. Method of Lines versus Operator Splitting for reaction–diffusiuon systems with fast chemistry. *Environmental Modelling & Software*, 15(6):673–679, 2000.

[123] B. Sportisse and R. Djouad. Use of proper orthogonal decomposition for the reduction of atmospheric chemical kinetics. *Journal of Geophysical Research: Atmospheres*, 112(D6), 2007.

[124] G. Strang. On the construction and comparison of difference schemes. *SIAM Journal on Numerical Analysis*, 5(3):506–517, 1968.

[125] K. Strehmel and R. Weiner. *Linear-implizite Runge-Kutta-Methoden und ihre Anwendungen (Teubner-Texte zur Mathematik)*. Vieweg+Teubner Verlag, 1992.

[126] K. Strehmel, R. Weiner, and H. Podhaisky. *Numerik gewöhnlicher Differentialgleichungen: nichtsteife, steife und differential-algebraische Gleichungen*. Springer Science & Business Media, 2012.

[127] T. Ström. On logarithmic norms. *SIAM Journal on Numerical Analysis*, 12(5):741–753, 1975.

[128] A. Tikhonov, A. Vasileva, and A. Sveshnikov. *Differential Equations*. Springer Berlin Heidelberg, 1985.

[129] L. N. Trefethen and D. Bau III. *Numerical Linear Algebra*, volume 50. Society for Industrial and Applied Mathematics, 1997.

[130] H. F. Trotter. On the product of semi-groups of operators. *Proceedings of the American Mathematical Society*, 10(4):545–551, 1959.

[131] T. Turanyi, A. Tomlin, and M. Pilling. On the error of the quasi-steady-state approximation. *The Journal of Physical Chemistry*, 97(1):163–172, 1993.

[132] M. Valorani and D. A. Goussis. Explicit time-scale splitting algorithm for stiff problems: auto-ignition of gaseous mixtures behind a steady shock. *Journal of Computational Physics*, 169(1):44–79, 2001.

[133] M. Valorani and S. Paolucci. Adaptive model reduction in chemical kinetics. *Italian Section of the Combustion Institute, Turin, Italy*, 2008.

[134] M. Valorani and S. Paolucci. The G-Scheme: A framework for multi-scale adaptive model reduction. *Journal of Computational Physics*, 228(13):4665–4701, 2009.

[135] A. Van den Bos. *Parameter Estimation for Scientists and Engineers*. John Wiley & Sons, 2007.

[136] J. Van Oijen and L. De Goey. Modelling of premixed laminar flames using flamelet-generated manifolds. *Combustion Science and Technology*, 161(1):113–137, 2000.

[137] J. Verwer and H. De Vries. Global extrapolation of a first order splitting method. *SIAM Journal on Scientific and Statistical Computing*, 6(3):771–780, 1985.

[138] A. Walter, E. Frind, D. Blowes, C. Ptacek, and J. Molson. Modeling of multicomponent reactive transport in groundwater: 1. model development and evaluation. *Water Resources Research*, 30(11):3137–3148, 1994.

[139] J. Warnatz, U. Maas, and R. W. Dibble. *Combustion*, volume 3. Springer, 2001.

[140] D. Wirtz, D. Sorensen, and B. Haasdonk. A posteriori error estimation for DEIM reduced nonlinear dynamical systems. *SIAM Journal on Scientific Computing*, 36(2):A311–A338, 2014.

[141] B. Yang and S. Pope. An investigation of the accuracy of manifold methods and splitting schemes in the computational implementation of combustion chemistry. *Combustion and Flame*, 112(1-2):16–32, 1998.

[142] A. Zagaris, H. G. Kaper, and T. J. Kaper. Analysis of the computational singular perturbation reduction method for chemical kinetics. *Journal of Nonlinear Science*, 14(1):59–91, 2004.

[143] Z. Zlatev, I. Dimov, I. Faragó, K. Georgiev, and Á. Havasi. Stability of the Richardson Extrapolation combined with some implicit Runge–Kutta methods. *Journal of Computational and Applied Mathematics*, 310:224–240, 2017.

[144] Z. Zlatev, I. Faragó, and Á. Havasi. Richardson Extrapolation combined with the sequential splitting procedure and the θ-method. *Open Mathematics*, 10(1):159–172, 2012.

Appendix

A.1 Measured data for Example 12

The measured data of Example 12 is

$$C^m(10^2\text{s}) = \left(1.5349 \cdot 10^{18},\quad 3.1936 \cdot 10^{16},\quad 5.7779 \cdot 10^{15},\quad 2.6236 \cdot 10^{12},\quad 8.1561 \cdot 10^{11}\right)^T,$$
$$C^m(10^3\text{s}) = \left(1.5048 \cdot 10^{18},\quad 2.2017 \cdot 10^{17},\quad 5.3567 \cdot 10^{16},\quad 3.4660 \cdot 10^{11},\quad 5.4667 \cdot 10^{12}\right)^T,$$
$$C^m(10^4\text{s}) = \left(7.7742 \cdot 10^{17},\quad 1.5852 \cdot 10^{18},\quad 3.9191 \cdot 10^{17},\quad 2.7219 \cdot 10^{10},\quad 4.1159 \cdot 10^{13}\right)^T.$$

A.2 Derivation of the local error constants of the considered splitting schemes

Lie-Trotter splitting

Functions $f_1(\Delta t)$ and $f_2(\Delta t)$ are defined in order to derive the error constants of the Lie-Trotter splitting. The defined functions and their derivatives are

$$\begin{aligned} f_1(\Delta t) &:= e^{A\cdot\Delta t}e^{B\cdot\Delta t}, \\ f_1'(\Delta t) &= e^{A\cdot\Delta t}Ae^{B\cdot\Delta t} + e^{A\cdot\Delta t}e^{B\cdot\Delta t}B, \\ f_1''(\Delta t) &= e^{A\cdot\Delta t}A^2e^{B\cdot\Delta t} + 2e^{A\cdot\Delta t}Ae^{B\cdot\Delta t}B + e^{A\cdot\Delta t}e^{B\cdot\Delta t}B^2, \\ f_1^{(3)}(0) &= A^3 + 3A^2B + 3AB^2 + B^3, \end{aligned}$$

and

$$\begin{aligned} f_2(\Delta t) &:= e^{(A+B)\cdot\Delta t}, \\ f_2'(0) &= (A+B), \\ f_2''(0) &= (A+B)^2 = A^2 + AB + BA + B^2, \\ f_2^{(3)}(0) &= (A+B)^3 = A^3 + A^2B + ABA + AB^2 + BA^2 + BAB + B^2A + B^3\,. \end{aligned}$$

The Taylor expansion for Δt close to 0 is used in order to obtain the error estimation. The error of the Lie-Trotter splitting applied to the linear equation (6.5) is

$$\begin{aligned} e_{LT}(\Delta t) &= \left(e^{A\cdot\Delta t}e^{B\cdot\Delta t} - e^{(A+B)\cdot\Delta t}\right)C_0 \\ &= \big(f_1(\Delta t) - f_2(\Delta t)\big)z_0 \\ &= \left(f_1(0) - f_2(0) + \big(f_1'(0) - f_2'(0)\big)\Delta t + \big(f_1''(0) - f_2''(0)\big)\frac{\Delta t^2}{2} + \mathcal{O}(\Delta t^3)\right)C_0 \\ &= \big(AB - BA\big)\frac{\Delta t^2}{2}C_0 + \mathcal{O}(\Delta t^3)\,. \end{aligned}$$

Strang splitting

In order to derive the error constants of the Strang splitting an additional function $f_3(\Delta t)$ is defined. The additional function and its derivatives are

$$
\begin{aligned}
f_3(\Delta t) &:= e^{A\cdot\Delta t/2}e^{B\cdot\Delta t}e^{A\cdot\Delta t/2},\\
f_3'(\Delta t) &= e^{A\cdot\Delta t/2}\frac{A}{2}e^{B\cdot\Delta t}e^{A\cdot\Delta t/2} + e^{A\cdot\Delta t/2}e^{B\cdot\Delta t}Be^{A\cdot\Delta t/2} + e^{A\cdot\Delta t/2}e^{B\cdot\Delta t}e^{A\cdot\Delta t/2}\frac{A}{2},\\
f_3''(\Delta t) &= e^{A\cdot\Delta t/2}\frac{A^2}{4}e^{B\cdot\Delta t}e^{A\cdot\Delta t/2} + 2e^{A\cdot\Delta t/2}\frac{A}{2}e^{B\cdot\Delta t}Be^{A\cdot\Delta t/2}\\
&\quad + 2e^{A\cdot\Delta t/2}\frac{A}{2}e^{B\cdot\Delta t}e^{A\cdot\Delta t/2}\frac{A}{2} + e^{A\cdot\Delta t/2}e^{B\cdot\Delta t}B^2e^{A\cdot\Delta t/2}\\
&\quad + 2e^{A\cdot\Delta t/2}e^{B\cdot\Delta t}Be^{A\cdot\Delta t/2}\frac{A}{2} + e^{A\cdot\Delta t/2}e^{B\cdot\Delta t}e^{A\cdot\Delta t/2}\frac{A^2}{4},\\
f_3^{(3)}(0) &= A^3 + B^3 + 3\frac{A^2B}{4} + 3\frac{AB^2}{2} + 3\frac{ABA}{2} + 3\frac{B^2A}{2} + 3\frac{BA^2}{4}.
\end{aligned}
$$

The Taylor expansion for Δt close to 0 is used in order to obtain the error estimation. The error of the Strang splitting applied to the linear equation (6.5) is

$$
\begin{aligned}
e_S(\Delta t) &= \left(e^{A\cdot\Delta t/2}e^{B\cdot\Delta t}e^{A\cdot\Delta t/2} - e^{(A+B)\cdot\Delta t}\right)C_0\\
&= \big(f_3(\Delta t) - f_2(\Delta t)\big)C_0\\
&= \left(f_3(0) - f_2(0) + \big(f_3'(0) - f_2'(0)\big)\Delta t + \big(f_3''(0) - f_2''(0)\big)\frac{\Delta t^2}{2} + \big(f_3^{(3)}(0) - f_2^{(3)}(0)\big)\frac{\Delta t^3}{6} + \mathcal{O}(\Delta t^4)\right)C_0\\
&= \frac{-A^2B + 2ABA + 2AB^2 - BA^2 - 4BAB + 2B^2A}{24}\Delta t^3C_0 + \mathcal{O}(\Delta t^4).
\end{aligned}
$$

Extrapolated Lie-Trotter splitting

An additional function $f_4(\Delta t)$ is defined in order to derive the error constants of the extrapolated Lie-Trotter splitting. The additional function and its derivatives are

$$
\begin{aligned}
f_4(\Delta t) &:= e^{A\cdot\Delta t/2}e^{B\cdot\Delta t/2}e^{A\cdot\Delta t/2}e^{B\cdot\Delta t/2},\\
f_4'(\Delta t) &= \frac{1}{2}\Big(e^{A\cdot\Delta t/2}Ae^{B\cdot\Delta t/2}e^{A\cdot\Delta t/2}e^{B\cdot\Delta t/2} + e^{A\cdot\Delta t/2}e^{B\cdot\Delta t/2}Be^{A\cdot\Delta t/2}e^{B\cdot\Delta t/2}\\
&\qquad + e^{A\cdot\Delta t/2}e^{B\cdot\Delta t/2}e^{A\cdot\Delta t/2}Ae^{B\cdot\Delta t/2} + e^{A\cdot\Delta t/2}e^{B\cdot\Delta t/2}e^{A\cdot\Delta t/2}e^{B\cdot\Delta t/2}B\Big),\\
f_4''(\Delta t) &= \frac{1}{4}\Big(e^{A\cdot\Delta t/2}A^2e^{B\cdot\Delta t/2}e^{A\cdot\Delta t/2}e^{B\cdot\Delta t/2} + 2e^{A\cdot\Delta t/2}Ae^{B\cdot\Delta t/2}Be^{A\cdot\Delta t/2}e^{B\cdot\Delta t/2}\\
&\qquad + 2e^{A\cdot\Delta t/2}Ae^{B\cdot\Delta t/2}e^{A\cdot\Delta t/2}Ae^{B\cdot\Delta t/2} + 2e^{A\cdot\Delta t/2}Ae^{B\cdot\Delta t/2}e^{A\cdot\Delta t/2}e^{B\cdot\Delta t/2}B\\
&\qquad + e^{A\cdot\Delta t/2}e^{B\cdot\Delta t/2}B^2e^{A\cdot\Delta t/2}e^{B\cdot\Delta t/2} + 2e^{A\cdot\Delta t/2}e^{B\cdot\Delta t/2}Be^{A\cdot\Delta t/2}Ae^{B\cdot\Delta t/2}\\
&\qquad + 2e^{A\cdot\Delta t/2}e^{B\cdot\Delta t/2}Be^{A\cdot\Delta t/2}e^{B\cdot\Delta t/2}B + 2e^{A\cdot\Delta t/2}e^{B\cdot\Delta t/2}e^{A\cdot\Delta t/2}Ae^{B\cdot\Delta t/2}B\\
&\qquad + e^{A\cdot\Delta t/2}e^{B\cdot\Delta t/2}e^{A\cdot\Delta t/2}A^2e^{B\cdot\Delta t/2} + e^{A\cdot\Delta t/2}e^{B\cdot\Delta t/2}e^{A\cdot\Delta t/2}e^{B\cdot\Delta t/2}B^2\Big),\\
f_4^{(3)}(0) &= \frac{1}{8}\Big(8A^3 + 8B^3 + 15A^2B + 15AB^2 + 6ABA + 3BA^2 + 3B^2A + 6BAB\Big).
\end{aligned}
$$

The error of the extrapolated Lie-Trotter splitting applied to the linear equation (6.5) is

$$
\begin{aligned}
e_{RE}(\Delta t) &= \Big(2e^{A\cdot\Delta t/2}e^{B\cdot\Delta t/2}e^{A\cdot\Delta t/2}e^{B\cdot\Delta t/2} - e^{A\cdot\Delta t}e^{B\cdot\Delta t} - e^{(A+B)\cdot\Delta t}\Big)C_0\\
&= \big(2f_4 - f_1(\Delta t) - f_2(\Delta t)\big)C_0\\
&= \Big(2f_4(0) - f_1(0) - f_2(0) + \big(2f_4'(0) - f_1'(0) - f_2'(0)\big)\Delta t + \big(2f_4''(0) - f_1''(0) - f_2''(0)\big)\frac{\Delta t^2}{2}\\
&\quad + \big(2f_4^{(3)}(0) - f_1^{(3)}(0) - f_2^{(3)}(0)\big)\frac{\Delta t^3}{6} + \mathcal{O}(\Delta t^4)\Big)C_0\\
&= \frac{2ABA + 2BAB - A^2B - AB^2 - B^2A - BA^2}{24}\Delta t^3 C_0 + \mathcal{O}(\Delta t^4).
\end{aligned}
$$

Wissenschaftlicher Werdegang

Axel Ariaan Lukassen,

geboren am 19.01.1989 in Aachen

seit 07/2014	Wissenschaftlicher Mitarbeiter der Technischen Universität Darmstadt am Fachbereich Mathematik, Arbeitsgruppe Numerik und wissenschaftliches Rechnen
04/2012 – 03/2014	Studium Mathematik mit Nebenfach Physik an der Technischen Universität Darmstadt. *Abschluss:* Master of Science
04/2009 – 03/2012	Studium Mathematik mit Nebenfach Physik an der Technischen Universität Darmstadt. *Abschluss:* Bachelor of Science
07/2008 – 03/2009	Grundwehrdienst in Walldürn und Speyer
06/2008	Abitur an der Christian-With-Schule in Usingen