# Inaugural-Dissertation

zur Erlangung der Doktorwürde
der
Naturwissenschaftlich-Mathematischen Gesamtfakultät
der
Ruprecht-Karls-Universität
Heidelberg

vorgelegt von
Diplom-Mathematiker Thomas Stricker
aus Leutkirch

Tag der mündlichen Prüfung: 5. Okt. 2006

# Spectral Model Reduction as Preconditioner and Adaptive Solver Component in Chemical Reaction Systems

27.06.2006

1. Gutachter: Prof. Dr. Rolf Rannacher

2. Gutachter: Prof. Dr. Dr. h.c. Hans Georg Bock

## Abstract

Aim of this work is the analysis and application of the reduction methods QSSA and ILDM. Both methods were developed in the context of the modeling of chemical combustion processes. They have in common that they reduce highly the number of species to be modeled, which leads to an enormous reduction of the computation time. The produced systematic error prevents clearly the obtained solution to solve the detailed problem. In some applications the reduced solution has therefore doubtful physical sense.

One focal point of this thesis is hence the investigation of the applicability of the reduction methods as preconditioner for the solver. The treated problems are stationary systems of equations. This way of the application ensures the convergence to a solution with physical sense, given that convergence occurs at all. Unfortunately, it can and will be shown that under no circumstances convergence can occur, if the reduction methods are applied as preconditioners.

A second focal point is therefore the application of the reduction methods to instationary problems. A given accuracy is to be reached with as many reduced time steps as possible and with only few detailed time steps in order to have the result be accurate enough. The technique of dual solutions leads to a strategy, which leads to a solution, which is by factors closer to the detailed than the reduced calculated solution. Fortunately, the same strategy can be used for an adaptive time stepping with almost no additional costs.

## Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit der Analyse und der Anwendung der Reduktionsmechanismen QSSA und ILDM, die beide im Rahmen der Modellierung von chemischen Verbrennungsprozessen entwickelt worden sind. Beide Mechanismen reduzieren die Anzahl der zu modellierenden Spezies erheblich, was eine bedeutende Kostenersparnis bei der numerischen Behandlung zur Folge hat. Der dadurch entstehende systematische Fehler bedingt, dass eine so erhaltene Lösung nicht notwendigerweise eine Lösung des detaillierten Problems ist. In einigen Anwendungen führt der reduzierte Lösungsweg daher zu physikalisch fragwürdigen Ergebnissen.

Ein Schwerpunkt dieser Arbeit ist daher die Untersuchung der Anwendbarkeit dieser Reduktionsmechanismen als Vorkonditionierer für Löser von stationären Problemen. Diese Art der Anwendung impliziert die Konvergenz der Berechnungen zu einem physikalisch sinnvollen Grenzwert, gegeben dass Konvergenz vorliegt. Unglücklicherweise kann gezeigt werden, dass Konvergenz mit den Reduktionsmechanismen als Vorkonditionierer unter keinen Umständen vorliegen kann.

Ein weiterer Untersuchungsschwerpunkt liegt in der adaptiven Anwendung der Reduktionsmechanismen bei instationären Problemen. Mit so vielen reduzierten Zeitschritten wie möglich und so wenigen detaillierten Schritten wie nötig soll eine zuvor vorgegebene Genauigkeit erreicht werden. Mit Hilfe der Technik der dualen Lösungen kann ein Verfahren entwickelt werden, das eine Lösung liefert, die um einen beliebig kleinen Faktor von der detailliert gerechneten Lösung abweicht. Das Verfahren ist so konstruiert, dass mit gleichem Rechenaufwand auch die Zeitschrittweite adaptiv gesteuert werden kann.

# Contents

# 1 Introduction

The simulation of phenomena involving chemical reactions, e.g. of chemical reactive flows, is still very time consuming due to the stiffness of the equations and the enormous number of degrees of freedom. They result from the large number of species, which are generally involved in the reaction and the large grids, which are to be applied for the calculation of PDEs in order to obtain decent numerical results. For advanced solving techniques like the multilevel technique, the computation time grows linearly with the number of grid nodes, but cubically with the number of species involved. Therefore reduction mechanisms have to be applied in order to reduce the number of calculated species and hence the computational costs.

This thesis will concentrate on two major reduction methods, namely the Quasi Steady State Assumption (QSSA) and the method of the Intrinsic Low Dimensional Manifolds (ILDM). The quasi steady state assumption takes advantage of the different timescales for the species. Species with fast timescales are assumed to relax quickly to the so called quasi steady state, whereas species with slow timescales drive the reaction. For an extensive discussion and the application of QSSA in combustion systems, see for example [30, 39, 46]. The technique of the ILDM is similar to the QSSA, but the idea is based on the differentiation of fast and slow processes in a chemical reaction system rather than on fast and slow species. This technique does not only reduce the number of equations, but also the stiffness. The technique of the ILDM was introduced by Maas and Pope in 1992 [28] and is widely used, see for example [8, 11, 34, 37, 44].

Both reduction methods QSSA and ILDM are introduced in detail in chapter 3, but the basic ideas are already presented here by considering the simple example

$$\begin{aligned} \dot{u}_1 &= -u_1 \\ \dot{u}_2 &= 1000(1/u_1 - u_2) \end{aligned}$$

with appropriate initial conditions. The solution to this system is given by

$$\begin{aligned} u_1(t) &= e^{-t} \\ u_2(t) &= \frac{1000}{1001} e^t + c e^{-1000t}, \end{aligned}$$

where the initial condition $u_1(0) = 1$ is assumed. The information of the initial value for $u_2$ is contained in the constant $c$. The solution shows clearly, that the manifold

$$u_2 = \frac{1000}{1001 u_1}$$

attracts the solution exponentially and independently of the initial value $u_2(0)$. Assume now that one is able to identify this manifold a priorily, then one finds itself in the advantageous situation that only the solution for $u_1$ is to be calculated and $u_2$ can be obtained from the function $\psi$. The so obtained solution is of course not exact, but the difference to the exact solution is only $\Delta u_2(t) = ce^{-1000t}$, which is close to zero even for small values of $t$ and can therefore be neglected in most cases.

The reduction methods QSSA and ILDM are concerned with the approximation of these manifolds. With the methods introduced in chapter 3, these approximations are calculated to be

$$\psi_Q(u_1) = \frac{1}{u_1} \quad \text{and} \quad \psi_I(u_1) = \frac{998}{999u_1}.$$

Clearly, the function $\psi_I$ obtained with ILDM is a better approximation to the exact manifold parameterized by $\psi$ than $\psi_Q$. The accuracy of the approximating manifolds depends strongly on the source term for the fast variable $u_2$. Replacing the factor 1000 by only 10 leads to the exact representation of the attracting manifold

$$u_2 = \frac{10}{11u_1},$$

whereas the manifolds calculated from QSSA and ILDM read

$$\psi_Q(u_1) = \frac{1}{u_1} \quad \text{and} \quad \psi_I(u_1) = \frac{8}{9u_1}.$$

Again, the manifold created with ILDM is of higher accuracy as the QSSA–manifold, but the performance is in both cases much worse than above, where the source term for the fast variable included the factor 1000. This indicates a relation between the accuracy of the obtained manifolds and the spectral gap of the gradient of the source term. The difference between the largest and smallest eigenvalue is now $\lambda_1 - \lambda_2 = 9$, whereas it is $\lambda_1 - \lambda_2 = 999$ in the above example.

Other reduction mechanisms are also possible to reduce the computational costs, but they are not treated in this thesis. Think for example of the theory of partial equilibria (e.g. [5, 30, 42, 43]), the method of computational singular perturbation [25, 26] or reduction methods based on optimization approaches [40]. All reduction strategies are in principal based on ideas of the Nobel price winner (1956) N.N. Semenov [45] and Bodenstein [7], who proposed the Bodenstein quasi stationarity.

All these reduction methods have in common that they reduce the computational effort for a chemical reaction system in the one or the other way. They all lead to an algebraic relation of the so called process variables to the fast variables. Therefore in dynamical systems for example, only the equations for the process variables are to be solved, the values for the fast variables can be obtained from the reduction method. This procedure introduces of course systematic errors, which leads directly to the goal of this thesis.

The aim of this thesis is to explore, how the reduction methods QSSA and ILDM can be applied, such that the computational costs are reduced without systematic errors. This will be done for various situations:

- QSSA and ILDM as preconditioners: Consider for example the preconditioned Richardson iteration for a system of linear equations

$$x^{n+1} = x^n + P(b - Ax^n) = x^n + h,$$

where $P$ is obtained by the application of QSSA and ILDM to the system

$$Ah = b - Ax^n.$$

The conjecture is that this preconditioner leads to exact results with small computational effort.

- The reaction mechanisms define a quasi–Newton method: For the given problem $f(u) = 0$, solve the linear equation

$$Df(u^n)\delta u = -f(u^n)$$

arising from the Newton method by the application of the manifolds obtained from QSSA or ILDM. This strategy will clearly converge to the exact solution, given that convergence occurs at all.

- Apply the reduction methods adaptively: Think of the ODE $\dot{u} = f(u)$, which is solved by an ODE solver, which leads clearly to errors. The application of QSSA and ILDM is therefore not necessarily required to lead to exact results. The methods may still be applied, if the occurring error is of the same order as the discretization error. Take for example the above ODE. Then the application of a detailed solver for the first time steps and using a reduced solver thereafter might easily lead to results with almost the same accuracy as a detailed solver for the whole interval.

The outline of this thesis is as follows: First a broad introduction to the equations modeling a chemical reaction is given in chapter 2. Then the reduction methods QSSA and ILDM will be explained in detail. The basic ideas and the theoretical background will be presented as well as a possible algorithm, which computes a single QSSA– or ILDM–point. A possibility, how the obtained results may be stored, is also shown.

Chapter 4 gives an overview to the techniques, which may be taken to solve equations involving chemistry. Special emphasis will be put to situations, where the reduction process does not lead to systematic errors. It can be shown that for relevant examples the application of the ILDM–method is exact.

The reader is then lead to the preconditioning of systems of linear equations. The main result of chapter 5 is that the preconditioners obtained from QSSA and ILDM prevent the iteration from convergence. An analytic result will show that convergence can practically never occur. The effort, which has to be undertaken in order to force convergence anyway, turns out to be as expensive as the detailed solution process.

Chapter 6 is then concerned with the quasi–Newton method, where the linear problem of the Newton iteration is solved with a reduced solver. The surprising result will be that the hereby reduced Newton iteration is only in very few examples comparable to the detailed Newton method, even if the considered problem contains only the chemical source term. Disturbed

problems are even more difficult to be solved, therefore convergence does in general not occur. Even though analytical results cannot be presented, the behavior of the reduced Newton method will be explained by an eigenvalue analysis.

Finally, the adaptive application of the reduction methods to ODEs will be investigated in chapter 7. The goal is to find criteria, which indicate, whether the solution process is to be switched from a reduced to the detailed strategy and vice versa. These criteria can be presented on a basis of dual solutions, which indicate the influence of the residual to the difference of the exact and calculated solution.

# 2 The Chemical Model

Historically, the ILDM–method was broad up for chemical reaction systems, especially, if combustion is involved. In this thesis, the application is therefore also from the field of chemical reactions and an overview over the generally used models and the underlying equations is given. They have the property that conservation laws, for example the mass conservation, are contained in the equations. A possibility will be presented, how these conservation laws can a priori be identified.

## 2.1 Modeling a homogeneous reactor

A homogeneous reactor is a model for a perfectly stirred reactor without any movement of the species, so effects like convection or diffusion are neglected. The modeling equations are therefore a system of $n+1$ ordinary differential equations, one equation describing the temperature and $n$ equations modeling the chemical species.

### 2.1.1 Modeling the chemical species

Let the species of the reactions be denoted by $A_s$, where $s$ is an index of the set $S$ of chemical species involved in the reaction system. The index set $R$ is supposed to contain all possible elementary reactions in the reactor.

The reaction system can then be formulated by the reactions

$$\sum_{s\in S} \alpha_{r,s} A_s \xrightarrow{k_r} \sum_{s\in S} \beta_{r,s} A_s \qquad \forall\, r \in R \tag{2.1}$$

with the reaction coefficients $\alpha_{r,i}$ and $\beta_{r,i}$. The factor $k_r$ denotes the reaction rate. Consider for example the oxidation of hydrogen atoms

$$H + O_2 \longrightarrow HO + O, \tag{2.2}$$

then

$$\begin{aligned} \alpha_H &= 1 & \beta_H &= 0 \\ \alpha_{O_2} &= 1 & \beta_{O_2} &= 0 \\ \alpha_{HO} &= 0 & \beta_{HO} &= 1 \\ \alpha_O &= 0 & \beta_O &= 1. \end{aligned}$$

The mathematical formulation of the reaction system (2.1) can be described by an ODE in terms of concentrations $c_i$ of species $A_i$ (compare [14, 6.3]):

$$\frac{dc_i}{dt} = \omega_i := \sum_{r \in R} (\beta_{r,i} - \alpha_{r,i}) k_r \prod_{s \in S} c_s^{\alpha_{r,s}}. \tag{2.3}$$

The reaction rates $k_r$ are highly temperature dependent and are generally modeled by the Arrhenius law [1]

$$k_r = A T^b e^{-\frac{E}{RT}}.$$

The coefficients $A$, $b$ and $E$ are obtained from experiments for every reaction, $R$ denotes the gas constant $R = 8.3145\, J/(mol \cdot K)$, see [31].

Other formulations to model the chemical reactions are also possible, take for example the formulation in mass fractions $y_i$ with $c_i = \frac{\rho y_i}{m_i}$, which is also widely used:

$$\frac{d(\rho y_i)}{dt} = m_i \omega_i.$$

Here, $\rho$ denotes the density of the mixture and $m_i$ the specific molar mass of species $A_i$.

### 2.1.2 A simplification

The formulation of the reactions in terms of mass fractions has the advantage that the physically side condition

$$\sum_i y_i = 1$$

can be used to simplify the ODEs to

$$\rho \frac{dy_i}{dt} = m_i \omega_i, \tag{2.4}$$

because $\rho$ is constant in time: In the following section it will be shown that the mass conservation implies $\sum_i m_i \omega_i = 0$, therefore $\sum_i \rho y_i = \rho \sum_i y_i = \rho$ must be constant in time.

### 2.1.3 Modeling the temperature

The equation for the temperature is a direct consequence of the conservation laws for the energy. The derivation can be found in [14, 3.2, 3.3] and leads to

$$\rho c_p \frac{dT}{dt} = -\sum_{i \in S} h_i m_i \omega_i.$$

The specific heat capacity $c_p$ of the mixture can be calculated by $c_p = \sum_i y_i c_{p,i}$ with the heat capacities $c_{p,i}$ of species $A_i$. The heat capacities $c_{p,i}$ of the species are usually calculated by polynomial fits to experimental data, the specific enthalpies $h_{p,i}$ are obtained via the relation

$$h_{p,i} = \frac{\partial c_{p,i}}{\partial T}.$$

The dependent variable for the polynomial fit is the temperature $T$.

### 2.1.4 The governing equations

Let $u \in \mathbb{R}^{n+1}$ denote the state vector of a chemical reactor, where $u_0$ denotes the temperature $T$ and $u_i$ describes the mass fraction $y_i$ of species $A_i$ for all $i = 1, ..., n$. Then the homogeneous reactor can be modeled by the system of ordinary differential equations

$$\frac{du}{dt} = f(u), \quad u(0) = u^0,$$

where $f : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ is given by

$$f(u) = \begin{pmatrix} -\frac{1}{c_p} \sum_{i \in S} h_i m_i \omega_i \\ m_1 \omega_1 \\ \vdots \\ m_n \omega_n \end{pmatrix} \tag{2.5}$$

## 2.2 Conservation laws

The modeling equations (2.4) contain several conservation laws, which are not formulated explicitely. As already mentioned in the previous chapter, we have the mass conservation law. A chemical reaction can neither create nor destroy mass, therefore the total mass of the products equals the total mass of the educts:

$$\sum_{i \in S} \alpha_{r,i} m_i = \sum_{i \in S} \beta_{r,i} m_i, \tag{2.6}$$

which leads with (2.3) directly to

$$\sum_{i \in S} m_i \omega_i = 0$$

and with (2.4) to

$$\frac{d}{dt} \sum_i \rho y_i = 0.$$

This implies with $\sum_i y_i = 1$ directly the constance of the density $\rho$ in time.

The chemical reactions do not only preserve mass, also the number of the elements is conserved in every reaction. Let $E_j$ denote the elements, which are involved in the species of the chemical reactor, and let $n_e$ be the total number of elements. Furthermore, let the coefficient $\mu_{i,j}$ denote the number of elements $E_j$ in species $A_i$. In water for example, these coefficients are clearly $\mu_{H,H_2O} = 2$ and $\mu_{O,H_2O} = 1$. Then every species $A_i$ can formally be written in the form

$$A_i = \sum_{j=1}^{n_e} \mu_{i,j} E_j \tag{2.7}$$

and the chemical reaction (2.1) reads

$$\sum_{s \in S} \alpha_{r,s} \sum_{j=1}^{n_e} \mu_{s,j} E_j \xrightarrow{k_r} \sum_{s \in S} \beta_{r,s} \sum_{j=1}^{n_e} \mu_{s,j} E_j.$$

The mathematical formulation of the element conservation in terms of reaction rates is

$$\sum_{s\in S} \alpha_{r,s}\mu_{s,j} = \sum_{s\in S} \beta_{r,s}\mu_{s,j} \quad \forall j \in \{1,...,n_e\}. \tag{2.8}$$

This equation means that the number of elements on the educt side equals the number of elements on the product side for every element in every reaction. The element conservation can be used to calculate the left eigenvectors to the eigenvalue $\lambda = 0$ of the Jacobian $\nabla f$ of the source term $f$ given by (2.5). This result will later be used to formulate the so called center manifolds.

**Lemma 2.2.1** *Let $f : \mathbb{R}^{n+1} \to \mathbb{R}^{n+1}$ denote the source term of a chemical reaction (2.5) with $n_e$ elements involved in the species of the mixture. Then the Jacobian $\nabla f$ has an eigenvalue $\lambda = 0$ with the corresponding left eigenspace of dimension $n_c \geq n_e$.*

*Moreover, an $n_e$–dimensional subspace of the left eigenspace is spanned by the vectors*

$$v_i = \left(0, \quad \tfrac{\mu_{1,i}}{m_1}, \quad ..., \quad \tfrac{\mu_{n,i}}{m_n}\right), \quad i = 1, ..., n_e \tag{2.9}$$

*where $\mu_{k,i}$ denotes the element composition as in (2.7).*

**Proof:** The scalar product of $v_i$ and the chemical source term $f(u)$ gives

$$\begin{aligned} v_i \cdot f(u) &= 0 + \sum_{s=1}^{n} \frac{\mu_{s,i}}{m_s} m_s \omega_s \\ &= \sum_{s=1}^{n} \mu_{s,i} \sum_{r\in R} (\beta_{r,s} - \alpha_{r,s}) k_r \prod_{i\in S} c_i^{\alpha_{r,i}} \\ &= \sum_{r\in R} \underbrace{\sum_{s=1}^{n} \mu_{s,i}(\beta_{r,s} - \alpha_{r,s})}_{=0 \text{ (with (2.8))}} k_r \prod_{i\in S} c_i^{\alpha_{r,i}}. \end{aligned}$$

Since the reaction coefficients $\alpha$ and $\beta$ are independent of the state vector $u$, the same calculation is valid, if $f(u)$ is replaced by a column vector of $\nabla f(u)$, therefore

$$v_i \cdot Df(u) = 0,$$

which finishes the proof. ■

**Remark:** In general, the dimension of the left eigenspace equals exactly the number of elements forming the species. The vector $v = \begin{pmatrix} 0 & 1 & ... & 1 \end{pmatrix}$, which describes the conservation of mass, is obviously also a left eigenvector to the eigenvalue 0. It is a linear combination of the vectors $v_i$ denoted by (2.9). In order to prove this remark, let the specific mass of element $E_i$ be denoted by $m_{E_i}$. Then

$$\begin{aligned} \sum_{i=1}^{n_e} m_{E_i} v_i &= \sum_{i=1}^{n_e} \left(0, \quad \tfrac{m_{E_i}\mu_{1,i}}{m_1}, \quad ..., \quad \tfrac{m_{E_i}\mu_{n,i}}{m_n}\right) \\ &= \left(0, \quad \tfrac{\sum_{i=1}^{n_e} m_{E_i}\mu_{1,i}}{m_1}, \quad ..., \quad \tfrac{\sum_{i=1}^{n_e} m_{E_i}\mu_{n,i}}{m_n}\right) \\ &= \begin{pmatrix} 0, & 1, & ..., & 1 \end{pmatrix}. \end{aligned}$$

This proof shows, that the conservation of the elements is only a more detailed description of the mass conservation.

## 2.3 Example: Ozone Reaction

The ozone reaction is one of the easiest realistic examples, because only three species are involved, namely $O$, $O_2$ and $O_3$. Moreover, these species consist of only one element, which makes it easy to identify the conservation laws.

The ozone mechanism can be formulated by

$$M + 2O \underset{k_2}{\overset{k_1}{\rightleftharpoons}} M + O_2$$
$$M + O_2 + O \underset{k_4}{\overset{k_3}{\rightleftharpoons}} M + O_3$$
$$O + O_3 \underset{k_6}{\overset{k_5}{\rightleftharpoons}} 2O_2$$

with the reaction rates $k_i$ calculated from the Arrhenius coefficients in table 2.1. Other reaction formulations containing ozone can be found in [36].

| | $A\ [cm/(mol \cdot s)]$ | $b$ | $E\ [kJ/mol]$ |
|---|---|---|---|
| $k_1$ | $2.9 \cdot 10^{17}$ | $-1$ | 0 |
| $k_2$ | $6.772 \cdot 10^{18}$ | $-1$ | 496 |
| $k_3$ | $3.426 \cdot 10^{13}$ | 0 | $-4.234$ |
| $k_4$ | $9.5 \cdot 10^{14}$ | 0 | 95.03 |
| $k_5$ | $5.2 \cdot 10^{12}$ | 0 | 17.38 |
| $k_6$ | $4.381 \cdot 10^{12}$ | 0 | 414.39 |

**Table 2.1:** The Arrhenius coefficients for the reaction rates of the ozone reaction.

The $M$ inside the above reaction mechanism is a symbol for every species in the mixture and is used as a third body. The first reaction for example is in fact a system of reactions describing the following situation:

$$O + 2O \underset{k_{1,2}}{\overset{k_{1,1}}{\rightleftharpoons}} O + O_2$$
$$O_2 + 2O \underset{k_{1,4}}{\overset{k_{1,3}}{\rightleftharpoons}} O_2 + O_2$$
$$O_3 + 2O \underset{k_{1,6}}{\overset{k_{1,5}}{\rightleftharpoons}} O_3 + O_2.$$

The reaction rates $k_{1,i}$ herein differ only slightly from $k_1$. It is assumed that the Arrhenius coefficients $b$ and $E$ remain the same as in table 2.1, whereas $A$ may be changed by a constant factor, the so called third body efficiency.

Figure 2.1 shows the behavior of the ozone reaction in a homogeneous reactor. As initial conditions, the temperature is set to $T = 700K$, the mass fractions for $O$, $O_2$ and $O_3$ to 0.0,

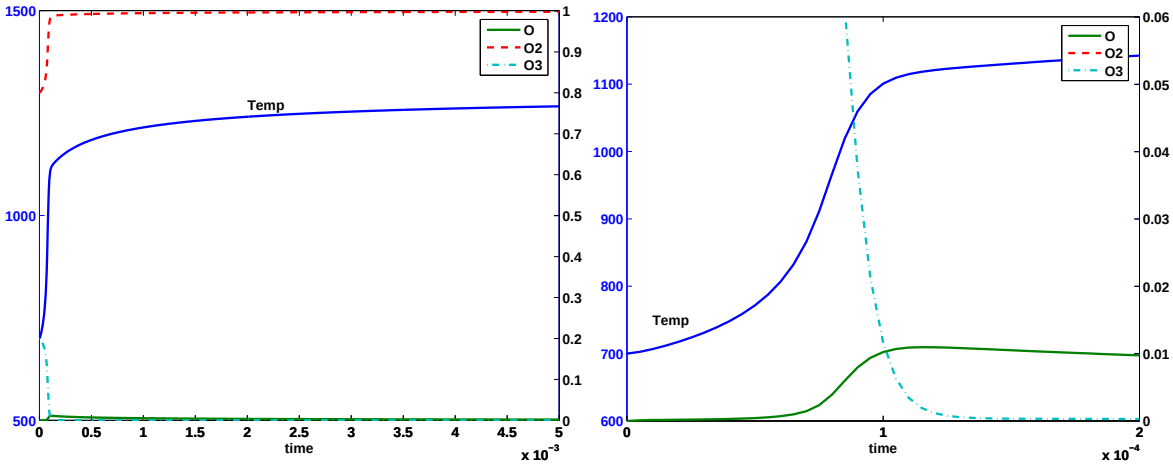


**Figure 2.1:** The ozone reaction in a homogeneous reactor. The right figure focuses on the ignition of the reaction. The atomic oxygen $O$ appears with very small mass fractions at the same time, when the rate of change for the temperature is maximal. The existence of $O$ can therefore be used as an indicator of the ignition time.

0.8 and 0.2 respectively. Obviously $O$ is only an intermediate species, a so called radical. It is generated only to a very small amount, if the temperature increases with maximal slope. The species $O$ can therefore be used to identify the flame front of the ozone combustion.

# 3 Introduction to QSSA and ILDM

The reduction methods QSSA (Quasi Steady State Assumption) and ILDM (Intrinsic Low Dimensional Manifold) are based on the idea that the state of a chemical reaction can basically be described by the mass fraction of only very few variables, the so–called process variables or slow variables. The mass fraction of the fast variables can then be obtained by an algebraic relation independently of effects like convection or diffusion. This can be interpreted to that effect that the fast species react, before they change their position due to diffusion or similar effects.

This chapter will therefore consider the theory of these reduction methods and explain the technical details. It will start with a small explanation of problems, which are typically reduced with QSSA or ILDM, namely the spectral gap of the source term's gradient. Then techniques are presented, how the value $\psi(u_1)$ can be computed and the difficulties are discussed. Also center manifolds are introduced as well as aspects on tabulation strategies.

## 3.1 Properties of stiff differential equations

Stiff systems of ordinary differential equations

$$u'(t) = f(u)$$

are characterized by the fact that the gradient of the source term $\nabla f(u)$ has eigenvalues with big differences in their real part. This means that very different time scales exist. Due to this phenomenon it is difficult to solve stiff ODEs numerically and the solution process with explicit solvers produces a lot of overhead, because the time step $k$ has to be small enough to get reasonable results for all variables.

But an advantage can be taken out of the stiffness, if there is a big gap in the spectrum, which means that disturbances of some variables relax much faster then disturbances of others. Then the the state vector $u$ can be split into so called fast and slow variables $u_2$ and $u_1$. Assume further that an algebraic relation can be found a priori, which maps $u_1$ to $u_2$ by the function

$$\psi : \mathbb{R}^{n_1} \to \mathbb{R}^{n_2} : u_1 \mapsto u_2. \tag{3.1}$$

Here $n_1$ is the number of slow variables combined in the vector $u_1$, whereas the fast variables in the vector $u_2$ have magnitude $n_2$.

Take for example the system of ODEs

$$u'(t) = \begin{pmatrix} -1 & 1 \\ 1 & -M \end{pmatrix} u(t) \tag{3.2}$$

with $M > 0$. This system has a steady state

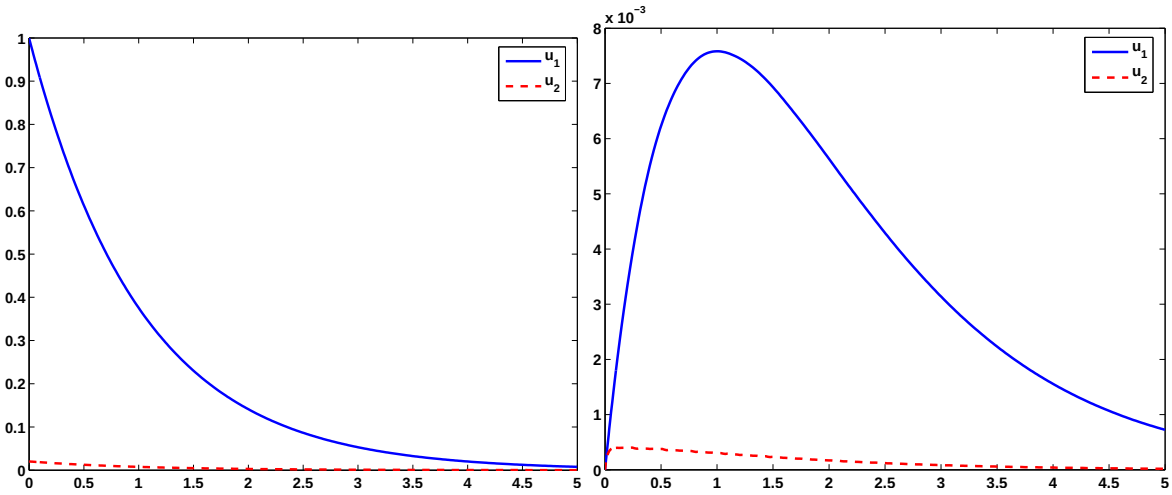


**Figure 3.1:** Left: Solution to the ODE (3.2) with $M = 50$ and initial conditions $u_0 = (1, 1/50)$. Right: The error produced by applying the reduction (3.3)

$$u_\infty = 0$$

and $u$ converges to this steady state for all initial values, if $M > 0$. An easy calculation shows that $u_2$ relaxes to half of its initial value already after short time, if $M$ is large, whereas it takes much longer for the variable $u_1$. Therefore it seems to be reasonable to assume that $u_2$ does not change much for small values of $t$ for constant $u_1$. The algebraic relation

$$u_2(t) = \psi(u_1(t))$$

transforms the original ODE to the DAE

$$\begin{aligned} u_1'(t) &= -u_1(t) + \psi(u_1(t)) \\ u_2(t) &= \psi(u_1(t)). \end{aligned}$$

Here we take

$$\psi(u_1) = \frac{1}{M} u_1 \tag{3.3}$$

following the idea of the quasi steady state assumption QSSA, which will be introduced in the following section. Now the calculation effort to obtain a solution decreases, because the number of equations is reduced and the time steps can be increased due to loss of stiffness.

Obviously the reduction causes a loss of accuracy, as can be seen in figure 3.1. The accuracy of the reduced problem depends on the spectral gap of the gradient $\nabla f(u)$. The bigger the spectral gap, i.e. the difference between the real parts of the eigenvalues, the better the approximation of the reduced to the detailed solution. In problem (3.2) the spectral gap

$$\sigma = \sqrt{M^2 - 2M + 5}$$

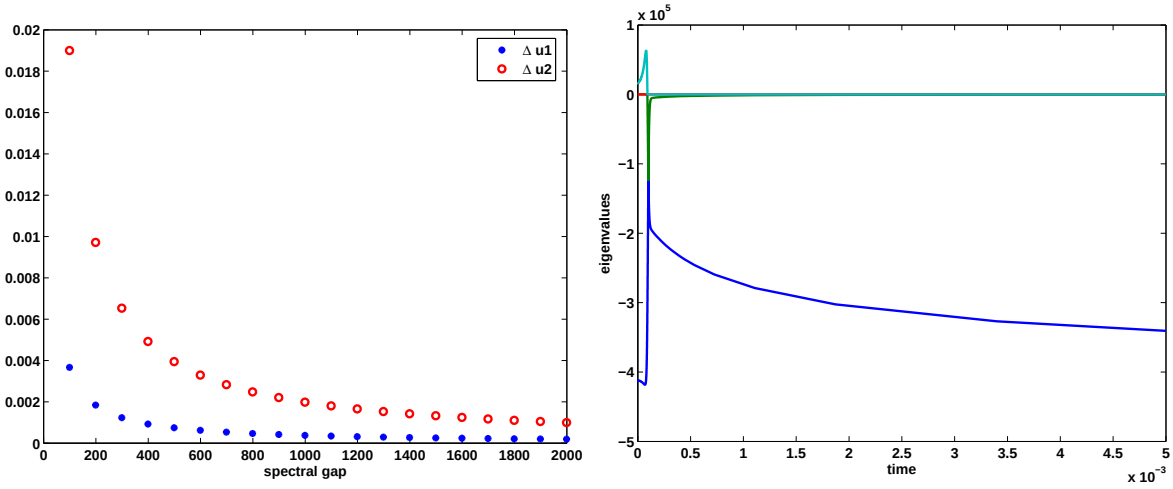


**Figure 3.2:** Left: The maximal difference between the reduced and the detailed solution of problem (3.2) on the interval $[0, 5]$ plotted against the spectral gap depending on $M$. Right: The evolution of the eigenvalues of the gradient of the source term of the ozone reaction. The evaluation point is the currently reached state in an homogeneous reactor, cp. figure 2.1. Clearly, one eigenvalue has large negative values.

depends almost linearly on $M$ and figure 3.2 shows the increasing accuracy with increasing spectral gap.

In chemical reaction systems the spectral gap can be reasonably big, if combustion is involved. In general, most eigenvalues have large negative real parts, whereas the real part of only few eigenvalues is around $-1$. This suggests the existence of an algebraic relation between the slow and fast variables (3.1).

For example, the ozone reaction described in section 2.3 can be modeled by 4 variables, namely the temperature and the mass fractions of $O$, $O_2$ and $O_3$. An eigenvalue analysis of the gradient of the source term (cp. figure 3.2) shows that one eigenvalue has a large negative value, whereas three eigenvalues are reasonably bigger for almost all times $t$. In addition, one eigenvalue is exactly zero (up to rounding errors) and highlights the existence of a conservation law. Therefore only two variables have to be calculated in an ODE, the remaining two variables can be calculated algebraically by applying the function $\psi$ introduced by (3.1), which shrinks the number of variables for the ODE by the factor $1/2$.

This factor decreases even further, when more advanced reaction systems are treated. In the methane–air combustion for example, more than 30 species depend algebraically on only four variables. This reduces the number of unknowns for the ODE by the factor $1/8$.

## 3.2 QSSA and ILDM

In this thesis two different methods to obtain the algebraic relation $u_2 = \psi(u_1)$ are considered, namely the Quasi Steady State Assumption (QSSA) and the Intrinsic Low Dimensional Manifolds (ILDM). In both cases the function $\psi$ is given implicitly by an equation of the form

$$F_2(u_1, \psi(u_1)) = 0,$$

where the definition of $F_2$ differs in the two methods. Whereas $F_2$ depends for QSSA directly on the fast part of the source term $f_2$, the eigenvectors of $\nabla f$ are included in $F_2$ in case of ILDM.

### 3.2.1 QSSA

By applying the QSSA to a system of ordinary differential equations, one distinguishes fast and slow variables by the magnitude of their source terms. Variables with big source terms are called fast variables and have the property that they relax quickly to a steady state, when the slow variables are kept constant. In this steady state for the fast variables (the quasi steady state for the system of ODEs) the time derivatives for these variables are zero:

$$u_2'(t) = 0.$$

This equality is now assumed to hold for every state vector $u(t) = (u_1(t), u_2(t))$, because the difference in terms of time between a given state and its corresponding quasi steady state is supposed to be very small.

The idea of QSSA transforms therefore the original ODE

$$\begin{aligned} u_1'(t) &= f_1(u_1, u_2) \\ u_2'(t) &= f_2(u_1, u_2) \end{aligned}$$

to the DAE

$$\begin{aligned} u_1'(t) &= f_1(u_1, u_2) \\ 0 &= f_2(u_1, u_2), \end{aligned}$$

where the algebraic equation $f_2(u_1, u_2) = 0$ describes a manifold, which can be parameterized by (3.1). The function $\psi$ is therefore given implicitly by the function $F_2 = f_2$.

There are basically two possibilities, how the variables can be characterized with respect to "slow" and "fast". The first one is based on trial and error, where a decent knowledge of the equations and the underlying mechanisms is very much advantageous. Since this knowledge can in general not be assumed, a mathematical method is needed. A successful method is an eigenvalue analysis of the gradient of the source term. The fast variables can then be characterized by the dominating entries in the eigenvectors of the fast eigenvalues.

The major drawback of this assumption is the fact that in almost no system of ODEs, "slow" and "fast" variables can be sharply distinguished. In general, all variables are involved also in

fast and in slow processes, therefore by using QSSA the size of the problem can be reduced, but the reduced system might still be stiff. The advantage of QSSA is the relative low expenses to obtain a QSSA–point $\psi(u_1)$.

### 3.2.2 ILDM

In contrast to the QSSA–assumption one assumes by applying ILDM that the system of ODEs describes fast and slow processes instead of fast and slow variables. In order to distinguish these processes, the system is transformed to the basis of the eigenvectors of the gradient $\nabla f(u)$ of the source term $f$.

For a brief description let the function $f$ be linear and be defined by the matrix $A$ with $f(u) = Au$. Let further $\Lambda$ and $V$ be the eigenvalue analysis of $A$ with

$$AV = V\Lambda$$

assuming the diagonalisability of $A$. Then the original ODE

$$\dot{u} = Au$$

can be transformed to

$$V^{-1}\dot{u} = V^{-1}Au$$

which can be simplified to

$$\dot{y} = V^{-1}AVy = \Lambda y$$

with $y = V^{-1}u$.

Since $\Lambda$ is diagonal, the last equation describes a system of totally decoupled ODEs, where the fast variables can be sharply distinguished from the slow variables in terms of $y$. This means that if an algebraic relation between the fast and slow variables can be found, the reduced system for the slow variables $y_1$ is not stiff anymore.

The algebraic relation between $y_1$ and $y_2$ can be found by

$$\left[\Lambda \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}\right]_2 = 0,$$

which reads

$$\left[V^{-1}A \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}\right]_2 = 0$$

in terms of $u$.

The nonlinear equivalence to this formulation leads to the general ILDM–condition

$$\left[V^{-1}f(u)\right]_2 = 0, \tag{3.4}$$

where $V$ and $\Lambda$ are given by the eigenvalue analysis of the source term's gradient:

$$\nabla f(u)V(u) = V(u)\Lambda(u).$$

Here, the difference between slow and fast variables is not as important as in the case of QSSA. The only requirement on the process variables $u_1$ is the uniqueness of the calculation of the variables $u_2$, and therefore the injectivity of the function $\psi$. This is in general given, if the variables are distinguished in the same way as in the case of QSSA. This has the advantage, that the comparison of these can easily be done. Further, it shows that QSSA and ILDM are equivalent, if the slow eigenspace is orthogonal to the fast eigenspace.

Compared to the QSSA, the method of ILDM has the big advantage to reduce the stiffness of the resulting equations even further. The price for this advantage is the calculation of the left eigenvectors $V(u)^{-1}$ of the gradient $\nabla f$ of the source term, which is numerically instable and expensive.

These instabilities can be circumvented by considering the Schur factorization $\nabla f = QRQ^T$ instead of the eigenvalue analysis $\nabla f = V\Lambda V^{-1}$. The Schur factorization is numerically stable and leads to the same results in the calculation of the ILDM–points $\psi(u_1)$. In order to see this equivalence, we first need a method to obtain the eigenvalue analysis of a matrix $A = V\Lambda V^{-1}$ assuming that the Schur factorization $A = QRQ^T$ is already calculated:

**Lemma 3.2.1** *Let $R \in \mathbb{R}^{n\times n}$ be of upper triangular form. Then the eigenvalue analysis leads to $R = W\Lambda W^{-1}$ with*

$$W = \begin{pmatrix} w_{11} & w_{12} & \dots & w_{1n} \\ 0 & w_{22} & \dots & w_{2n} \\ \vdots & & \ddots & \vdots \\ 0 & \dots & 0 & w_{nn} \end{pmatrix} \quad \textit{and} \quad \Lambda = \begin{pmatrix} r_{11} & 0 & \dots & 0 \\ 0 & r_{22} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \dots & 0 & r_{nn} \end{pmatrix}$$

**Proof:** Let $W^{-1}RW = \Lambda$ be diagonal. Then the last row of $RW = W\Lambda$ reads

$$r_{nn}(w_{n1}, ..., w_{nn}) = (r_{11}w_{n1}, ..., r_{nn}w_{nn}),$$

therefore

$$w_{nk} = 0 \quad \forall k < n.$$

The equality $w_{ij} = 0$ for $i < j < n$ is then shown by induction. ■

This Lemma shows that for the given Schur factorization of a matrix $A = QRQ^T$ there exist a diagonal matrix $\Lambda$ and an upper triangular matrix $W$ such that $A = (QW)\Lambda(QW)^{-1}$ is an eigenvalue analysis of $A$. Moreover, $W$ and $\Lambda$ are the result of an eigenvalue analysis of $R$. This phenomenon can directly be used to obtain the above mentioned equivalence.

**Lemma 3.2.2** *Let $f : \mathbb{R}^n \to \mathbb{R}^n$ be continuously differentiable such that*

$$\nabla f(u) = Q(u)R(u)Q(u)^T = V(u)\Lambda(u)V(u)^{-1}$$

*is the Schur and the eigenvalue transformation respectively with $\Lambda(u)$ being diagonal for all $u$. Let further $P_2$ be in $\mathbb{R}^{n_2\times n}$ the trivial projection from the space of the physical variables onto the space of the fast variables. Then the problem*

$$P_2V(u)^{-1}f(u) = 0$$

*is equivalent to*

$$P_2 Q(u)^T f(u) = 0.$$

**Proof:** For simplicity assume $P_2 = \begin{pmatrix} 0 & \mathbb{1} \end{pmatrix}$. Otherwise reorder the eigenvalues in $R$ and $\Lambda$. Let $W$ and $R$ be in $\mathbb{R}^{n \times n}$ such that

$$R = W \Lambda W^{-1}.$$

Lemma 3.2.1 shows that $W^{-1}$ is of upper triangular form and can be written as

$$W^{-1} = \begin{pmatrix} W_{11} & W_{12} \\ 0 & W_{22} \end{pmatrix}$$

with $W_{22} \in \mathbb{R}^{n_2 \times n_2}$ invertible. Then

$$\begin{aligned} [V^{-1} f]_2 = P_2 V^{-1} f &= P_2 W^{-1} Q^T f \\ &= W_{22} P_2 Q^T f = W_{22} [Q^T f]_2. \end{aligned}$$

Since $W_{22}$ is invertible, the proof is finished. ■

Lemma 3.2.2 shows that the ILDM formulation (3.4) can be reduced to

$$\left[ Q(u_1, u_2)^T f(u_1, u_2) \right]_2 = 0 \tag{3.5}$$

and the instable calculation of the eigenvectors is replaced by the numerically more stable Schur factorization.

The function $F_2$, which defines the manifold implicitly, is therefore given by

$$F_2(u_1, u_2) = [Q^T(u_1, u_2) f(u_1, u_2)]_2$$

and calculates the fast variables $u_2$ from the process variables $u_1$ by

$$F_2(u_1, u_2) = 0$$

for given values of $u_1$. Note that even though ILDM distinguishes fast and slow processes, the result is still a relation from process variables to fast variables.

The major advantage of the ILDM is that the fast and slow processes can be distinguished sharply. This leads to a reduction, where not only the size of the problem, but also the stiffness is reduced. The price for this advantage is the treatment of the Schur factorization. Its calculation is numerically expensive and makes the function $F_2$ loose its continuity.

### 3.2.3 Numerical examples

We apply QSSA and ILDM to the example (3.2). In order to obtain the function $\psi$ by QSSA,

$$u_1 - M u_2 = 0$$

has to be solved, hence

$$\psi_{\mathrm{QSSA}}(u_1) = {}^{u_1}/_{M}.$$

Consider now the approximation of the value $u_2$ by ILDM. The fast eigenvalue of $A$ is $\lambda_2 = -(M+1+\sqrt{M^2-2M+5})/2$ with the corresponding left eigenvector given by $v_2 = \left(-2, \quad M-1+\sqrt{M^2-2M+5}\right)$. The equation to be solved is therefore

$$[v_2 Au]_2 = (M+1+\sqrt{M^2-2M+5})u_1 - 2(M^2-M+2+M\sqrt{M^2-2M+5})u_2 = 0,$$

which leads to the solution

$$\psi_{\mathrm{ILDM}}(u_1) = \frac{1}{M}u_1 + O(M^{-2}) = \psi_{\mathrm{QSSA}}(u_1) + O(M^{-2}).$$

This shows that the difference between QSSA and ILDM is reasonably small and decreases with increasing spectral gap.

A more interesting example consists of a mixture containing the three species $A$, $B$ and $C$. The nonlinear reaction

$$2A \underset{1}{\overset{2}{\rightleftharpoons}} 2B \underset{90}{\overset{1000}{\rightleftharpoons}} 2C \tag{3.6}$$

which is modeled by the ODE

$$\dot{y} = \begin{pmatrix} -4y_A^2 + 2y_B^2 \\ 4y_A^2 - 2002y_B^2 + 180y_C^2 \\ 2000y_B^2 - 180u_C^2 \end{pmatrix}, \qquad y(0) = y_0,$$

assuming that the species $A$, $B$ and $C$ have the same specific molar mass.

Obviously $B$ is only an intermediate product, which reacts to C almost instantaneously. Therefore the mass fraction $y_B$ of species $B$ is taken as a fast variable and will be approximated by the function $\psi$, which has to fulfill the equation

$$4y_A^2 - 2002\psi(y_A, y_C)^2 + 180y_C^2 = 0, \tag{3.7}$$

when QSSA as reduction mechanism is applied. Figure 3.3 shows the approximating function $\psi_{QSSA}$ and the relative small difference of the approximations with QSSA and ILDM.

Even though the difference of the values $\psi_Q$ and $\psi_I$ is rather small, the difference in an application is surprisingly big. The reduction with ILDM turns out to perform much better than problems reduced with the QSSA–formalism. Figure 3.4 shows the difference of ILDM and QSSA when applied as a reduction to the ODE modeling the reaction system (3.6). Clearly, the errors are very big even in the picture norm. The perfomance will be improved by considering the center manifolds in the following sections. But still, the figure depicts clearly that ILDM leads to better approximations than QSSA.

### 3.2.4 Problems

The manifold described by QSSA is much easier to be calculated than the values of $\psi_{ILDM}$, but still difficulties arise:

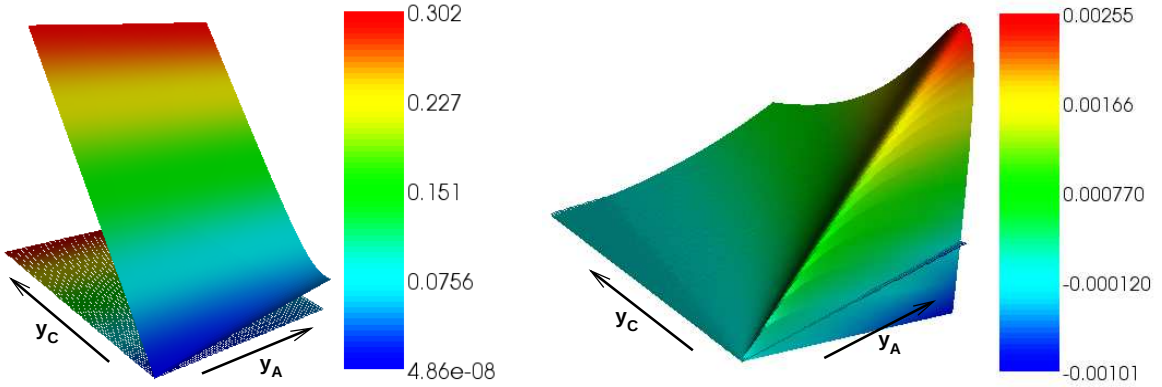


**Figure 3.3:** Left: The value $\psi_I(y_A, y_C)$ obtained by the ILDM for reaction (3.6). Right: The difference $\psi_I - \psi_Q$. In both figures, the $x$–axis shows the mass fraction of $A$, the $y$–axis of $B$ in the range $[0, 1]$. Note that the figure is purely mathematical, the condition $\sum_i y_i = 1$ is neglected and will be taken care of in the upcoming section.

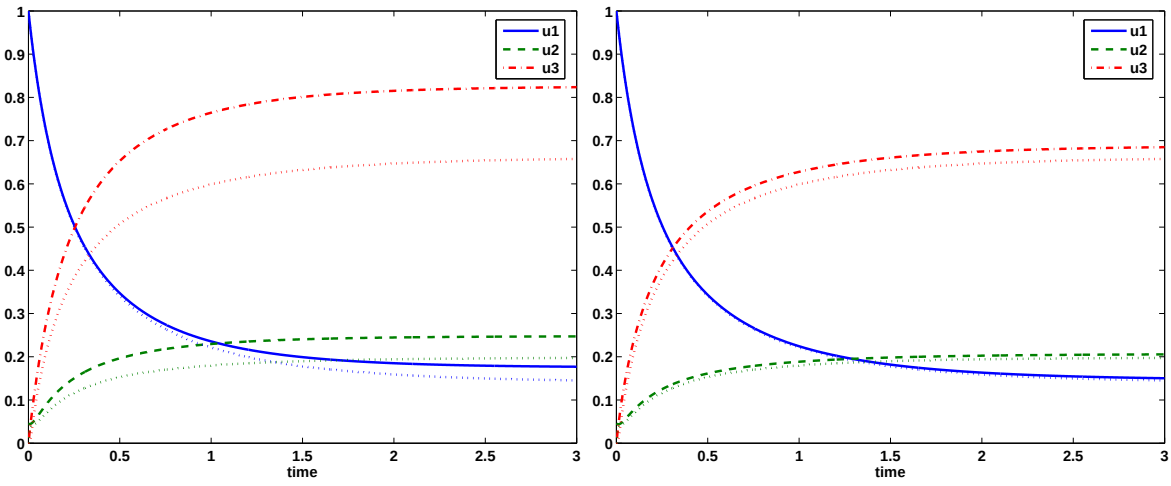


**Figure 3.4:** The reaction mechanism (3.6) in a homogeneous reactor with initial conditions $y_A(0) = 1$, $y_B(0) = y_C(0) = 0$. The dotted lines represent the detailed solution, the full lines are the solution to the reduced DAE. In the left figure the manifold obtained from QSSA is used, in the right figure the ILDM–manifold.

- Even if there is a $u_1$ and a $u_2$ such that $f(u_1, u_2) = 0$, there is no guarantee that there is a $u_2$ for any given $u_1$, such that $f_2(u_1, u_2) = 0$. And if there is a solution, then it is not necessarily unique. Take for example the function

$$f(u_1, u_2) = \begin{pmatrix} u_1 \\ -u_1 - u_2^2 \end{pmatrix}.$$

  Then $f_2(u_1, u_2) = 0$ for $u_1 = u_2 = 0$, but there is no $u_2$ to solve $f_2(1, u_2) = 0$, whereas the solution to $f_2(-1, u_2) = 0$ is not unique ($u_2 = \pm 1$).

- The equation $f_2(u_1, u_2) = 0$ is in general highly nonlinear and stiff, therefore a good starting point for the Newton method is required. One acceptable possibility to circumvent this problem is doing a few time steps for the problem

$$u_2'(t) = f_2(u_1, u_2), \quad u_2(0) = u_2^0. \tag{3.8}$$

  The starting point $u_2^0$ might for example be the second part of the steady state $f(u_1^\infty, u_2^\infty) = 0$ or an already calculated solution $u_2^0 = \psi(v_1)$, where $|v_1 - u_1|$ is small.

  Doing a few time steps has the big advantage that the solution $\widehat{u}_2$ of $f_2(u_1, u_2) = 0$ is stable from an ODE point of view, i.e. the solution to problem (3.8) with initial condition $u_2^0 = \widehat{u}_2 + \varepsilon$ relaxes again to $\widehat{u}_2$ for small $\varepsilon$.

- The physical conservation laws for the state $u = (u_1, \psi(u_1))$ should not be violated. When chemical reaction systems are treated, this means, that for example the mass fractions of the species should sum up to one. Also the number of atoms of the elements cannot change in a chemical reactor. These effects are summarized in the so called center manifold, which is treated separately.

  Figure 3.4 shows the effect, when the center manifold is neglected in the calculation of $\psi$. Whereas the detailed solution is on the center manifold (it solves $u_1(t) + u_2(t) + u_3(t) = 0$ for all $t \geq 0$), both the solutions obtained from QSSA and ILDM are not.

- There might be side conditions for the state vector $u$, which should not be violated by the new state vector $(u_1, \psi(u_1))$. In chemical reaction systems for example, the mass fractions $y_i$ must fulfill the inequality $0 \leq y_i \leq 1$. In general, this is not the case for the state vector $(u_1, \psi(u_1))$ and has to be forced heuristically.

  As an example, reconsider the nonlinear reaction (3.6) with QSSA and conservation laws. The function $\psi$ is then a function, which maps $y_A$ to $(y_B, y_C)$ and has to solve the equations

$$\begin{aligned} 4y_A^2 - 2002\psi_1(y_A)^2 + 180\psi_2(y_A)^2 &= 0, \\ y_A + \psi_1(y_A) + \psi_2(y_A) &= 1. \end{aligned} \tag{3.9}$$

  The function $\psi$ has then the solution shown in figure 3.5, where the function values have to be adapted manually for large $y_A$ in order to assure the physically necessary condition $\psi > 0$.

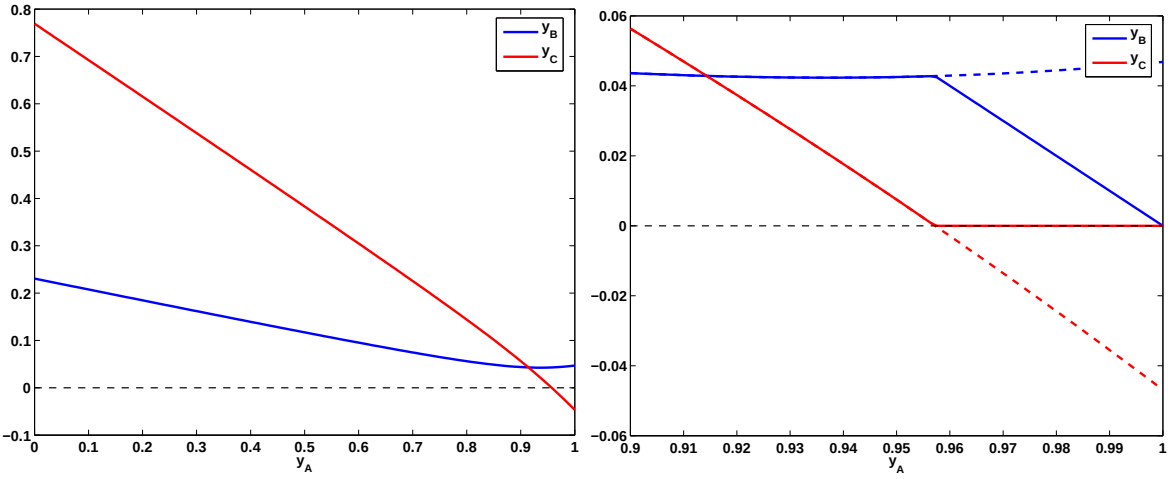


**Figure 3.5:** The solution of equations (3.9). Since $\psi$ is a reduction for a chemical model, the function values should be positive for all $y_A \in [0, 1]$. This is not the case for $y_A > 0.96$. The right figure shows one possibility to fix this problem: In the domain where $y_C$ is negative, it is set to $y_C = 0$ and the difference is subtracted from $y_B$ in order not to conserve the relation $y_A + y_B + y_C = 1$.

On top of the problems arising from the QSSA, there is a much more serious problem, when the ILDM is calculated. Whereas the source term $f$ is continuously differentiable, the function $Q^T f$ is not even continuous.

The reason for the discontinuities is the existence of multiple eigenvalues, a problem, which cannot be circumvented nicely. Consider two eigenvalues $\lambda_1(t)$ and $\lambda_2(t)$ of $\nabla f(u(t))$, where $u(t)$ describes a path such that $\lambda_1(t) < \lambda_2(t)$ for $t < t_0$ and $\lambda_1(t) > \lambda_2(t)$ for $t > t_0$. Since the matrix $R$ is sorted on its diagonal with respect to the real part of $\lambda_i$, this means that the two eigenvalues $\lambda_1$ and $\lambda_2$ change their position on the diagonal of $R$ at $t = t_0$. Then, of course, the rows of $Q$ have to be reordered at $t = t_0$, and this cannot be done in a continuous way.

In order to find a solution to the ILDM formulation (3.5) anyway, very good starting points for the Newton method are required in order to prevent the eigenvalues to intersect during the solution process. Another idea might be to reuse the Schur factorization for a few Newton steps, which avoids the discontinuities but may also reduce the convergence rates.

## 3.3 Center Manifolds

As already seen in 2.2, the modeling equations for chemical reaction systems contain conservation laws, which are not formulated explicitely. These implicitly given conditions in the equations reduce the number of degrees of freedom for the variables, which ensures that the set of all physically possible variables forms a manifold in the state space, the so called center manifold. If the initial condition of an ODE is on this manifold, the whole trajectory defined by the equations lies on the manifold as well.

**Remark:** Two small remarks for chemical reaction systems:

- In chemical reaction systems, the center manifold can be identified by lemma 2.2.1.
- From a physical point of view, two different types of conservation laws are to be distinguished. The first type has to be fulfilled by the initial value, whereas in the second type the initial value defines partly the center manifold. Consider the reaction (3.10) with the initial condition $y_0 = (y_A^0, y_B^0, y_{AB}^0)$. For a realistic simulation of this reaction, the sum $y_A^0 + y_B^0 + y_{AB}^0$ has to be equal to one, but the number of $A$– and $B$–atoms in the reactor is defined only by the initial value.

As an easy example, consider the simple reaction

$$A + B \longrightarrow AB \tag{3.10}$$

for the elements $A$ and $B$ with the specific molar masses $m_A$ and $m_B$. This reaction can be modeled by the ODE

$$\begin{pmatrix} \dot{y}_A(t) \\ \dot{y}_B(t) \\ \dot{y}_{AB}(t) \end{pmatrix} = \begin{pmatrix} -m_A y_A y_B \\ -m_B y_A y_B \\ (m_A + m_B) y_A y_B \end{pmatrix}, \quad y(0) = y_0,$$

where $y$ describes the vector of the species mass fractions $(y_A, y_B, y_{AB})$. There are two obvious conservation laws in the model, namely

$$\dot{y}_A + \frac{m_A}{m_A + m_B}\dot{y}_{AB} = 0 \quad \text{and} \quad \dot{y}_B + \frac{m_B}{m_A + m_B}\dot{y}_{AB} = 0,$$

therefore only one degree of freedom is to be calculated by the ODE, see figure 3.6.

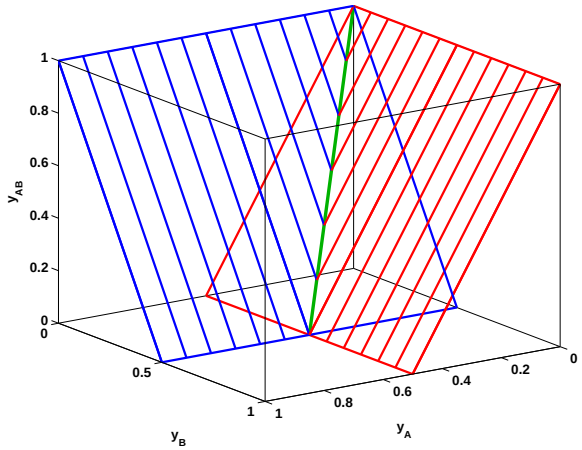


**Figure 3.6:** The center manifold for reaction (3.10) with initial conditions $y_A(0) = y_B(0) = 1/2$. The blue colored surface describes the manifold for the conservation of $y_A + \frac{1}{2}y_{AB}$, the red surface the manifold for $y_B + \frac{1}{2}y_{AB}$. The intersection yields the green center manifold.

The chemical interpretation of these conservation laws is the conservation of the element mass fractions. The term $\frac{m_A}{m_A+m_B}y_{AB}$ describes the element mass fraction of $A$, given that $A$ belongs to the molecule $AB$. This means that the total mass fraction of the element $A$ is denoted by $y_A + \frac{m_A}{m_A+m_B}y_{AB}$, which does not change in time, therefore the amount of $A$–atoms remains constant in the reaction. This is obviously also valid for the element $B$, as can be seen by the second conservation law.

In general, those conservation laws are not preserved in the calculation of the manifolds from QSSA and ILDM. In order not to change the physical meaning of the equations and their solution, the conservation laws have to be forced by hand.

To assure that the state vector $(u_1, \psi(u_1))$ is on the center manifold, the term $F_2(u_1, u_2)$ does not only contain the equations describing the fast processes of the model, but also the equations defining the center manifold. Reconsidering reaction (3.6), this means for the reduction method by QSSA, that the function $\psi$ is now a function from $\mathbb{R}$ to $\mathbb{R}^2$ and has to solve the equations

$$\begin{aligned} 4y_A^2 - 2002\psi_1(y_A)^2 + 180\psi_2(y_A)^2 &= 0 \\ y_A + \psi_1(y_A) + \psi_2(y_A) &= 1, \end{aligned}$$

if $y_A$ is taken as a process variable.

The introduction of the center manifolds to the reduction function $\psi$ has the big advantage, that the results from the reduced system is much more reliable, as can be seen in figure 3.7. But the center manifolds make the evaluation of $\psi(u_1)$ numerically more difficult, because

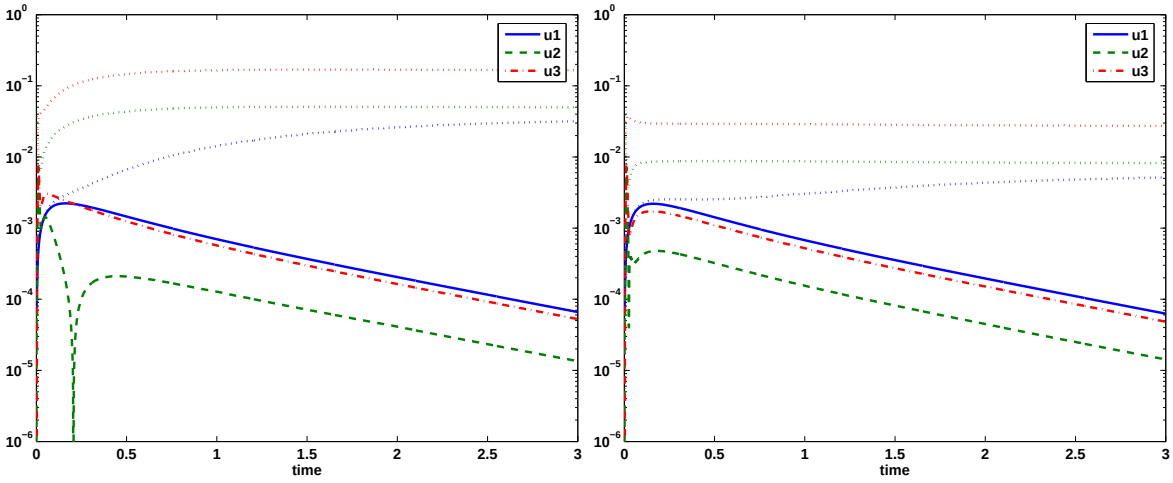


**Figure 3.7:** The reaction (3.6) reconsidered and the difference between the reduced and the detailed solution is plotted. The dotted lines represent the error, when the center manifold is neglected. The full lines are the occurring error, when the manifolds are now obtained by treating not only the fast variable $y_B$, but also the center manifold, which is in this case given by the equation $y_A+y_B+y_C = 1$. In the left figure, the reduction is based on QSSA, in the right figure on ILDM.

the equations for the center manifold introduce the eigenvalue 0 to the system (3.5), which increases its stiffness.

## 3.4 Calculation of QSSA– and ILDM–points

One possibility to evaluate the function $\psi$ for a given $u_1$ is the following algorithm containing three major steps:

- Finding an initial guess for the iteration
- Finding the zero by iterative methods (Newton's method and minimization)
- Post-processing of the solution $u_2$

### 3.4.1 The initial guess

As shown in section 3.2, the major problem of evaluating the function $\psi$ at a given point $u_1$ is identifying a decent initial guess for the Newton method. This is in fact the most time consuming part in the algorithm.

The initial guess for evaluating $\psi$ at a point $u_1$ is obtained by doing time steps for the problem

$$u_2'(t) = F_2(u_1, u_2), \quad u_2(0) = u_2^0, \tag{3.11}$$

where $F_2$ is defined by $f_2$ when the manifold is based on QSSA or by $[Q^T f]_2$ in case of ILDM. Reasonable results were obtained by choosing $u_2^0$, such that the fast variables (in chemical problems: the radicals) have the initial value 0, whereas the variables, which are to be calculated from the center manifolds, are chosen to have $u_2^0$ as close as possible to the center manifold. In chemical reaction systems they may be chosen, such that the sum over all mass fractions equals one.

For the time stepping a reasonable step size has to be chosen as well as the number of time steps, which are to be done. This is in fact a difficult task, simply because the system (3.11) is still stiff and highly nonlinear. For this reason the process of finding an initial guess is very time consuming.

When evaluating an ILDM–point, it is also possible to use the already calculated QSSA–point as an initial guess. In the considered examples, this turned out to be very successful, but in general there is no guarantee, that the existence of a QSSA–point implies the existence of an ILDM–point and vice versa.

Take for example the problem

$$\begin{aligned} \dot{u}_1 &= -u_1 \\ \dot{u}_2 &= 1000(1 - u_1 u_2). \end{aligned}$$

The parameterizations for the manifolds are clearly

$$\psi_Q(u_1) = {}^1/_{u_1} \quad \text{and} \quad \psi_I(u_1) = \frac{1000u_1 - 1}{1001u_1^2 - u_1}.$$

A QSSA–point does therefore exist for all $u_1 \in ]0, 1]$, whereas no ILDM–point can be found for $u_1 = {}^1/_{1001}$.

### 3.4.2 The iteration

The heart or the iteration is of course the Newton iteration. With the initial guess as described above, the method has in most cases nice convergence properties. But unfortunately, convergence is not guaranteed, even if the starting point is close to the existing steady state of (3.11), because the convergence radius for the Newton method is due to the high nonlinearity of $F_2$ very small.

In case of divergence of the Newton, a minimization strategy is applied to the function $\|f_2(u_1, u_2)\|_2$ or $\left\| \left[Q^T(u_1, u_2) f(u_1, u_2)\right]_2 \right\|_2$ respectively. The difficulty for the common minimization algorithms like BFGS [41, Chapter 1.6] or the gradient free simplex method [32] is again the stiffness of $f_2$, in general the result is only a poor approximation to the searched zero $u_2$. Anyway, this approximation can be used as a second initial guess to the Newton's method, which was successful in the considered examples.

### 3.4.3 Post–processing of the solution

Since center manifolds are included in the zero–finding process, the conservation laws of the original problem are fulfilled for the state vector $u = (u_1, \psi(u_1))$, but there is no reason to have the values of $\psi(u_1)$ to be in the interval $[0, 1]$. As already mentioned, there are only heuristical solutions to this problem. The entries of $\psi(u_1)$ have to be changed somehow, if one or more values are outside of $[0, 1]$.

One possibility is to apply a minimization strategy as above and add some constraints to the functional. In practice, this leads to problems, again due to the stiffness of the function $f_2$ or $[Q^T f]_2$. In fact, the standard minimization algorithms minimize mainly the fast part of the functional, neglecting the equations resulting from the center manifold with slow properties.

The second and more successful possibility is based on pure heuristics. If one variable $u_k$ turns out to be outside of the allowed interval $[0, 1]$, this variable is set to either 0, if $u_k < 0$ or 1 otherwise. Then the values of the other variables have to be changed such that the conservation laws are not violated. This method turns out to work pretty well with small examples with only a small number of variables, but it will definitely fail in big reaction systems. A confident strategy to find a physically decent solution to this problem is still part of research.

## 3.5 Tabulation of the manifold

In the previous section we have seen that the manifolds derived from both the quasi steady state assumption and the intrinsic low dimensional manifolds are able to reduce big systems of ordinary differential equations. The resulting systems, which are to be solved, are much smaller (only 10% of the original size, when the methane–air combustion is considered) and they are usually of nice behavior, because the stiffness is reduced.

This might give the impression that using these reduction methods makes the calculation much faster and less memory consuming, but the contrary is true, as Deuflhard already mentioned:

> *In order not to raise wrong expectations: in all of our experiments, the direct numerical integration of the unprepared stiff ODE system was much faster than the integration of the split DAE system [...]* [12]

The reason for the high numerical costs is the solving process for the equations

$$f_2(u_1, u_2) = 0 \quad \text{or} \quad \left[Q^T(u_1, u_2) f(u_1, u_2)\right]_2 = 0,$$

which has to be performed in every single time step. In addition to the solving process the Schur factorization has to be done in the ILDM case, which is also very expensive.

A solution to this problem is the a priori calculation of the function $\psi$. Since in only very few problems the function $\psi$ is given analytically, the value $u_2 = \psi(u_1)$ is calculated for all possible values of $u_1$, before the solution process for the ODE is started. The resulting vectors $u_2$ have to be stored in a file which provides quick access to the data, whenever the function $\psi$ has to be evaluated at a point $u_1$. This file is commonly called a table.

Obviously no computer is able to keep the function values for all possible vectors $u_1$, therefore only some vectors $u_2$ are calculated and the value $\psi(u_1)$ for an arbitrary $u_1$ is calculated by interpolation of the surrounding points. One possibility to identify the stored points is the discretization of the domain $D = [a_1, b_1] \times ... \times [a_{n_1}, b_{n_1}]$ of the function $\psi$ by

$$U_1 = \left\{ x \in D \;\middle|\; x = \begin{pmatrix} a_1 \\ \vdots \\ a_{n_1} \end{pmatrix} + \begin{pmatrix} k_1 \cdot h_1 \\ \vdots \\ k_{n_1} \cdot h_{n_1} \end{pmatrix}, \quad 0 \le k_i \le N_i \right\},$$

where $h_i$ is given by $h_i = \frac{b_i - a_i}{N_i}$. Other possibilities, for example with adaptively refined domains, are discussed in [34].

The big advantage of this possibility is that the stored points $u_1 \in U_1$ can be addressed by so called integer variables $k_i$, and for two elements $x_a, x_b \in U_1$, $x_a$ can be compared with $x_b$ without discrepancies due to rounding errors. Together with an alphabetical equivalence relation, this makes it possible to tabulate the function values $\psi(U_1)$ in a binary tree.

The structure of the binary tree ensures quick access to the data. The reading time grows in fact only with the logarithm of the number of tabulated points, see [22, 3.2], and the size of the table has therefore only little effect on the total calculation time.

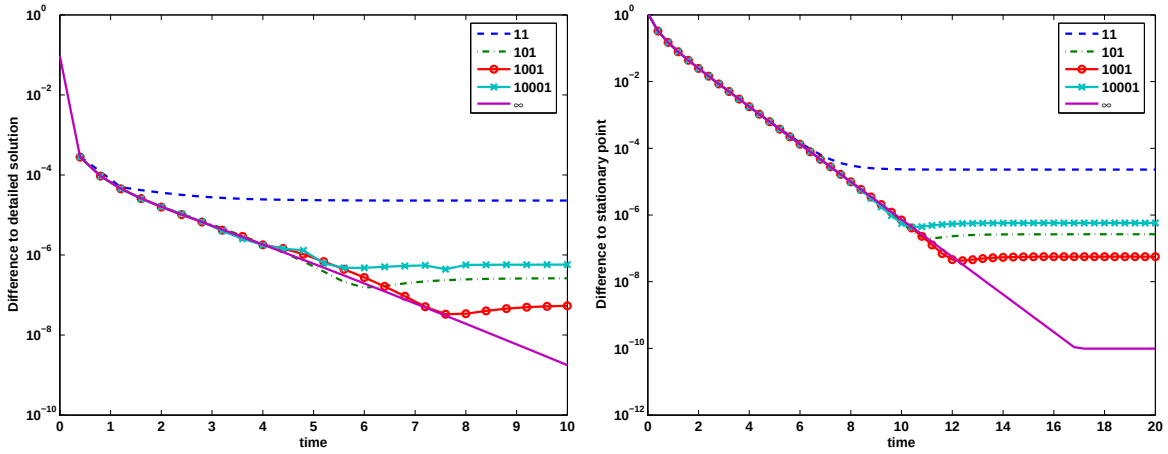


**Figure 3.8:** Both figures show the reaction (3.6) reduced with five different QSSA–tables, each of them contains a different number of points. The table, which is denoted by $\infty$, is in fact an analytically given parameterization. The left figure shows the difference to the detailed solution $\|u(t) - u_h(t)\|_2$, the right figure the difference of the currently reached state to the analytically calculated steady state $\|u(\infty) - u_h(t)\|_2$.

The interpolation error is clearly of order $h^2$, if linear interpolation is applied, and decreases therefore quadratically (or faster for interpolation strategies of higher order) with the number of equidistantly tabulated points. Therefore large tables are expected to lead to better performances. This is unfortunately only partly true. First of all, the logarithmic order of the access time is only valid, if the whole table is stored in the memory, which leads to an upper limit for the number of stored points. And secondly, the interpolation introduces rounding errors, which cannot be circumvented by bigger tables, if the accuracy of the calculation of the ILDM– or QSSA–points is not increased. This is especially in the calculation of derivatives of $\psi$ a serious problem. These derivatives are necessary in the Newton iteration, as will be shown in the following sections.

One drawback of the tabulation is the enormous requirement of memory. A table for the methane–air reaction mechanism for example with four process variables and 501396 stored points as provided by the workgroup of Prof. Warnatz takes around 320 MB of disk space. Due to the frequent access, which is required within the calculation of a reaction system, the table has to be copied into the work space. The size of the table is therefore limited, especially the number of process variables. Up to now, only tables with at most 4 process variables can be treated.

A second possibility for the storage of the manifolds is based on orthogonal polynomials. Details can be found in [35].

### 3.5.1 Criteria for the performance of the table

As shown in chapter 3.2.3, small differences in the function $\psi$ can have big influences on the accuracy of the solution of the reduced problem. Therefore the approximation goodness of the discrete table to the continuously defined function $\psi$ cannot be the only criterion for the performance of the table.

Other criteria consider therefore the effect of the table to practical problems, here, two possibilities are considered, both acting on ODEs. The first criterion is the error produced in a certain time interval. Secondly, the steady state is considered and compared to the detailed solution.

Figure 3.8 shows the effect of the size of a QSSA–table to the solution of an ODE. Here four tables with $N = 10^p$, $p = 1, ..., 4$ are considered. The figure shows that the size of the table has nearly no influence to the solution in early stages, when the table with 11 points is excluded. This is the region, where the systematic error produced by the reaction process dominates the errors from the interpolation process. The contrary is true, if the solution is considered close to the stationary point. Here bigger tables seem to perform much better than the smaller ones up to a certain number of tabulated points. If the table contains too many points, the interpolation error increases with increasing number of stored points. In the considered case, where the tables are stored with an accuracy of 6 digits, the QSSA–table with 1001 stored points seems to be the best. Even though the error of order $10^{-4}$ is still accurate enough for many practical purposes, the bigger tables are to be preferred in order to obtain higher accuracies.

# 4 Solution Process with QSSA and ILDM

The goal of this chapter is to lead the reader to the difference of the detailed and the reduced solution of systems of algebraic equations and develop then strategies to solve partial differential equations with the reduction techniques.

First, linear reaction mechanisms are considered. On the one hand, they have the advantage that many results can be proven analytically. On the other hand, the investigation of the quasi–Newton method in chapter 6 is based on the reduction of the arising system of linear equations. These linear equations are similar to equations with linear reaction mechanisms, which motivates the separate treatment of linear reactions. It will be shown that ILDM performs much better than QSSA, there are even relevant examples, where the reduction with ILDM does not produce any errors.

Thereafter, the obtained results will be extended as far as possible to nonlinear reactions. For both the linear and the nonlinear reactions, the analysis includes the treatment of exact reductions, the error analysis for general equations and the consideration of a reaction–diffusion equation as a more realistic problem. In both cases, mechanisms with and without center manifolds are treated separately.

## 4.1 Description of the reduction strategy

Assume that the solution of the algebraic equation

$$L(u) + F(u) = b$$

with $F$ denoting a chemical source term is to be calculated numerically with a black box solver. If this solver is based on the Newton iteration, it has (at least) two interfaces for the user. One interface addresses the residual of the underlying equations, the other interface provides the solver with the Jacobi matrix. By implementing these two functions, the black box solver can be used to solve the considered system of algebraic equations.

In order to use this solver for the reduced equation

$$L_1(u_1, u_2) + F_1(u_1, u_2) = b_1$$

with a given function $u_2 = \psi(u_1)$, a black box solver can be applied by implementing interfaces between the solver, which iterates only on the process variables $u_1$, and the functions providing the residual and the matrix of the system, see figure 4.1. These interfaces compute the total state vector by applying the function $\psi$ to $u_1$ and send the resulting vector $u$ to the user–provided functions calculating the residual and its derivative. The obtained residual $res$ and matrix $M$ is then splitted according to the process variables, and $res_1$ and $M_{11}$ can be provided for the solver.

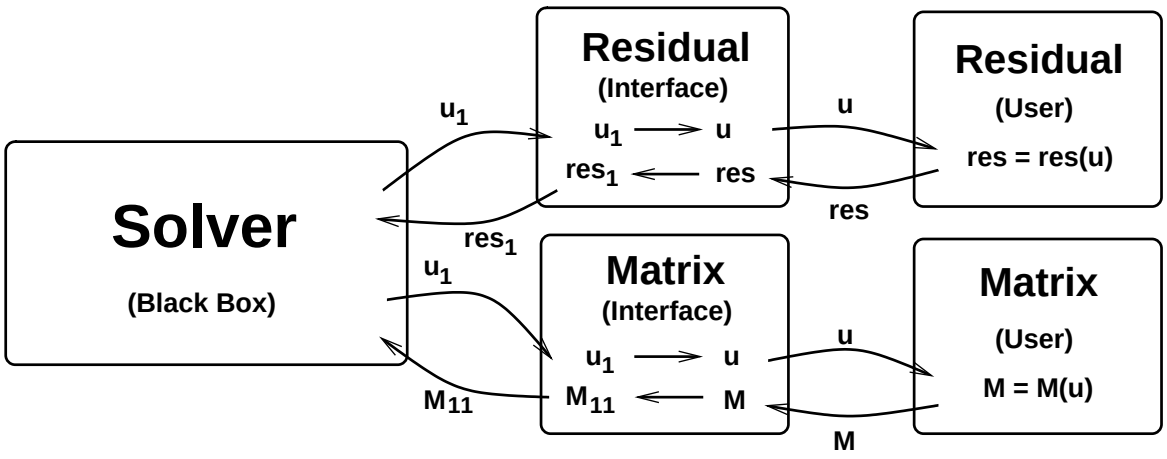


**Figure 4.1:** Solution strategy with reduction methods: The solver computes the solution for the process variables only. If information about the residual or its derivative is required, an interface has to get the total state vector via the tables, then calculate the required entity and return only the slow part to the solver.

## 4.2 Linear equations without conservation laws

In order to understand the principles of QSSA and ILDM, consider a linear reaction mechanisms without conservation laws first. Then the source term $f : \mathbb{R}^n \to \mathbb{R}^n$ can be described by a matrix $A$:

$$f(u) = Au,$$

where $A$ is due to the lack of conservation laws necessarily invertible.

This section will first introduce the analytic description of the QSSA– and ILDM–table and their effect as reduction methods to systems of linear equations. Then criteria are given, when the reduction methods lead to exact solutions without systematic errors. This result will be extended to ODEs and PDEs, but also equations, where systematic errors occur, will be investigated. This section concludes with a deeper analysis of a reaction–diffusion equation and investigations on the computational costs.

The main result of this section is given in theorem 4.2.1 about the possibilities to reduce systems of linear equations without any errors. A consequence of this theorem is that linear ODEs are exactly solved by the application of ILDM, whereas QSSA introduces systematic errors. The same is true for special examples of PDEs.

### 4.2.1 The parameterization of the manifolds

Let the general formulation of the ODE be

$$\begin{pmatrix} \dot{u}_1 \\ \dot{u}_2 \end{pmatrix} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix}$$

with the process variables $u_1$ and the fast variables $u_2$. Let further $V, \Lambda \in \mathbb{R}^{n\times n}$ such that $\Lambda$ is diagonal and

$$VA = \Lambda V = \begin{pmatrix} \Lambda_1 & 0 \\ 0 & \Lambda_2 \end{pmatrix} \begin{pmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{pmatrix}.$$

Then the rows of $V$ describe the left eigenvectors of the matrix $A$ and the entries of $\Lambda$ its eigenvalues. The eigenvalues $\lambda_i$ of $A$ shall be ordered on the diagonal of $\Lambda$ such that $\mathrm{Re}\lambda_1 \geq ... \geq \mathrm{Re}\lambda_n$. If $A$ has a spectral gap, i.e. $\mathrm{Re}\lambda_{n_1} \gg \mathrm{Re}\lambda_{n_1+1}$, then $\Lambda_1$ shall be in $\mathbb{R}^{n_1\times n_1}$ and $\Lambda_2 \in \mathbb{R}^{n_1\times n_2}$ with $n_1 + n_2 = n$. Let further the equations be ordered such that $u_1$ describes the slow process variables.

The parameterizations of the manifolds are then given by the functions

$$\psi_Q(u_1) = -A_{22}^{-1} A_{21} u_1 \tag{4.1}$$

for the QSSA–parameterization and

$$\psi_I(u_1) = -(V_{21}A_{12} + V_{22}A_{22})^{-1}(V_{21}A_{11} + V_{22}A_{21})u_1 = -V_{22}^{-1}V_{21}u_1$$

for the ILDM–parameterization, assuming the invertibility of $A_{22}$ and $V_{22}$.

As an example, take the linear reaction

$$sp_1 \underset{1}{\overset{2}{\rightleftharpoons}} sp_2 \underset{90}{\overset{1000}{\rightleftharpoons}} sp_3 \xrightarrow{500} sp_4, \tag{4.2}$$

where the mass fractions of only the species $sp_1$, $sp_2$ and $sp_3$ are modeled in the ODE and the mass fraction of $sp_4$ is obtained by $y_4 = 1 - y_1 - y_2 - y_3$. The parameterizations read then

$$\psi_Q(u_1) = \begin{pmatrix} 0.002357 \\ 0.003995 \end{pmatrix} u_1 \quad \text{and} \quad \psi_I(u_1) = \begin{pmatrix} 0.002364 \\ 0.004021 \end{pmatrix} u_1.$$

In the upcoming parts, the calculation of the stationary points of the problem

$$\dot{u} = (A + \alpha B)u - \beta b$$

is considered with the matrix $A$ describing the source term for the above chemical reaction, for which the reduction mechanisms were calculated. The constants $\alpha$ and $\beta$ and the (almost) arbitrary disturbance matrix $B$ is supposed to be chosen, such that $A + \alpha B$ is invertible. The vector $b$ shall also be arbitrary.

First of all, conditions for the disturbances $B$ and $b$ are given, such that the approximation by the parameterizations does not produce any systematic errors. Then two possibilities for the application of the parameterization $\psi$ for arbitrary values of $B$ and $b$ will be considered, namely the direct application of the original parameterization $\psi$, and the creation of a new parameterization $\phi$ on the basis of the original parameterization will be treated.

### 4.2.2 Description of reduced problems

The reduced problem to $(A + \alpha B)u = \beta b$ reads now

$$\begin{aligned}(A_{11} + \alpha B_{11})u_1 + (A_{12} + \alpha B_{12})\psi(u_1) &= \beta b_1\\ u_2 &= \psi(u_1),\end{aligned}$$

hence

$$\begin{aligned}(A_{11} + \alpha B_{11} - (A_{12} + \alpha B_{12})A_{22}^{-1}A_{21})u_1 &= \beta b_1\\ u_2 &= -A_{22}^{-1}A_{21}u_1\end{aligned}$$

for the QSSA–parameterization and

$$\begin{aligned}(A_{11} + \alpha B_{11} - (A_{12} + \alpha B_{12})V_{22}^{-1}V_{21})u_1 &= \beta b_1\\ u_2 &= -V_{22}^{-1}V_{21}u_1,\end{aligned}$$

if the ILDM–parameterization is applied. The solution to these reduced problems solve in case of QSSA the equations

$$\begin{pmatrix} A_{11} + \alpha B_{11} & A_{12} + \alpha B_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} \beta b_1 \\ 0 \end{pmatrix}, \tag{4.3}$$

and

$$\begin{pmatrix} A_{11} + \alpha B_{11} & A_{12} + \alpha B_{12} \\ \Lambda_2 V_{21} & \Lambda_2 V_{22} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} \beta b_1 \\ 0 \end{pmatrix}$$

for ILDM.

Clearly, the reduced equations and the original equation do in general have different solutions. But under certain circumstances, the reduced equations lead to the same solution as the detailed problem, as the following theorem shows.

**Theorem 4.2.1** *Let $A \in \mathbb{R}^{n\times n}$ be arbitrary but invertible. Define the function $\psi : \mathbb{R}^{n_1} \to \mathbb{R}^{n_2}$ such that the vector $u = (u_1, \psi(u_1))^T$ is on the manifold defined by QSSA (ILDM) for all values of $u_1$. If $B \in \mathbb{R}^{n\times n}$ and $r \in \mathbb{R}^n$ are such that $u$ is on the QSSA–(ILDM–)manifold for the function $f(u) = Bu - r$, then the solution to the reduced problem*

$$\begin{aligned}(A_{11} + B_{11} + (A_{12} + B_{12})\psi)u_1 &= r_1\\ u_2 &= \psi(u_1)\end{aligned}$$

*is also a solution of the original problem $(A + B)u = r$.*

**Proof:** The proof will be split into two cases. In the first case, $\psi$ is assumed to be created by the QSSA–method, in the second case, $\psi$ represents the ILDM–parameterization.

1. Let $u = (u_1, u_2)^T$ be the solution of the reduced problem, where the QSSA–parameterization was applied. Then the relation $u_2 = \psi(u_1)$ does necessarily hold. The fact that $\psi$ is also the QSSA–parameterization for the problem $\dot{u} = Bu - r$, leads to the relation $B_{21}u_1 + B_{22}u_2 = r_2$, and therefore to

$$(A+B)u = \begin{pmatrix} A_{11}+B_{11} & A_{12}+B_{12} \\ A_{21} & A_{22} \end{pmatrix} u + \begin{pmatrix} 0 & 0 \\ B_{21} & B_{22} \end{pmatrix} u = \begin{pmatrix} r_1 \\ 0 \end{pmatrix} + \begin{pmatrix} 0 \\ r_2 \end{pmatrix} = r.$$

   The reduced solution does therefore solve the detailed problem.

2. Let the function $\psi$ now represent the ILDM–parameterization for the problems

$$\dot{u} = Au \quad \text{and} \quad \dot{u} = Bu - r.$$

   This implies the relation

$$V_2 B u = V_2 r$$

   for all $u$ on the manifold. Let now $u$ be the solution of the reduced problem. Then $u$ is necessarily on the manifold and the following holds:

$$\begin{pmatrix} \mathbb{1} & 0 \\ V_{21} & V_{22} \end{pmatrix} (A+B)u = \begin{pmatrix} r_1 \\ V_2 B u \end{pmatrix} = \begin{pmatrix} \mathbb{1} & 0 \\ V_{21} & V_{22} \end{pmatrix} r,$$

   because $V_2 A u = 0$. The existence of an ILDM–table implies the invertibility of $V_{22}$, which leads to the equation $(A+B)u = r$.

■

A few remarks to this theorem:

- The above described conditions for $B$ and $r$ are not the only possibility to obtain solutions without systematic errors. Let for example $r = 0$. Then $u = 0$ is the solution of both the reduced and the detailed problem independently of $B$.
- An easy possibility to fulfill the conditions in the case of QSSA is $B_{21} = 0$, $B_{22} = 0$ and $r_2 = 0$.
- The approximation is also exact, if the fast variables are decoupled from the slow ones, i.e. $A_{21} = B_{21} = 0$, which implies $V_{21} = 0$. In addition, $b_2 = 0$ is required.
- The requirement

$$B_{21}u_1 + B_{22}u_2 = r_2$$

  in case of QSSA leads with $u_1 = (A_{11} + B_{11} - (A_{12} + B_{12})A_{22}^{-1}A_{21})^{-1}r_1$ and $u_2 = -A_{22}^{-1}A_{21}u_1$ to the condition

$$r_2 = (B_{21} - B_{22}A_{22}^{-1}A_{21})(A_{11} + B_{11} - (A_{12} + B_{12})A_{22}^{-1}A_{21})^{-1}r_1$$

  for the vector $r$.

- The necessary conditions for the ILDM–parameterizations are fulfilled, if $AB = BA$ (which implies that the eigenvectors of $A$ are also eigenvectors of $B$) and $r_2 = \psi(r_1)$.
- The beautiful condition $r_2 = \psi(r_1)$ can only be used for calculations with ILDM. It leads to systematic errors for the QSSA–parameterization.

### 4.2.3 The importance of the spectral gap

An interesting aspect of theorem 4.2.1 is the fact, that the existence of a spectral gap for the matrix $A$ was not required. In order to obtain the exact solution for linear problems, the ILDM method can be applied to all (invertible) matrices, given that the right hand side $r$ and the possible second matrix $B$ fulfill the conditions required in the theorem.

Nevertheless, the existence of a spectral gap leads still to advantages, if the required conditions are not fulfilled exactly. Let in case of ILDM the fast part of the right hand side be given by $r_2 = \psi(r_1) + \delta$. Then the error in the solution depends linearly on the perturbation $\delta$:

$$(A+B)^{-1}\begin{pmatrix} r_1 \\ \psi(r_1)+\delta \end{pmatrix} - \begin{pmatrix} A_{11}+B_{11} & A_{12}+B_{12} \\ V_{21} & V_{22} \end{pmatrix}^{-1}\begin{pmatrix} r_1 \\ 0 \end{pmatrix}$$
$$= (A+B)^{-1}\begin{pmatrix} r_1 \\ \psi(r_1)+\delta \end{pmatrix} - (A+B)^{-1}\begin{pmatrix} r_1 \\ \psi(r_1) \end{pmatrix} = (A+B)^{-1}\begin{pmatrix} 0 \\ \delta \end{pmatrix}.$$

Assume now that $A+B$ is of block–diagonal form. Then the error for the slow variables is zero and is for the fast variables given by $(A_{22}+B_{22})^{-1}\delta$. This error is of course smaller with larger values of $A_{22}+B_{22}$, therfore with large eigenvalues. This is supported by a big spectral gap. If $A+B$ is not of block–diagonal form, the error does also depend on the coupling between the fast and the slow variables, but is still independent of the block $A_{11}+B_{11}$.

The accuracy of the approximated solution of the problem

$$Au = b$$

with the QSSA method and $b_2 = \psi(b_1)$ does also depend on the size of the eigenvalues of the block $A_{22}$. As an example, consider the three matrices

$$A_1 = \begin{pmatrix} -2 & 1 & 0 \\ 2 & -1001 & 90 \\ 0 & 1000 & -2000 \end{pmatrix}, \quad A_2 = \begin{pmatrix} -2 & 1 & 0 \\ 2 & -101 & 9 \\ 0 & 100 & -200 \end{pmatrix}$$

and

$$A_3 = \begin{pmatrix} -2000 & 1 & 0 \\ 2 & -1001 & 90 \\ 0 & 1000 & -2000 \end{pmatrix}$$

with the right hand side $b = (1, \psi(1))^T$. These matrices have the eigenvalues $A_1$: $\lambda \in \{-2.0, -918, -2080\}$, $A_2 : \lambda \in \{-2.0, -93, -208\}$ and $A_3 : \lambda \in \{-918, -2000, -2080\}$. The first two matrices do therefore have a spectral gap of different size, whereas the matrix $A_3$ has only large negative eigenvalues without spectral gap. The difference between the detailed solution and the solution obtained by the QSSA method is then $|u - u_Q| = 3e-6$ for the first matrix and $|u - u_Q| = 3e-4$ and $|u - u_Q| = 3e-6$ for the second and third matrix. This suggests clearly, that for a decent approximation, the fast part has to be purely fast, but is independent of the slow part $A_{11}$ of the matrix and therefore of the spectral gap.

### 4.2.4 Reducing ordinary differential equations

A consequence of theorem 4.2.1 is that the ODE

$$\dot{u} = Au$$
$$u(0) = \begin{pmatrix} u_{1,0} \\ \psi(u_{1,0}) \end{pmatrix}$$

can be nicely reduced with the ILDM method: If the implicit Euler is applied for one time step of size $k$, the corresponding linear equation

$$(\mathbb{1} - k \cdot A)\, u(t+k) = u(t)$$

can be exactly reduced with ILDM in the first time step, because the table for the matrix $k \cdot A$ equals the table of $A$, the matrices $A$ and $\mathbb{1}$ commute and $u(0)$ is on the manifold. Clearly, the resulting state $u(k)$ fulfills the condition $u_2(k) = \psi(u_1(k))$ and is therefore also on the manifold generated by ILDM. But since $u(k)$ defines the right hand side for the linear equation for the next time step $u(2k)$, the following time steps are also exactly reduced by the ILDM method.

The reduction method QSSA cannot be used in the same way. In order to obtain the exact solution by the application of the QSSA method in the first time step, the initial conditions $u_0$ have to be chosen such that

$$u_{0,2} = -A_{22}^{-1} A_{21} (\mathbb{1} - k(A_{11} - A_{12} A_{22}^{-1} A_{21}))^{-1} u_{0,1}. \tag{4.4}$$

This can easily be verified by considering the reduced solution in the detailed equation:

$$(\mathbb{1} - kA) \begin{pmatrix} u_1 \\ -A_{22}^{-1} A_{21} u_1 \end{pmatrix} = \begin{pmatrix} u_1 - k(A_{11} - A_{12} A_{22}^{-1} A_{21}) u_1 \\ -A_{22}^{-1} A_{21} u_1 \end{pmatrix} = \begin{pmatrix} u_{0,1} \\ u_{0,2} \end{pmatrix},$$

which leads to the above condition for $u_{2,0}$.

Since the equations for the first time step are now exactly solved by the QSSA method, the state vector $u_1$ is clearly on the QSSA–manifold and fulfills the condition

$$u_{1,2} = -A_{22}^{-1} A_{21} u_{1,1}.$$

But in order to obtain an exact state in the second time step, the first time step has to fulfill a condition similar to (4.4), which is a contradiction.

This observation leads to the question, how big the differences are between the exact solution and the solution obtained by applying the QSSA method. The difference between the detailed solver and the solutions obtained from QSSA for the problem

$$\dot{u} = Au$$
$$u(0) = \begin{pmatrix} 1 \\ \psi(1) \end{pmatrix}$$

for $A$ as in (4.8) is shown in figure 4.2. The figure depicts clearly, that the error for QSSA is dominated by the difference of the fast variables, which leads to an almost constant error. The difference of the slow variable seems to increase linearly.

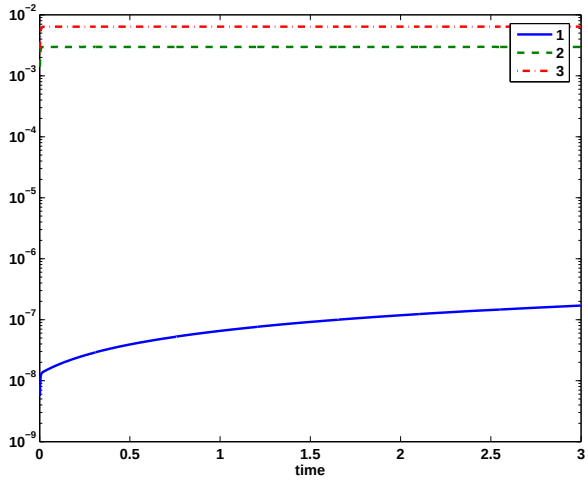


**Figure 4.2:** The relative difference between the exact and QSSA reduced solution of $\dot{u} = Au$, where $A$ defines the source term of reaction (4.2).

### 4.2.5 Reduction of a simple reaction–diffusion equation

Consider now in 1d a very simple form of a reaction–diffusion equation. The diffusion shall be modeled by the Laplacian with the same diffusion coefficient for all species, which leads to the equation

$$-\nu u''(x) = Au(x), \quad x \in \Omega := ]0, 1[. \tag{4.5}$$

Dirichlet boundary conditions will be considered. The state vector $u$ shall describe the mass fraction of the three species of the linear reaction (4.2), the domain $\Omega$ shall be equidistantly discretized by the grid points

$$\Omega_h := \{0, h, ..., (N-1)h, Nh\}$$

with $Nh = 1$.

This leads with the finite element method to the linear algebraic equations

$$(A_h + B_h)u_h = b_h \tag{4.6}$$

with blocked state vector $u_h$, where each block describes the mass fractions of the chemical species at the corresponding grid point. If the matrices are lumped, $A_h$ and $B_h$ read

$$A_h = \begin{pmatrix} \mathbb{1} & 0 & \cdots & \cdots & 0 \\ 0 & -hA & 0 & & 0 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & 0 & -hA & 0 \\ 0 & \cdots & \cdots & 0 & \mathbb{1} \end{pmatrix} \quad \text{and} \quad B_h = \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ -1/h\mathbb{1} & 2/h\mathbb{1} & -1/h\mathbb{1} & & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & -1/h\mathbb{1} & 2/h\mathbb{1} & -1/h\mathbb{1} \\ 0 & \cdots & \cdots & 0 & 0 \end{pmatrix},$$

and the right hand side $b_h$ is basically zero. Only the first and the last three entries are given by the boundary values of the differential equation. In order to investigate the influence of the parameterizations, the boundary values have to be chosen according to the value $\psi(u_1)$.

Clearly, $A_h$ and $B_h$ commute, which might under certain circumstances lead to the exact solution, if the ILDM–parameterization is applied. The trueness of this conjecture is shown in the following lemma.

**Lemma 4.2.2** *Let the matrix $A$ describe a linear chemical reaction without conservation laws and let the rows of $\begin{pmatrix} V_{21} & V_{22} \end{pmatrix}$ describe the left eigenvectors of the fast eigenvalues of $A$. Then the reduction of*

$$\left[ \begin{pmatrix} \mathbb{1} & & & & \\ & A & & & \\ & & \ddots & & \\ & & & A & \\ & & & & \mathbb{1} \end{pmatrix} + \begin{pmatrix} 0 & & & & \\ a\mathbb{1} & b\mathbb{1} & a\mathbb{1} & & \\ & \ddots & \ddots & \ddots & \\ & & a\mathbb{1} & b\mathbb{1} & a\mathbb{1} \\ & & & & 0 \end{pmatrix} \right] u = \begin{pmatrix} r \\ 0 \\ \vdots \\ 0 \\ r \end{pmatrix}$$

*by the ILDM–parameterization leads to the same solution $u$ as the original problem, if the fast part of $r$ is chosen to be $r_2 = \psi(r_1)$.*

**Proof:** The lemma can be proven inductively. Clearly the exact solution on the first and last node is on the manifold. The exact solution fulfills on the inner nodes the relation

$$(A + b)u = -a(u^{upper} + u^{lower}), \tag{4.7}$$

which is solved exactly by the ILDM method (see theorem 4.2.1), because both $u^{upper}$ and $u^{lower}$ are on the manifold by induction and therefore $-a(u^{upper} + u^{lower})$ as well. ■

This lemma shows that the above simple reaction–diffusion equation is reduced exactly with the ILDM method. More advanced diffusion models, take for example Fick's law [17], do in general depend not only on one, but on several species, which introduces additional coupling to the blocks of $B_h$. This contradicts the exact reduction by the ILDM method.

Note that a similar lemma for QSSA–parameterizations cannot exist: The exactness of the QSSA–reduced problem requires $r_2 = -bA_{22}^{-1}A_{21}(A_{11} + b\mathbb{1} - A_{12}A_{22}^{-1}A_{21})^{-1}r_1$ (compare remark to theorem 4.2.1) in order to obtain correct values for $u$ on the boundary nodes. This implies $u_2 = \psi(u_1)$ at the boundary, and (4.7) is not reduced exactly anymore.

This observation raises clearly the question to the size of the errors introduced by the application of the QSSA method. An answer will be given in section 4.2.7.

### 4.2.6 Reduction of general equations

The previous sections showed clearly that the application of the reduction mechanisms does not necessarily lead to systematic errors. For special examples, the reduced solutions equal the exactly obtained solutions.

Now the effect of the reduced solution processes to general problems shall be investigated. For that reason, the difference between the detailed and reduced solution to the problem $(A + \alpha B)u = \beta b$ with

$$A = \begin{pmatrix} -2 & 1 & 0 \\ 2 & -1001 & 90 \\ 0 & 1000 & -590 \end{pmatrix}, \quad B = \begin{pmatrix} -2 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \tag{4.8}$$

will be considered. This difference does in fact depend strongly on the influence of $B$ and the right hand side $b$ and can be seen in figure 4.3. In this example, the matrix $A$ describes the source term of the linear reaction (4.2). The error seems to depend linearly on $\beta$ and

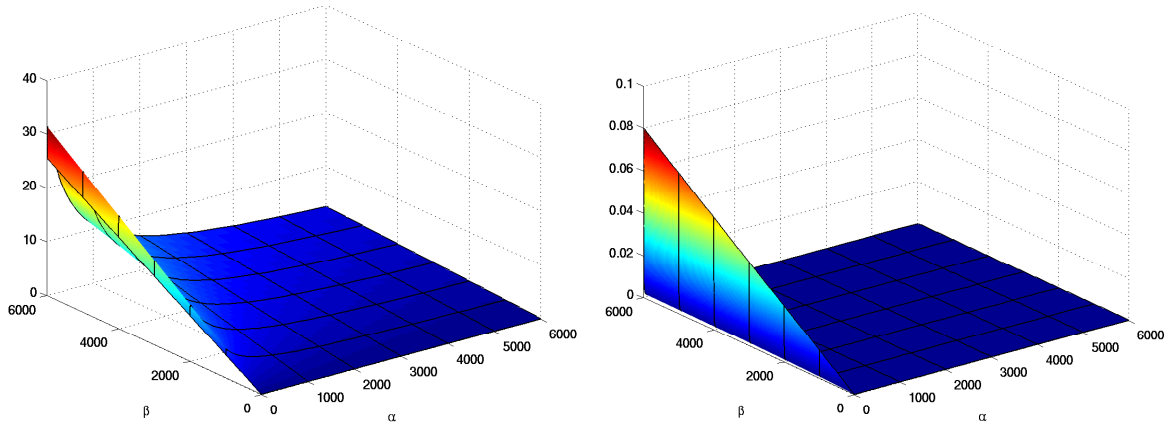


**Figure 4.3:** The difference between the detailed and reduced solutions to $(A+\alpha B)u = \beta b$, if the original ILDM–parameterization $\psi_I$ is applied. Here, $A$, $B$ and $b$ are defined by (4.8). The right figure shows the difference of the solutions obtained by the application of QSSA and ILDM.

decreases up to a certain limit with increasing $\alpha$, therefore with increasing influence of the second operator $B$.

### Create a modified parameterization

The big errors produced by the application of the reduction methods to arbitrary equation systems leads to the question, if the accuracy can be increased by allowing additional computations in the evaluation of the tables. For that reason, a new parameterization will be created on the basis of the original parameterization $\psi$. This new parameterization $\phi$ is supposed to fulfill the conditions for QSSA and ILDM for the perturbed problem. The function $\phi$ is therefore supposed to contain information not only on the chemical source term, but also on the second matrix $B$ and the right hand side $b$. In addition, the evaluation of $\phi$ shall be based on the already existing parameterization $\psi$, such that the additional computational effort for the calculation of $\phi$ is small.

In the best case, the new parameterization $\phi_Q$ for the QSSA method solves the equation

$$(A_{21} + \alpha B_{21})u_1 + (A_{22} + \alpha B_{22})\phi_Q(u_1) = \beta b_2,$$

so the value of $\phi_Q(u_1)$ is calculated to be

$$\phi_Q(u_1) = (A_{22} + \alpha B_{22})^{-1}(A_{22}\psi(u_1) - \alpha B_{21}u_1 + \beta b_2). \tag{4.9}$$

This formulation contains all the required information on the fast parts of $B$ and $b$, but requires the inversion of the matrix $A_{22} + B_{22}$. Two possibilities arise:

- Accept the additional computational price.
- Reduce the above formulation to

  $$\phi_Q(u_1) = \psi_Q(u_1) + \beta A_{22}^{-1} b_2.$$

  Clearly, the inversion of $A_{22}$ is equally expensive to the inversion of the matrix $A_{22} + B_{22}$, if the matrices describe the coupling of the species in one node. But if $A$ and $B$ describe discretized differential operators on a large grid, then the chemical source term $A$ is of block diagonal form, whereas $B$ contains also the coupling of the nodes. Then the inversion of $A_{22} + B_{22}$ is much more expensive than the inversion of $A_{22}$, which can be performed node–wise.

**Consider the second possibility first.** Note that the formulation of the renewed parameterization for ILDM–tables reads

$$\phi_I(u_1) = \psi_I(u_1) + \beta(\Lambda_2 V_{22})^{-1} \begin{pmatrix} V_{21} & V_{22} \end{pmatrix} b.$$

The solution $u$ to the reduced problem resulting from the application of $\phi$ to the original problem $(A + \alpha B)u = \beta b$ is the solution to the system

$$\begin{pmatrix} A_{11} + \alpha B_{11} & A_{12} + \alpha B_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} \beta b_1 \\ \beta b_2 \end{pmatrix}$$

for QSSA and

$$\begin{pmatrix} A_{11} + \alpha B_{11} & A_{12} + \alpha B_{12} \\ \Lambda_2 V_{21} & \Lambda_2 V_{22} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \beta \begin{pmatrix} b_1 \\ \widetilde{b_2} \end{pmatrix}$$

with $\tilde{b}_2 := \begin{pmatrix} V_{21} & V_{22} \end{pmatrix}$, if the ILDM–parameterization $\phi_I$ is applied.

The absence of $B_{21}$ and $B_{22}$ in the formulations above leads of course to errors. These errors are visualized in figure 4.4. The figure suggests that the error depends linearly on the influence of the right hand side $\beta$, as in the case of the original parameterization, compare figure 4.3. But an increasing $\alpha$ and therefore an increasing effect of $B$ leads now to an increasing error.

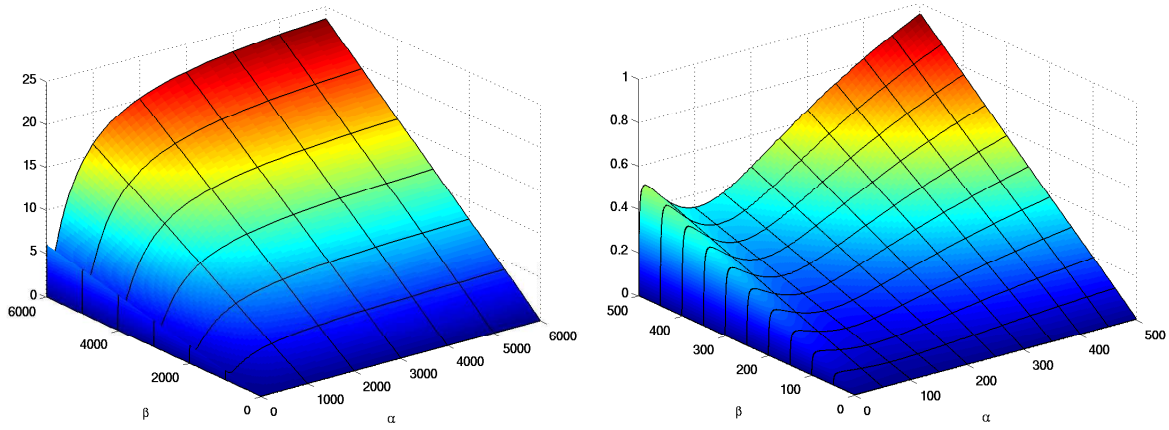


**Figure 4.4:** The error produced by the application of the renewed ILDM–parameterization $\phi_I$ to the problem $(A+\alpha B)u = \beta b$ with $A$, $B$ and $b$ as in (4.8). The right figure shows the same error with a different scale.

This phenomenon is probably best understood, if both $A$ and $B$ are supposed to be of block diagonal form and the QSSA–parameterization is applied. Then the solution to the original problem

$$\begin{aligned}(A_{11} + \alpha B_{11})u_1 &= \beta b_1\\ (A_{22} + \alpha B_{22})u_2 &= \beta b_2\end{aligned}$$

is to be compared with the two approximations

$$\begin{aligned}(A_{11} + \alpha B_{11})u_1 &= \beta b_1\\ A_{22}u_2 &= 0\end{aligned} \qquad \text{and} \qquad \begin{aligned}(A_{11} + \alpha B_{11})u_1 &= \beta b_1\\ A_{22}u_2 &= \beta b_2,\end{aligned}$$

where the first system describes the application of $\psi$ and the second the application of the renewed parameterization $\phi$. Let now $\beta$ be constant. Then the fast part of the detailed solution $u_2$ tends to zero for increasing $\alpha$, which is perfectly approximated by $\psi$, but not by $\phi$. This leads to the increasing error in figure 4.4 for constant $\beta$, whereas the error decreases in figure 4.3.

A similar analysis can be done for constant $\alpha$ with increasing $\beta$. This explains the linear dependence of the error on the size of $b_2$.

The difference between the performance of the ILDM– and the QSSA–parameterization is again rather moderate, as figure 4.5 shows. The figure shows clearly, that QSSA and ILDM perform equivalently, if

- $\beta = 0$: Both reduction methods lead to the exact solution $u = 0$. This explains also, why the error is zero for $\beta = 0$ in figure 4.4.
- $\alpha = 0$: If $u$ solves the system

$$\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} u_1 \\ u_1 \end{pmatrix} = \beta \begin{pmatrix} b_1 \\ b_2 \end{pmatrix},$$

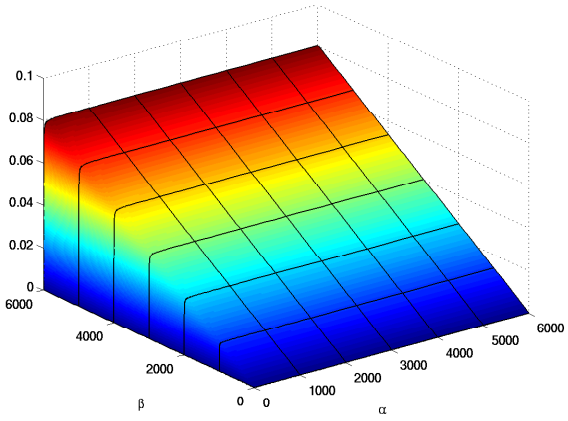


**Figure 4.5:** The difference between the application of the renewed parameterizations $\phi_Q$ and $\phi_I$ to the problem $(A + \alpha B)u = \beta b$.

then it solves also the equations

$$\begin{pmatrix} A_{11} & A_{12} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \beta b_1$$
$$\begin{pmatrix} V_{21} & V_{22} \end{pmatrix} A \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \beta \begin{pmatrix} V_{21} & V_{22} \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \end{pmatrix}$$

and vice versa. This is also the reason for the exactness of the reduction for $\alpha = 0$, compare figure 4.4.

**Consider now the first possibility.** Now, the performance of the possibility, where the inversion of $A_{22} + B_{22}$ is allowed, will be investigated.

Instead of solving the equation $(A + \alpha B)u = \beta b$, the solution process reads now

$$\begin{aligned} (A_{11} + \alpha B_{11})u_1 - (A_{12} + \alpha B_{12})\psi(u_1) &= \beta b_1 \\ (A_{22} + \alpha B_{22})u_2 &= \beta b_2 - (A_{21} + \alpha B_{21})u_1. \end{aligned} \tag{4.10}$$

The main information obtained from the table is therefore the splitting of the state vector $u$ into a slow and a fast part. The application of the function $\psi$ for the first equation guarantees that the original problem is splitted in two decoupled subsystems.

The solution the equation $(A + \alpha B)u = \beta b$ is by the above solution process approximated by

$$\begin{pmatrix} A_{11} + \alpha B_{11} - (A_{12} + \alpha B_{12})Y_{22}^{-1}Y_{21} & 0 \\ A_{21} + \alpha B_{21} & A_{22} + \alpha B_{22} \end{pmatrix} u = \beta b, \tag{4.11}$$

where $Y_{2i}$ has to be replaced by $A_{2i}$ or $\Lambda_2 V_{2i}$ according to the type of table. The inexactness of this possibility is on the one hand the difference in the Schur complement, which is created with $Y_{22}$ and $Y_{21}$ instead of the original blocks $A_{22} + B_{22}$ and $A_{21} + B_{21}$. On the other hand,

the original solution process considers the effect of $b_2$ to the slow part $u_1$, whereas the reduced slow variables are independent of the fast part of the right hand side.

The errors produced by this solution technique can be seen in figure 4.6. The figure shows

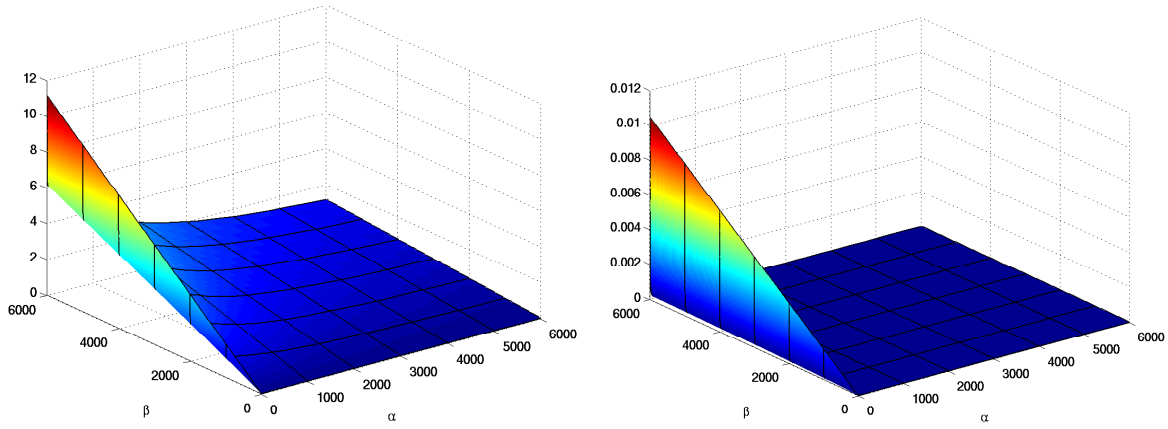


**Figure 4.6:** The left figure shows the difference of the solution process (4.10) to the detailed solution of the problem $(A+\alpha B)u = \beta b$ with matrices and right hand side given by (4.8). The figure on the right hand side visualizes the difference between the application of QSSA and ILDM.

clearly the same behavior as the pure application of $\psi$ in figure 4.3, but the error is smaller.

### 4.2.7 A PDE example

Reconsider the 1d differential equation (4.5) in order to investigate the numerical errors occurring by the application of the different reduction strategies. The exact solutions for the boundaries corresponding to the ILDM–parameterization can be seen in figure 4.7.

Applying the parameterizations $\psi_Q$ and $\psi_I$ to the linear algebraic system reduces first of all the size of the problem. The reduced system reads then

$$(A_{h,11} + B_{h,11})u_{h,1} + (A_{h,12} + B_{h,12})\psi(u_{h,1}) = b_{h,1}, \tag{4.12}$$

where the evaluation of $\psi$ is to be interpreted node–wise. The original $3N \times 3N$–system is now reduced to a system containing only $N$ equations.

As already seen in lemma 4.2.2, the ILDM method reduces the system of equations without systematic errors, which is not the case for QSSA. Therefore the systematic error of the reduction is to be investigated for the QSSA–table and its variations. Also the errors produced by the ILDM method will be calculated in order to see the effect of errors in the evaluation of the tables.

The difference to the exact solution can be seen in table 4.1. This table substantiates clearly the result of lemma 4.2.2. Note that the error for ILDM can be even further reduced by

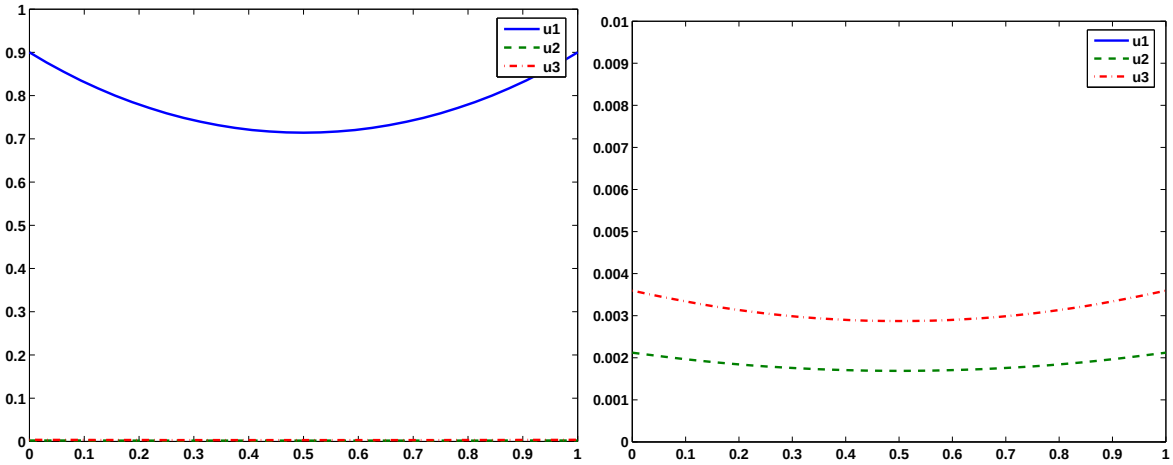


**Figure 4.7:** The detailed solution to problem (4.5) on a grid with 65 nodes. The right figure shows the same plot with a different scale.

| | | analytic param. | | tabulated param. | |
|---|---|---|---|---|---|
| # points | $h$ | QSSA | ILDM | QSSA | ILDM |
| 5 | 0.25 | $2.06e-5$ | $1.39e-9$ | $1.85e-5$ | $4.2e-9$ |
| 17 | 0.0625 | $2.15e-5$ | $1.39e-9$ | $1.93e-5$ | $3.97e-9$ |
| 65 | 0.015625 | $2.16e-5$ | $1.39e-9$ | $1.95e-5$ | $4.73e-9$ |

**Table 4.1:** The systematic error produced by the reduction of problem (4.5) with the QSSA and ILDM method. In both cases, tables containing 1001 points were used, but tables with only 11 points lead basically to the same results.

considering boundary values, which evaluate the parameterizations $\psi$ with higher precision, whereas the error for QSSA remains the same, compare lemma 4.2.2.

Note that the nice performance of the reduction methods depends strongly on the properties of the second operator $B_h$ and the right hand side $b_h$ of the algebraic equation. Here, the matrix $B_h$ does not change "slow" and "fast", therefore the slow variables of the purely chemical problem described by $A_h$ are still considered to be slow in the problem described by $A_h + B_h$. The right hand side $b_h$ plays also an important role. First of all, the parameterization $\psi$ was created for right hand sides with $b_{h,2} = 0$ for the non–boundary nodes, which is in the considered example the case. External sources for the slow process variables can also be treated with the parameterizations $\psi$, but sources for the fast variables lead to big errors. And secondly, the boundary values of the original problem have to be chosen appropriately. Since the values of the fast variables depend algebraically on the slow variables, the boundaries of the original problem have to fulfill the property $u_2 = \psi(u_1)$. If for example the vector $u = (0.9, 0.1, 0.0)$ is taken as boundary conditions, the error is for both QSSA and ILDM approximately 0.098 for all grid sizes.

The choice of the boundaries in the considered example and the partial differential operator, which does conserve the characteristics "slow" and "fast" for the variables ensures the good approximation of the reduced iteration. Taking the boundary values as initial guess for the

Newton iteration leads to an initial residual of order $10^{-1}$ for the process variables and of order $10^{-7}$ for the fast variables. The reduced solution process leads to an approximation, which has a residual of order $10^{-13}$ for the process variables, whereas the fast residual remains constant with $10^{-3}$. These values were produced on a grid with 65 nodes.

Consider now the more expensive possibility and evaluate the tables by

$$\phi(u_1) = \psi(u_1) + A_{22}^{-1} b_2$$

and allow therefore the additional inversion of the block–diagonal matrix $A_{22}$ in each iteration. The solution process reads therefore

$$\begin{aligned}(A_{h,11} + B_{h,11} + (A_{h,12} + B_{h,12})\psi)u_{h,1} &= b_{h,1} - (A_{h,12} + B_{h,12})A_{h,22}^{-1} b_{h,2} \\ u_{h,2} &= \psi(u_{h,1}) + A_{h,22}^{-1} b_{h,2}.\end{aligned}$$

The application of this technique to the global problem does not lead to new results, simply because the fast part of the right hand side is zero as well as the matrix block $B_{21}$. This means that the modified tables have the same effect on the reduction of the PDE as the original tables.

This is different, if not only $A_{22}$, but also the matrix block $A_{22}+B_{22}$ is allowed to be inverted. The reduced solution process for the reaction–diffusion equation (4.5) reads now

$$\begin{aligned}(A_{h,11} + B_{h,11} + (A_{h,12} + B_{h,12})\psi)u_1 &= b_{h,1} \\ (A_{h,22} + B_{h,22})u_{h,2} &= b_{h,2} - (A_{h,21} + B_{h,21})u_{h,1}.\end{aligned}$$

This improves of course the accuracy, because the effect of the fast part of $B_h$ is also treated. The difference between the detailed and the reduced solution is shown in table 4.2. This

| # points | $h$ | QSSA | ILDM |
|---|---|---|---|
| 5 | 0.25 | $5.79e-7$ | $1.05e-8$ |
| 17 | 0.0625 | $5.88e-7$ | $1.10e-8$ |
| 65 | 0.015625 | $5.89e-7$ | $1.10e-8$ |

**Table 4.2:** The difference between the detailed and the reduced solution of (4.5), if in addition to the application of the parameterization $\psi$ the matrix $A_{h,22} + B_{h,22}$ is allowed to be inverted.

table shows clearly that the inversion of $A_{h,22} + B_{h,22}$ improves the application of the parameterizations $\psi$ a lot, if QSSA is applied. The results for ILDM are at least comparable, see also table 4.1. Also of importance is the fact that the error is basically independent of the grid size.

#### Variation of the parameters

Clearly, different parameters for the reaction and different viscosities $\nu$ do not only change the solutions, but also the accuracy of their approximations. Table 4.3 shows the influence of

the viscosities to the accuracy of the reduced solution for three different reaction mechanisms given by

$$sp_1 \underset{1}{\overset{2}{\rightleftharpoons}} sp_2 \underset{90\,k}{\overset{1000\,k}{\rightleftharpoons}} sp_3 \xrightarrow{500\,k} sp_4.$$

The reduction mechanism QSSA was applied to create this table. The influence of the

| $k$ | $\nu = 1$ | $\nu = 10^{-2}$ | $\nu = 10^{-4}$ | $\nu = 1$, ILDM |
|---|---|---|---|---|
| 1 | $\Delta u_1 = 2e-5$<br>$\Delta u_2 = 6e-7$<br>$res_2 = 1e-3$ | $\Delta u_1 = 2e-5$<br>$\Delta u_2 = 6e-7$<br>$res_2 = 3e-4$ | $\Delta u_1 = 4e-6$<br>$\Delta u_2 = 4e-7$<br>$res_2 = 3e-5$ | $\Delta u_1 = 6e-8$<br>$\Delta u_2 = 4e-9$<br>$res_2 = 4e-6$ |
| 10 | $\Delta u_1 = 2e-7$<br>$\Delta u_2 = 7e-9$<br>$res_2 = 1e-5$ | $\Delta u_1 = 2e-7$<br>$\Delta u_2 = 2e-8$<br>$res_2 = 3e-5$ | $\Delta u_1 = 4e-8$<br>$\Delta u_2 = 1e-8$<br>$res_2 = 3e-6$ | $\Delta u_1 = 6e-10$<br>$\Delta u_2 = 4e-10$<br>$res_2 = 7e-7$ |
| 100 | $\Delta u_1 = 3e-9$<br>$\Delta u_2 = 7e-11$<br>$res_2 = 1e-5$ | $\Delta u_1 = 2e-9$<br>$\Delta u_2 = 2e-10$<br>$res_2 = 3e-6$ | $\Delta u_1 = 4e-10$<br>$\Delta u_2 = 2e-10$<br>$res_2 = 3e-7$ | $\Delta u_1 = 8e-12$<br>$\Delta u_2 = 4e-11$<br>$res_2 = 5e-7$ |

**Table 4.3:** The table shows in each block the difference of the detailed to the reduced solutions of the reaction–diffusion equation (4.5). The source term has the eigenvalues on the left hand side, the viscosity constants for the species shall be the same for all three variables. The blocks are divided into three parts, the first two parts are the difference of the slow and fast variables, whereas $res_2$ denotes the residual of the reduced solution for the fast variables.

viscosity is rather small compared to the influence of the spectral gap. The viscosity $\nu = 1$ for all three species in the considered example is much bigger than the viscosities are in general in reality. Smaller values for $\nu$ do not change the performance of the tables significantly, even though the solution differs a lot. Note that the performance is slightly increased, if smaller values for the viscosity of the fast variables are considered, whereas bigger viscosities leads to minor approximations.

Even though lemma 4.2.2 proves the exactness of the reduced solution, if ILDM is applied, still errors occur due to errors for example in the evaluation of the table. These errors are much smaller than the errors computed for the QSSA–reduction, compare the last column of 4.3. The increasing spectral gap seems to have an improving effect on the accuracy.

The only possibility to decrease the performance significantly is to choose the viscosities such that the process variables for the matrix $A_h$ cannot be used as process variables for $A_h + B_h$. This can be achieved by changing the original equation (4.5) for example to

$$\begin{pmatrix} -10^{-3} & 0 & 0 \\ 0 & -10^{2} & 0 \\ 0 & 0 & -10^{2} \end{pmatrix} u''(x) = Au(x), \tag{4.13}$$

which leads to the results in figure 4.8. Clearly, the reduction methods resulting from QSSA cannot be applied anymore. Again due to lemma 4.2.2, the reduced solution obtained by the ILDM method should be exact. But the non-existing spectral gap of $A_h + B_h$ leads to almost the same effect as in the case of QSSA.

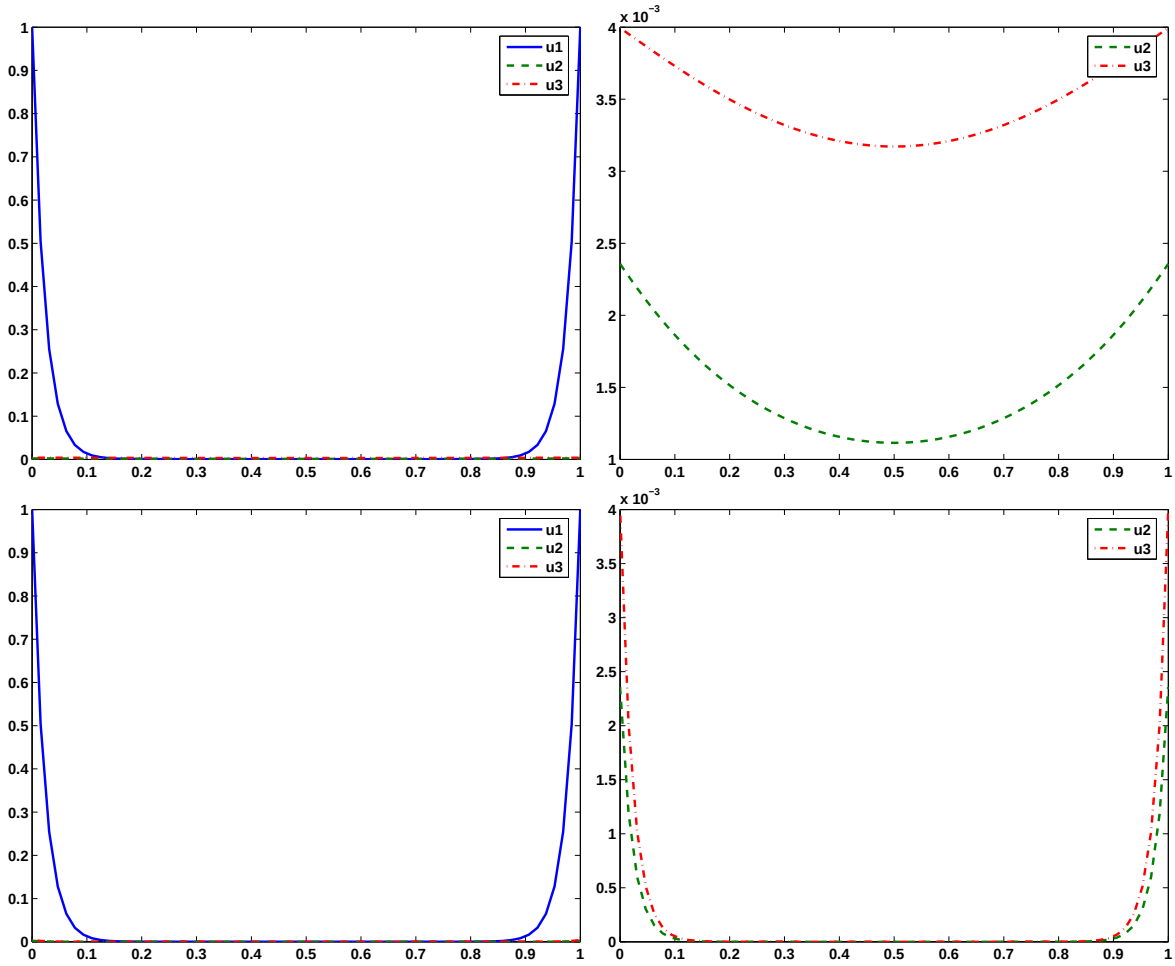


**Figure 4.8:** The solution to problem (4.13). The upper figures show the detailed, the lower two the reduced solution. The two figures on the right hand side show the same plot as the left two figures, but with a different scale.

### 4.2.8 Numerical costs

In order to state the numerical costs for an algorithm, this section is divided into the consideration of the computation time and the calculation of the consumed memory.

The considered system of linear equations shall be denoted by

$$(A + B)u = b,$$

where $A$ and $B$ are the finite element matrices similar as in section 4.2.7. The pattern of the matrix $A + B$ depends obviously on the dimension of the domain of the underlying PDE and the discretization of the system. If the PDE is considered in one dimension in space and the discretization is obtained by the finite element method with the trapezoid rule as integration formula, the matrix $A + B$ is of block–tridiagonal form. The same discretization of a two dimensional domain leads to a matrix with blocks on the diagonal and basically two

off–diagonals. Each block is of size $n_1 + n_2$, the number of blocks on the diagonal equals the size of the grid.

### The computation time for PDEs

Consider the problem

$$-\Delta u = f(u)$$

in two dimensions with $f$ denoting the source term of a linear reaction. Here, this problem is solved by the Newton's method, even though it is purely linear. The solution process includes therefore the computation of the defect and the multilevel solver for the Newton update. The smoothing will be done by a preconditioned Richardson iteration, the preconditioner is provided by the ILU.

Let now the domain be discretized with 1024 cells. The computation time can then be seen in table 4.4 for reactions with 3, 5, 10 and 15 species. The chemical source term describes reactions of the form (4.2) with additional fast variables as intermediate species between $sp_3$ and $sp_4$. The table suggests clearly that the computational expenses for the

| $n_2$ | detailed | reduced | reduced analytic | preconditioned |
|---|---|---|---|---|
| 2 | 3.3 | 35 | 12 | 36 |
| 4 | 44 | 39 | 15 | 58 |
| 9 | 555 | 55 | 24 | 432 |
| 14 | 2657 | 71 | 37 | 2160 |

**Table 4.4:** The computation time in seconds for the problem $-\Delta u = f(u)$ with one process variable and $n_2$ fast variables. The compared solution techniques are the detailed solving strategy without any reduction, the solution process with direct application of the tables, the direct application of the analytically given tables and the preconditioned method, where the inversion of $A_{22} + B_{22}$ is allowed. The computation was done on a Pentium IV with 1.4 GHz and 2 GB memory.

detailed problem grow with more than $n^3$. It also shows that the reduced solution process is very expensive, especially for reaction mechanisms, where the ration between the fast and slow variables $n_2/n_1$ is small. The power of the application of the tables can only be seen, if reaction mechanisms with many fast variables are considered. The computational costs can be reduced, if the parameterizations of the manifolds is given analytically. The preconditioned solver is slightly cheaper than the detailed solution process, but the numerical costs do also grow with $n^3$.

The computation times in more detail can be seen in table 4.5. Clearly, the most time-consuming part of the detailed solution process is the calculation of the ILU, which grows faster than with $n^3$. The assembling of the matrix is relatively cheap, but grows also cubically with the number of species. Only the calculation of the nonlinear residual seems to be of almost the same costs independently of $n$. This behavior changes, if the reduction mechanisms are applied. The calculation of the residual is still almost independent of $n$, but the costs are much higher. The same with the assembling of the matrix, which is quite

| | detailed | | | reduced | | |
|---|---|---|---|---|---|---|
| $n_2$ | matrix | ilu | residual | matrix | ilu | residual |
| 2 | 0.76 | 0.10 | 0.42 | 19.3 | 0.11 | 7.00 |
| 4 | 3.7 | 31.7 | 0.53 | 22.2 | 0.11 | 7.8 |
| 9 | 22.2 | 516 | 0.56 | 31.9 | 0.11 | 11.0 |
| 14 | 79.7 | 2551 | 0.8 | 43.3 | 0.11 | 12.8 |

**Table 4.5:** The computation times of table 4.4 divided into the time to assemble the matrix, to calculate the ILU and to compute the residual in the nonlinear iteration.

expensive, but grows less than linearly in $n$. The costs of the computation of the ILU do of course not change, because the resulting linear system contains for all reaction mechanisms only one variable.

### Memory consumption for PDEs

In order to give a feeling for the memory usage of the solution processes, assume that the matrix is stored as well as three vectors for the solution of a linear system. The three vectors provide space for the right hand side, the iterate $u$ and the current residual. The detailed solver needs therefore

- $5Nn^2[double]$ byte for the matrix ($5N$ blocks containing $n^2$ doubles. Note that the 5 is only correct in two dimensions with lumping.)
- $Nn[double]$ byte for each vector

for a sparse linear system of equations. Let the computation be performed on a computer with 1 GB memory and consider a reaction mechanism similar to the methane–air combustion with $n_1 = 4$ slow and $n_2 = 32$ fast variables. This leads to a maximal grid size of 19 000 nodes. In two dimensions of space, grids with $10^5$ and more grid points are considered for reactive flows.

The reduction of the total system leads to a memory consumption of $(5n_1^2 + 3n_1)N[double]$ bytes in addition to the table. Consider again a reaction mechanism similar to the methane flame. Then grids with up to 950 000 nodes can be considered on a computer with 1 GB memory. Here, the table is assumed to require 300 MB space.

The treatment of the system

$$\begin{aligned}(A_{11} + B_{11} + (A_{12} + B_{12})\psi)u_1 &= b_1\\ (A_{22} + B_{22})u_2 &= b_2 - (A_{21} + B_{21})u_1\end{aligned}$$

instead of the original problem $(A + B)u = b$ leads to two decoupled linear systems and can be treated separately. The solution process for the first equation requires as much disk space as the direct application of the tables, this means $(5n_1^2 + 3n_1)N[double]$ bytes in addition to the space required by the table. The second equation needs $(5n_2^2 + 3n_2)N[double]$ bytes, but the table $\psi$ is not necessary anymore. The limiting size is therefore the size of the second equation. The maximal grid size, which can be considered with a reaction mechanism of the

size of the methane–air reaction, is therefore $N = 24\,000$ nodes, which is approximately 25% more than with detailed solvers.

## 4.3 Linear equations including conservation laws

The goal of this section is to provide a deeper understanding of the effect of center manifolds to reduction methods. The outline of this section is therefore comparable to section 4.2. The main difference to the previous chapter is that analytical results are slightly more tricky to prove. The proof is often based on the fact that $V_3^0$ is invertible, for details see lemma 4.3.1. The existence of the center manifold leads also to a different behavior, if general linear equations are treated, where the reduction methods produce systematic errors. This difference can probably best be seen by comparing figures 4.4 and 4.11.

### 4.3.1 The parameterization of the manifolds

Consider now a linear reaction mechanism, where conservation laws occur. The modeling ODE may then be given in the form

$$\begin{pmatrix} \dot{u_1} \\ \dot{u_2} \\ \dot{u_3} \end{pmatrix} = \underbrace{\begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix}}_{=:A} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix},$$

where $u_3 \in \mathbb{R}^{n_3}$ are the variables, which might be calculated from the process variables $u_1 \in \mathbb{R}^{n_1}$ and the fast variables $u_2 \in \mathbb{R}^{n_2}$ by the conservation laws. The eigenvalue analysis of $A$ leads to

$$V = \begin{pmatrix} V_{11} & V_{12} & V_{13} \\ V_{21} & V_{22} & V_{23} \\ V_1^0 & V_2^0 & V_3^0 \end{pmatrix} \quad \text{and} \quad \Lambda = \begin{pmatrix} \Lambda_1 & 0 & 0 \\ 0 & \Lambda_2 & 0 \\ 0 & 0 & 0 \end{pmatrix}$$

with $VA = \Lambda V$.

The matrix $A$ has several nice properties:

- The rows of
$$\begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \end{pmatrix}$$
are linearly independent.

- The rows of
$$\begin{pmatrix} A_{31} & A_{32} & A_{33} \end{pmatrix}$$
are linearly independent and depend linearly on the rows of $\begin{pmatrix} A_{11} & A_{12} & A_{13} \end{pmatrix}$ and $\begin{pmatrix} A_{21} & A_{22} & A_{23} \end{pmatrix}$, therefore there exist matrices $\Gamma_1 \in \mathbb{R}^{n_3 \times n_1}$ and $\Gamma_2 \in \mathbb{R}^{n_3 \times n_2}$ such that
$$\Gamma_1 \begin{pmatrix} A_{11} & A_{12} & A_{13} \end{pmatrix} + \Gamma_2 \begin{pmatrix} A_{21} & A_{22} & A_{23} \end{pmatrix} + \begin{pmatrix} A_{31} & A_{32} & A_{33} \end{pmatrix} = 0.$$

- The matrix $V_3^0$ is invertible, as the following lemma shows.

**Lemma 4.3.1** *Let the rows of*

$$\begin{pmatrix} V_1^0 & V_2^0 & V_3^0 \end{pmatrix}$$

*be the left eigenvectors to the eigenvalue $\lambda = 0$ of the block matrix*

$$A = \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ A_{31} & A_{32} & A_{33} \end{pmatrix}$$

*of size $(n_1 + n_2 + n_3) \times (n_1 + n_2 + n_3)$, where the last $n_3$ rows are linearly independent and depend linearly on the upper $n_1 + n_2$ rows. Then $V_3^0$ is invertible.*

**Proof:** The dimension of the eigenspace to the eigenvalue $\lambda = 0$ is

$$\dim E(0) = \dim \ker(A) = n_3,$$

therefore do the rows of $\begin{pmatrix} V_1^0 & V_2^0 & V_3^0 \end{pmatrix}$ form a basis of $E(0)$. So do the rows of

$$\begin{pmatrix} \Gamma_1 & \Gamma_2 & \mathbb{1} \end{pmatrix}.$$

This implies the existence of an invertible matrix $\mathcal{M} \in \mathbb{R}^{n_3 \times n_3}$ with

$$\begin{pmatrix} \Gamma_1 & \Gamma_2 & \mathbb{1} \end{pmatrix} = \mathcal{M} \begin{pmatrix} V_1^0 & V_2^0 & V_3^0 \end{pmatrix},$$

hence

$$\mathbb{1} = \mathcal{M} V_3^0.$$

The invertibility of $\mathcal{M}$ implies now the invertibility of $V_3^0$. ∎

The above formulation of the ODE is equivalent to the DAE

$$\begin{pmatrix} \dot{u}_1 \\ \dot{u}_2 \\ c \end{pmatrix} = \underbrace{\begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ V_1^0 & V_2^0 & V_3^0 \end{pmatrix}}_{=:\widetilde{A}} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}, \tag{4.14}$$

given that the rows of the matrix

$$\begin{pmatrix} V_1^0 & V_2^0 & V_3^0 \end{pmatrix}$$

span the left eigenspace of the matrix $A$ to the eigenvalue $\lambda = 0$ and that $c$ describes the conservation constant of the initial value

$$c = \begin{pmatrix} V_1^0 & V_2^0 & V_3^0 \end{pmatrix} \cdot u(0).$$

The QSSA–parameterization $\psi_Q$ and the ILDM–parameterization $\psi_I$ are then described by the functions

$$\psi_Q(u_1) = \begin{pmatrix} A_{22} & A_{23} \\ V_2^0 & V_3^0 \end{pmatrix}^{-1} \left( \begin{pmatrix} 0 \\ c \end{pmatrix} - \begin{pmatrix} A_{21} \\ V_1^0 \end{pmatrix} u_1 \right) \tag{4.15}$$

and

$$\psi_I(u_1) = \begin{pmatrix} \Lambda_2 V_{22} & \Lambda_2 V_{23} \\ V_2^0 & V_3^0 \end{pmatrix}^{-1} \left( \begin{pmatrix} 0 \\ c \end{pmatrix} - \begin{pmatrix} \Lambda_2 V_{21} \\ V_1^0 \end{pmatrix} u_1 \right).$$

As an example, take the mechanism

$$sp_1 \underset{1}{\overset{2}{\rightleftharpoons}} sp_2 \underset{90}{\overset{1000}{\rightleftharpoons}} sp_3, \tag{4.16}$$

and let the mass fractions of all three species $sp_1$, $sp_2$ and $sp_3$ be modeled. The source term is therefore given by the matrix

$$\begin{pmatrix} -2 & 1 & 0 \\ 2 & -1001 & 90 \\ 0 & 1000 & -90 \end{pmatrix}, \tag{4.17}$$

and the mass fraction of species $sp_3$ can be calculated by the conservation law $y_1+y_2+y_3 = 1$. The parameterizations are then given by

$$\psi_Q(u_1) = \begin{pmatrix} 0.082493 \\ 0.917507 \end{pmatrix} - \begin{pmatrix} 0.08066 \\ 0.91934 \end{pmatrix} u_1$$

and

$$\psi_I(u_1) = \begin{pmatrix} 0.082499 \\ 0.917501 \end{pmatrix} - \begin{pmatrix} 0.0808141 \\ 0.919186 \end{pmatrix} u_1.$$

In this section, the problem

$$(A + \alpha B)u = \beta b$$

will be treated with the above parameterizations, which were introduced for the problem $\dot{u} = Au$. The introduction of the matrix $B$ together with the disturbance $b$ leads to the problem that the original conservation laws do not have to be valid anymore. The conservation laws are only preserved, if both equations

$$\begin{pmatrix} V_1^0 & V_2^0 & V_3^0 \end{pmatrix} B = 0 \quad \text{and} \quad \begin{pmatrix} V_1^0 & V_2^0 & V_3^0 \end{pmatrix} b = 0$$

hold, which is in general not the case. But due to the existence of the operator $B$, the failure of these conditions does not contradict the existence of a steady state of the ODE $\dot{u} = (A + \alpha B)u - \beta b$.

In order to simplify the investigation of the influence of the center manifold, note the following technical lemma.

**Lemma 4.3.2** *Let $A$ be a block matrix given by*

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}$$

*with $A_{11}$ and $A_{22}$ invertible blocks. If $M := A_{11} - A_{12}A_{22}^{-1}A_{21}$ exists and is invertible, the inverse of $A$ is given by*

$$A^{-1} = \begin{pmatrix} M^{-1} & -M^{-1}A_{12}A_{22}^{-1} \\ -A_{22}^{-1}A_{21}M^{-1} & A_{22}^{-1}\left(\mathbb{1} + A_{21}M^{-1}A_{12}A_{22}^{-1}\right) \end{pmatrix}.$$

### 4.3.2 Description of reduced problems

The application of the QSSA–parameterization $\psi_Q$ introduces the conservation laws of $\dot{u} = Au$ to the problem $\dot{u} = (A + \alpha B)u - \beta b$. The reduced problem to $(A + \alpha B)u = b$ reads now

$$(A_{11} + \alpha B_{11}) + \begin{pmatrix} A_{12} + \alpha B_{12} & A_{13} + \alpha B_{13} \end{pmatrix} \psi(u_1) = b_1$$

for the process variables $u_1$ and with the evaluation of $\psi(u_1)$ also to $u_2$ and $u_3$. By lemma 4.3.2, the reduced problem is clearly equivalent to

$$\begin{pmatrix} A_{11} + \alpha B_{11} & A_{12} + \alpha B_{12} & A_{13} + \alpha B_{13} \\ A_{21} & A_{22} & A_{23} \\ V_1^0 & V_2^0 & V_3^0 \end{pmatrix} u = \begin{pmatrix} \beta b_1 \\ 0 \\ c \end{pmatrix}$$

with $c$ denoting the conservation constant

$$c = \begin{pmatrix} V_1^0 & V_2^0 & V_3^0 \end{pmatrix} u(0)$$

for the original problem $\dot{u} = Au$.

A similar analysis for the reduction with the ILDM–parameterization $\psi_I$ leads to the equations

$$\begin{pmatrix} A_{11} + \alpha B_{11} & A_{12} + \alpha B_{12} & A_{13} + \alpha B_{13} \\ \Lambda_2 V_{21} & \Lambda_2 V_{22} & \Lambda_2 V_{23} \\ V_1^0 & V_2^0 & V_3^0 \end{pmatrix} u = \begin{pmatrix} \beta b_1 \\ 0 \\ c \end{pmatrix},$$

which describe the solution, if the original problem is reduced with ILDM. Even though these two linear systems differ in the description of the fast variables, they are equivalent, if $\alpha = \beta = 0$, as the following lemma shows.

**Lemma 4.3.3** *Let $A$ and $V_i^0$ be as above. Then the equivalence*

$$\begin{pmatrix} A_{11} & A_{12} & A_{13} \\ A_{21} & A_{22} & A_{23} \\ V_1^0 & V_2^0 & V_3^0 \end{pmatrix} u = \begin{pmatrix} 0 \\ 0 \\ c \end{pmatrix} \iff \begin{pmatrix} A_{11} & A_{12} & A_{13} \\ \Lambda_2 V_{21} & \Lambda_2 V_{22} & \Lambda_2 V_{23} \\ V_1^0 & V_2^0 & V_3^0 \end{pmatrix} u = \begin{pmatrix} 0 \\ 0 \\ c \end{pmatrix}$$

*holds.*

**Proof:** Let $\begin{pmatrix} A_{11} & A_{12} & A_{13} \end{pmatrix} u = 0$. Then

$$\begin{aligned} &\begin{pmatrix} \Lambda_2 V_{21} & \Lambda_2 V_{22} & \Lambda_2 V_{23} \end{pmatrix} u \\ &\quad = \begin{pmatrix} V_{21} & V_{22} & V_{23} \end{pmatrix} Au \\ &\quad = \begin{pmatrix} V_{22} \begin{pmatrix} A_{21} & A_{22} & A_{23} \end{pmatrix} + V_{23} \begin{pmatrix} A_{31} & A_{32} & A_{33} \end{pmatrix} \end{pmatrix} u \end{aligned}$$

By lemma 4.3.1 we get

$$\begin{pmatrix} A_{31} & A_{32} & A_{33} \end{pmatrix} = -\left(V_3^0\right)^{-1} \left( V_1^0 \begin{pmatrix} A_{11} & A_{12} & A_{13} \end{pmatrix} + V_2^0 \begin{pmatrix} A_{21} & A_{22} & A_{23} \end{pmatrix} \right),$$

therefore we have

$$\begin{pmatrix} \Lambda_2 V_{21} & \Lambda_2 V_{22} & \Lambda_2 V_{23} \end{pmatrix} u = \left( V_{22} - V_{23} \left(V_3^0\right)^{-1} V_2^0 \right) \begin{pmatrix} A_{21} & A_{22} & A_{23} \end{pmatrix} u,$$

which finishes the proof. ∎

Theorem 4.2.1 can be reformulated for the case, where $A$ is not invertible.

**Theorem 4.3.4** *Let $A \in \mathbb{R}^{n\times n}$ be arbitrary. Define the function $\psi : \mathbb{R}^{n_1} \to \mathbb{R}^{n_2+n_3}$ such that the vector $u = (u_1, \psi(u_1))^T$ is on the manifold defined by QSSA (ILDM) for all values of $u_1$. Let further $B \in \mathbb{R}^{n\times n}$ and $r \in \mathbb{R}^n$ are such that $u$ is on the QSSA–(ILDM–)manifold for the function $f(u) = Bu - r$. If $A + B$ is invertible, then the solution of the reduced problem*

$$(A_{11} + B_{11} + \begin{pmatrix} A_{12} + B_{12} & A_{13} + B_{13} \end{pmatrix} \psi)u_1 = r_1$$
$$\begin{pmatrix} u_2 \\ u_3 \end{pmatrix} = \psi(u_1)$$

*is also a solution as the original problem $(A+B)u = r$. If $A+B$ is not invertible, the reduced solution equals the exact solution of $(A + B)u = r$ under the condition $V^0u = c$.*

**Proof:** The proof will be split into two cases. In the first case, $\psi$ is assumed to be created by the QSSA–method, in the second case, $\psi$ represents the ILDM–parameterization.

1. Let $\psi$ represent the manifold created by QSSA. The fact that $\psi$ is also the QSSA–parameterization for the problem $\dot{u} = Bu-r$, leads to the relations $B_{21}u_1+B_{22}u_2+B_{23}u_3 = r_2$ and $V^0Bu = V^0r$, if $u$ is on the manifold. Let $u = (u_1, u_2, u_3)^T$ be the solution of the reduced problem, where the QSSA–parameterization was applied. Then $u$ is necessarily on the manifold and the following holds:

$$\begin{pmatrix} \mathbb{1} & 0 & 0 \\ 0 & \mathbb{1} & 0 \\ V_1^0 & V_2^0 & V_3^0 \end{pmatrix} (A + B)u = \begin{pmatrix} r_1 \\ B_{21}u_1 + B_{22}u_2 + B_{23}u_3 \\ V^0Bu \end{pmatrix} = \begin{pmatrix} \mathbb{1} & 0 & 0 \\ 0 & \mathbb{1} & 0 \\ V_1^0 & V_2^0 & V_3^0 \end{pmatrix} \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix},$$

   because $\begin{pmatrix} A_{21} & A_{22} & A_{23} \end{pmatrix} u = V^0Au = 0$. The invertibility of $V_3^0$ (lemma 4.3.1) implies the equation $(A + B)u = r$, the reduced solution does therefore solve the detailed problem.

2. Let the function $\psi$ now represent the ILDM–parameterization for the problems

$$\dot{u} = Au \quad \text{and} \quad \dot{u} = Bu - r.$$

   This implies the relations

$$V_2(Bu - r) = 0 \quad \text{and} \quad V^0(Bu - r) = 0$$

   for all possible $u$ on the manifold. Let now $u$ be the solution of the reduced problem. Then $u$ is necessarily on the manifold and the following is true:

$$\begin{pmatrix} \mathbb{1} & 0 & 0 \\ V_{21} & V_{22} & V_{23} \\ V_1^0 & V_2^0 & V_3^0 \end{pmatrix} (A + B)u = \begin{pmatrix} r_1 \\ V_2Bu \\ V_0Bu \end{pmatrix} = \begin{pmatrix} \mathbb{1} & 0 & 0 \\ V_{21} & V_{22} & V_{23} \\ V_1^0 & V_2^0 & V_3^0 \end{pmatrix} r,$$

   because $V_2Au = 0$ and $V^0A = 0$. The existence of the ILDM–parameterization implies the invertibility of the matrix $\begin{pmatrix} V_{22} & V_{23} \\ V_2^0 & V_3^0 \end{pmatrix}$, which leads to the relation $(A + B)u = r$.

The linearity of the problems does in both cases assure the equivalence of the detailed and reduced formulations. ■

A few remarks to this theorem:

- In case of QSSA, the reduced solution is clearly exact, if $B_{2i} = 0$, $B_{3i} = V_i^0 - A_{3i}$, $b_2 = 0$ and $b_3 = c$.

- Exactness is also obtained, if the fast variables are decoupled from the slow and conservation variables. This means that $A_{21} = B_{21} = 0$ and $A_{23} = B_{23} = 0$ together with $B_{3i} = V_i^0 - A_{3i}$, $b_2 = 0$ and $b_3 = c$ is required.
- The conditions of the above theorem are fulfilled, if $AB = BA$ and the right hand side $r$ is on the ILDM manifold.
- The conditions are especially fulfilled for ILDM, if

  $$V_2 r = 0 \quad \text{and} \quad V^0 B u = c$$

  for all $u$ on the manifold.

### 4.3.3 Reducing ordinary differential equations

Similar as in section 4.2.4 it can be shown with theorem 4.3.4 that a linear ODE

$$\dot{u} = Au$$

with initial values on the manifold is exactly solved by the ILDM method, if the implicit Euler algorithm is applied. The QSSA method introduces systematic errors. A closer look to the argumentation in section 4.2.4 shows that the non–existing of the center manifolds was not explicitly required, the introduced arguments are therefore also valid for the current case with conservation laws and will not be repeated here.

The behavior of the systematic error introduced by the QSSA method is still interesting compared to ODEs without conservation laws. The difference between the detailed solver and the solutions obtained from QSSA for the problem

$$\begin{aligned} \dot{u} &= Au \\ u(0) &= \begin{pmatrix} 0.9 \\ \psi(0.9) \end{pmatrix} \end{aligned}$$

for $A$ as in (4.17) is shown in figure 4.9. The figure depicts clearly, that the error decreases fastly unlike in the case, where no center manifold existed, compare figure 4.2.

### 4.3.4 Reduction of a simple reaction–diffusion equation

Similar as in the case without center manifolds (compare section 4.2.5), equations of the form

$$-\nu u''(x) = Au(x), \quad x \in \Omega := ]0,1[ \tag{4.18}$$

with $A$ for example as in (4.17), can be exactly solved with the ILDM method, as the following lemma shows.

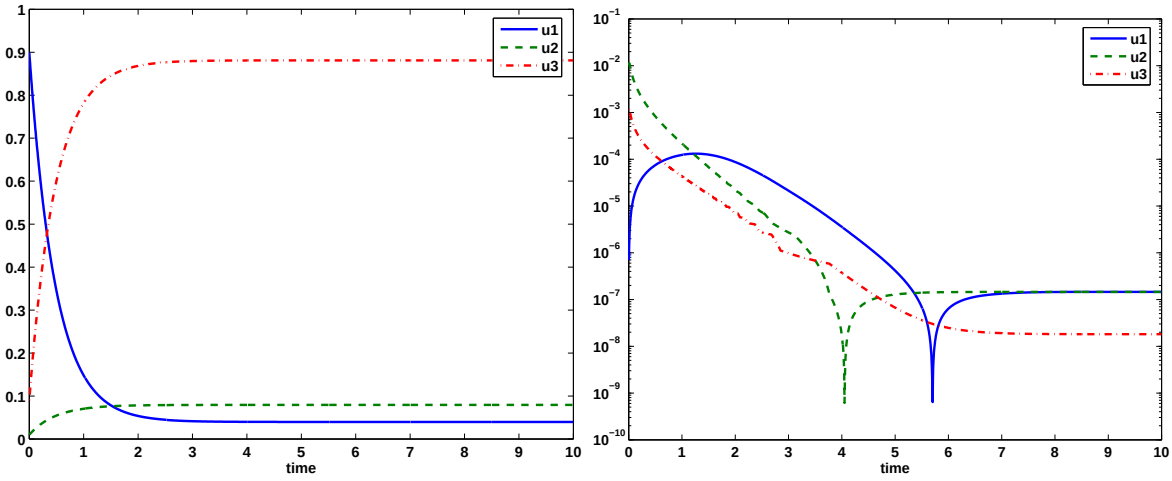


**Figure 4.9:** The solution (left figure) and the relative difference between the exact and QSSA reduced solution of $\dot{u} = Au$, where $A$ defines the source term of reaction (4.16).

**Lemma 4.3.5** *Let the matrix $A$ describe a linear chemical reaction with conservation laws. The rows of $\begin{pmatrix} V_{21} & V_{22} & V_{23} \end{pmatrix}$ describe the left eigenvectors of the fast eigenvalues of $A$, the rows of $\begin{pmatrix} V_1^0 & V_2^0 & V_3^0 \end{pmatrix}$ are the eigenvectors to $\lambda = 0$. Then the reduction of*

$$\left[ \begin{pmatrix} \mathbb{1} & & & \\ & A & & \\ & & \ddots & \\ & & & A & \\ & & & & \mathbb{1} \end{pmatrix} + \begin{pmatrix} 0 & & & & \\ -b/2\mathbb{1} & b\mathbb{1} & -b/2\mathbb{1} & & \\ & \ddots & \ddots & \ddots & \\ & & -b/2\mathbb{1} & b\mathbb{1} & -b/2\mathbb{1} \\ & & & & 0 \end{pmatrix} \right] u = \begin{pmatrix} r \\ 0 \\ \vdots \\ 0 \\ r \end{pmatrix}$$

*by the ILDM–parameterization leads to the same solution $u$ as the original problem, if $r$ fulfills the relation $(r_2, r_3) = \psi(r_1)$.*

**Proof:** The lemma can be proven inductively. Clearly, the exact solution on the first and last node is on the manifold. The solution on the inner nodes fulfill the equation

$$(A + b)u = b/2(u^{upper} + u^{lower}),$$

which is solved exactly by the ILDM method, if $u^{upper}$ and $u^{lower}$ are on the manifold, because then

$$\begin{pmatrix} V_{21} & V_{22} & V_{23} \end{pmatrix} (u^{upper} + u^{lower}) = 0$$

and

$$b/2 \begin{pmatrix} V_1^0 & V_2^0 & V_3^0 \end{pmatrix} (u^{upper} + u^{lower}) = bc,$$

compare the remarks of theorem 4.3.4. ■

Even though this lemma and its proof reminds on lemma 4.2.2, there is a major difference. In the case of linear reactions without conservation laws, the manifold created by ILDM is

linear, this means that $u + v$ is on the manifold, if both $u$ and $v$ are on the manifold as well. Here, the ILDM–manifold is only affine linear. If $u$ and $v$ are on the manifold, then $ku + (1 - k)v$ is on the manifold, but $u + v$ definitively not. This leads to the necessary condition for the disturbance matrix $B$ that the sum of the elements in each row has to be zero. Fortunately, this is the case for the discretized Laplacian.

By similar arguments as in section 4.2.5 it can be seen that the QSSA method does not lead to correct results. The produced systematic error will be analyzed in section 4.3.6.

### 4.3.5 Reduction of general equations

The previous sections showed clearly that there are (relevant) examples, which are exactly solved by the reduction methods. Now the produced errors for an arbitrary problem will be studied with an example.

For general disturbances $B$ and $b$ of the original problem, exactness cannot be expected and systematic errors occur. Figure 4.10 shows the error, which is produced by the application of the reduction $\psi$ to the system $(A + \alpha B)u = \beta b$ with

$$A = \begin{pmatrix} -2 & 1 & 0 \\ 2 & -1001 & 90 \\ 0 & 1000 & -90 \end{pmatrix}, \quad B = \begin{pmatrix} -2 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{pmatrix} \quad \text{and} \quad b = \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix}. \tag{4.19}$$

Here, $A$ describes the source term of the linear reaction (4.16). The figure suggests that the

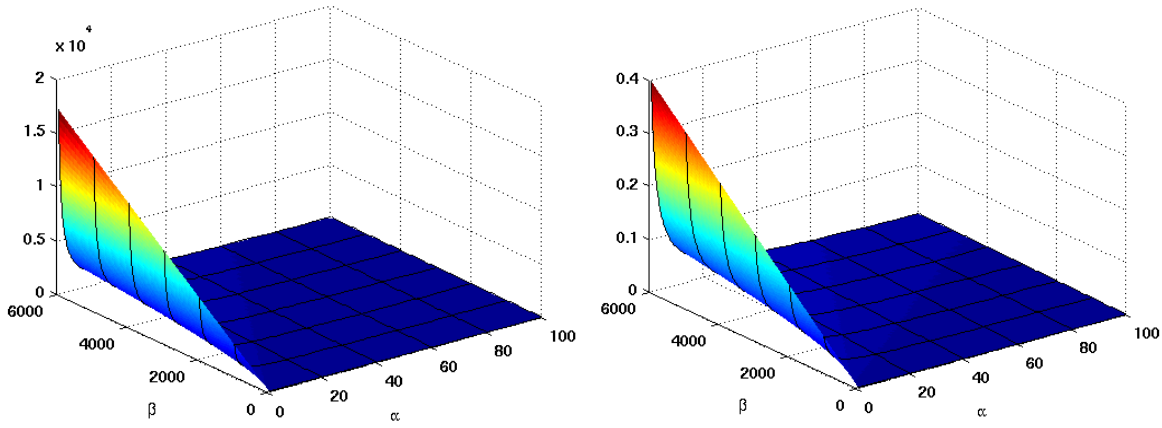


**Figure 4.10:** The left figure shows the error produced by the application of the ILDM–parameterization $\psi_I$ to problem $(A + \alpha B)u = \beta b$ defined by (4.19). The right figure shows the difference between the solutions obtained by the application of ILDM and the application of QSSA.

influence of the entries of $b$ is much bigger than of $B$. In fact, the error depends linearly on $\beta$, whereas the performance seems to increase with increasing influence of the second operator $B$.

Compared to the application of the parameterizations to problems without conservation laws, the error is enormously larger, especially for small values of $\alpha$, compare figure 4.3. This is not surprising, because the matrix $A+\alpha B$ becomes singular for $\alpha = 0$. But also for bigger $\alpha$ the difference is remarkable. Whereas the error tends to zero, if no conservation laws hold, the error is here bounded away from zero. In fact, the difference between the detailed and reduced solution is for $\alpha = 6000$ and $\beta = 1$ with $\Delta u \approx 4.2 \cdot 10^{-4}$ rather small in case of no conservation equations, whereas it is $\Delta u \approx 0.89$ for the problem with conservation laws.

#### Create a modified parameterization by the inversion of the chemical part

The original parameterization was created to solve the equation

$$\begin{pmatrix} A_{21} \\ V_1^0 \end{pmatrix} u_1 + \begin{pmatrix} A_{22} & A_{23} \\ V_2^0 & V_3^0 \end{pmatrix} \psi(u_1) = \begin{pmatrix} 0 \\ c \end{pmatrix}.$$

Now the parameterization shall be modified, such that information about the right hand side $b$ is contained in this formulation. As already seen in section 4.2.6, the consideration of the fast part of the matrix $B$ leads to a numerically very expensive calculation, so only the right hand side $b$ will be treated in the creation of a new parameterization.

The right hand side of the above equation may be changed to

- $\begin{pmatrix} \beta b_2 \\ c \end{pmatrix}$: The parameterization contains more information about the right hand side as the original parameterization and accepts the existence of the conservation laws. In case of ILDM, replace $b_2$ by $V_{21}b_1 + V_{22}b_2 + V_{23}b_3$.
- $\begin{pmatrix} 0 \\ 0 \end{pmatrix}$: The resulting parameterization omits the constant $c$. The conservation laws are therefore changed, but due to the existence of $B$, they might be invalid anyway.
- $\begin{pmatrix} \beta b_2 \\ 0 \end{pmatrix}$: This parameterization does also omit the center manifold, but contains more information on the right hand side of the problem as the parameterization introduced before. If an ILDM–parameterization is to be created, replace again $b_2$ by $V_{21}b_1 + V_{22}b_2 + V_{23}b_3$.

All three parameterizations are generated on the basis of $\psi$ by

$$\phi(u_1) = \psi(u_1) + \begin{pmatrix} Y_{22} & Y_{23} \\ V_2^0 & V_3^0 \end{pmatrix}^{-1} \begin{pmatrix} rhs_1 \\ rhs_2 - c \end{pmatrix}$$

with the values of $rhs_1$ and $rhs_2$ as above and $Y_{2i} = A_{2i}$ for QSSA–parameterizations and $Y_{2i} = \Lambda_2 V_{2i}$, if an ILDM–parameterization is modified.

These three possibilities lead to solutions $u$ of the reduced problem, which solve the equation

$$\begin{pmatrix} A_{11} + \alpha B_{11} & A_{12} + \alpha B_{12} & A_{13} + \alpha B_{13} \\ Y_{21} & Y_{22} & Y_{23} \\ V_1^0 & V_2^0 & V_3^0 \end{pmatrix} u = \begin{pmatrix} \beta b_1 \\ \begin{pmatrix} rhs_1 \\ rhs_2 \end{pmatrix} \end{pmatrix}.$$

They all neglect the effect of the second operator $B$ to the fast variables, because the formulation depends otherwise on the inverse of the fast and conservation part of $A+\alpha B$, similar as in (4.9), which is too expensive for practical purposes.

The absence of $B_{2i}$ and $B_{3i}$ leads to large errors, see figure 4.11. The difference between

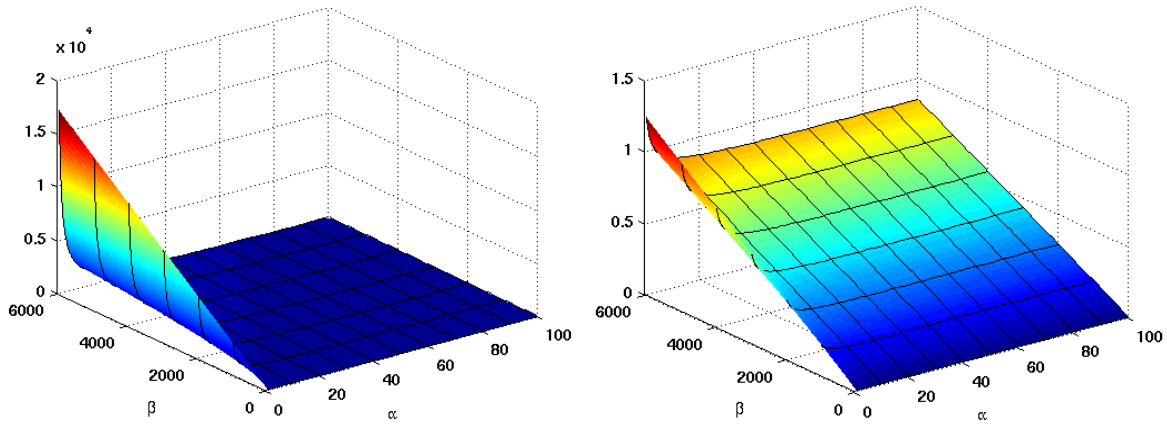


**Figure 4.11:** The error produced by the reduction of the problem $(A+\alpha B)u = \beta b$ with data as in (4.19) and the modified table with $(\beta b_2, c)^T$ on the right hand side.

the reduced solution with the original parameterization and the reduced solution with the renewed parameterizations is small, and no difference is visible in a plot. But there are differences, especially for large values of $\alpha$. Let for example $\alpha = 6000$ and $\beta = 1$ in the problem $(A+\alpha B)u = \beta b$ with $A$ and $B$ defined by (4.19). Then the difference of the reduced and detailed solutions is

- Original parameterization: $\Delta u \approx 0.8859$
- Parameterization considering the right hand side with conservation constant: $\Delta u \approx 0.8858$
- Parameterization neglecting both the right hand side and the conservation constant: $\Delta u \approx 4.8 \cdot 10^{-4}$
- Parameterization neglecting the conservation constant but considering the right hand side: $\Delta u \approx 4.9 \cdot 10^{-4}$

These results suggest clearly that it is advantageous to omit the conservation constant $c$.

The omitting of the conservation constant $c$ leads to a difference, which is constant in $\alpha$ and $\beta$:

$$\Delta u = \begin{pmatrix} A_{11}+\alpha B_{11} & A_{12}+\alpha B_{12} & A_{13}+\alpha B_{13} \\ Y_{21} & Y_{22} & Y_{23} \\ V_1^0 & V_2^0 & V_3^0 \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ 0 \\ c \end{pmatrix}$$

with $|\Delta u| \approx 0.89$ in example (4.19), if the ILDM–parameterization is used. Considering also the fast part of the right hand side $b_2$ or $V_{21}b_1 + V_{22}b_2 + V_{23}b_3$ respectively, leads to a

difference to the solution with the original parameterization, which is still constant in $\alpha$ and depends linearly on $\beta$:

$$\Delta u = \begin{pmatrix} A_{11}+\alpha B_{11} & A_{12}+\alpha B_{12} & A_{13}+\alpha B_{13} \\ Y_{21} & Y_{22} & Y_{23} \\ V_1^0 & V_2^0 & V_3^0 \end{pmatrix}^{-1} \begin{pmatrix} 0 \\ \beta b_2 \\ 0 \end{pmatrix}$$

with $|\Delta u| < 0.1$, if the ILDM–parameterization is applied to example (4.19) with $\alpha, \beta < 6000$. This means that no difference to figure 4.10 will be visible, if the error produced by the application of the renewed parameterization $\phi$ to this problem is plotted.

#### Create a modified parameterization by the inversion of the chemical and physical part

Assume now that equations of the type

$$\begin{pmatrix} A_{22}+B_{22} & A_{23}+B_{23} \\ A_{32}+B_{32} & A_{33}+B_{33} \end{pmatrix} u = b_2$$

may be solved in the solution process. The solution strategy for the problem $(A+\alpha B)u = \beta b$ reads then

$$\begin{aligned} (A_{11}+\alpha B_{11})u_1 + \begin{pmatrix} A_{12}+\alpha B_{12} & A_{13}+\alpha B_{13} \end{pmatrix} \psi(u_1) &= \beta b_1 \\ \begin{pmatrix} A_{22}+\alpha B_{22} & A_{23}+\alpha B_{23} \\ A_{32}+\alpha B_{32} & A_{33}+\alpha B_{33} \end{pmatrix} \begin{pmatrix} u_2 \\ u_3 \end{pmatrix} &= \beta \begin{pmatrix} b_2 \\ b_3 \end{pmatrix} - \begin{pmatrix} A_{21}+\alpha B_{21} \\ A_{31}+\alpha B_{31} \end{pmatrix} u_1. \end{aligned}$$

The information of the table is therefore only used to decouple the first equation from the fast variables.

The error produced by this solution strategy for the problem $(A+\alpha B)u = \beta b$ can be seen in figure 4.12. Clearly, the performance is much better than in the cases, where the fast variables are not calculated detailedly, compare figures 4.10 and 4.11.

### 4.3.6 A PDE example

Reconsider the 1d reaction–diffusion equation given by (4.18) with the source term for the reaction (4.16). Here, the conservation law $\sum_i u_i = 1$ is respected in the parameterizations. The solution to the exact equation with the boundary values $u = (0.9, \psi(0.9))$ for $x = 0$ and $x = 1$ can be seen in figure 4.13.

Similar as in the case without conservation laws, the existence of the matrix $B_h$ describing the discretized Laplacian introduces errors, which can be seen in table 4.6, where the difference between the detailed and reduced solution is shown for analytically given parameterizations and for tables, which introduce an additional interpolation error. The smallness of the errors produced by ILDM is explained by lemma 4.3.5. Clearly, the performance can be even further improved, if the function $\psi_I$ is evaluated with higher accuracy in order to obtain suitable boundary conditions. Similar as in the case without conservation laws, the nice behavior of the two reduction methods is destroyed for boundaries, which are not on the manifold.

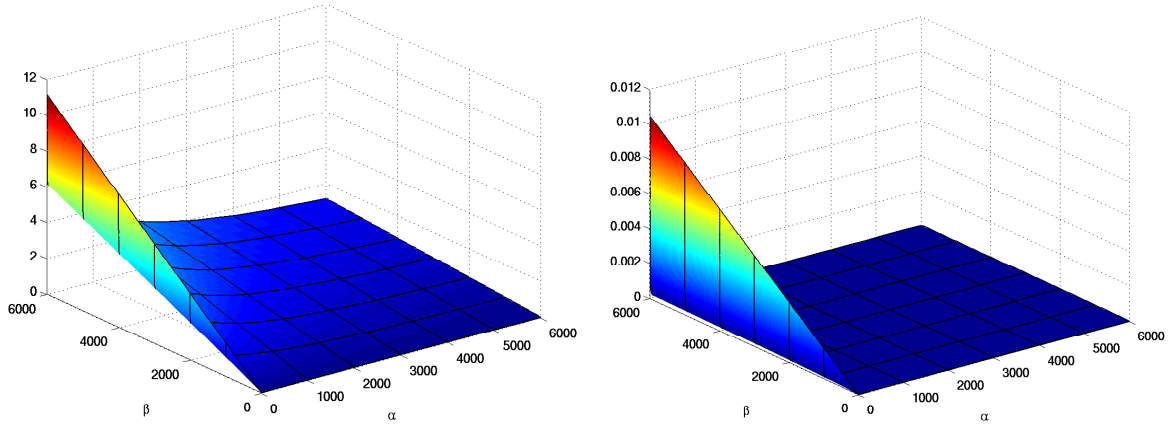


**Figure 4.12:** The error produced by the solution technique with higher effort is shown in the left figure. The difference between the application of the QSSA method instead of ILDM can be seen on the right hand side.

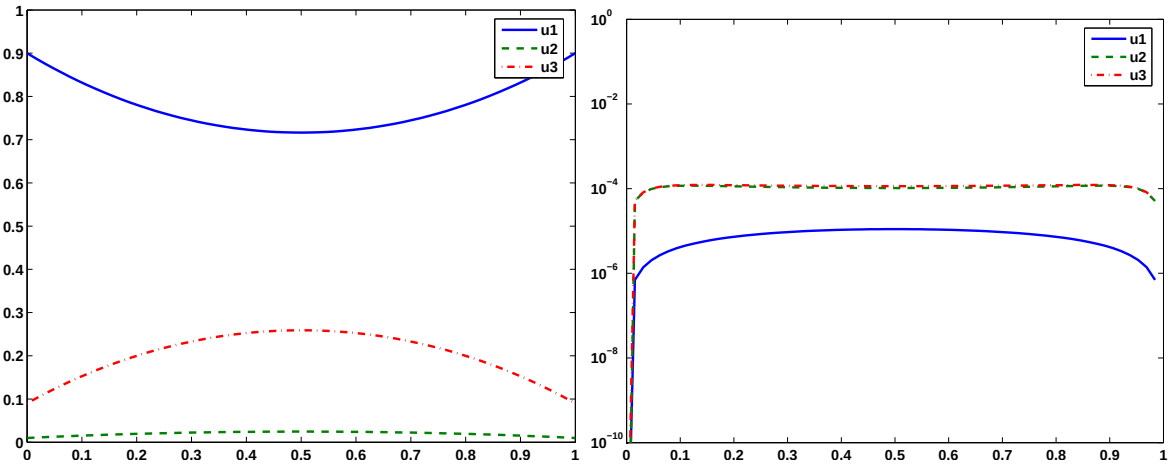


**Figure 4.13:** The solution to the problem $-u''(x) = Au(x)$ with corresponding boundary conditions on a grid with 65 nodes. The matrix $A$ describes the source term of reaction (4.16). The right figure shows the error, if the problem is reduced with QSSA.

| | | analytic param. | | tabulated param. | |
|---|---|---|---|---|---|
| # points | $h$ | QSSA | ILDM | QSSA | ILDM |
| 5 | 0.25 | $1.17e-4$ | $1.39e-9$ | $1.33e-4$ | $2.68e-7$ |
| 17 | 0.0625 | $1.21e-4$ | $1.39e-9$ | $1.33e-4$ | $3.30e-7$ |
| 65 | 0.015625 | $1.22e-4$ | $1.39e-9$ | $1.33e-4$ | $3.86.e-7$ |

**Table 4.6:** The systematic error produced by the reduction of problem (4.18) with reaction 4.16 with the parameterizations $\psi_Q$ and $\psi_I$ in the $L^\infty$–norm. The values are obtained by comparing the detailed solution and the reduced solution with analytically given and tabulated parameterizations.

## 4.4 Nonlinear equations without conservation laws

The results of the linear case presented in section 4.2 will be extended to nonlinear reaction mechanisms in this section. This is unfortunately not very successful, because the equivalent to theorem 4.2.1 raises only very few examples, which are reduced exactly by QSSA and ILDM. Unlike in the linear case, ODEs can for example only be reduced with systematic errors. The same is true for even very simple reaction–diffusion equations. Therefore the occurring errors are investigated in this section for algebraic and differential equations. It turns out that QSSA and ILDM show a similar performance.

### 4.4.1 The parameterization of the manifolds

Consider a system of nonlinear ODEs $\dot{u} = f(u)$ splitted into

$$\begin{aligned} \dot{u_1} &= f_1(u_1, u_2) \\ \dot{u_2} &= f_2(u_1, u_2), \end{aligned} \tag{4.20}$$

such that $u_1$ denote the slow and $u_2$ the fast variables. Calculate the matrices $V(u)$ and $\Lambda(u)$ such that

$$V(u)\nabla f(u) = \Lambda(u)V(u)$$

is an eigenvalue analysis of $\nabla f$ with $\Lambda(u)$ being diagonal with the entries ordered with decreasing real part.

The QSSA–parameterization $\psi_Q$ to this problem fulfills the equations

$$f_2(u_1, \psi_Q(u_1)) = 0,$$

the ILDM–parameterization $\psi_I$ solves the equations

$$\begin{pmatrix} V_{21}(u_1, \psi_I(u_1)) & V_{22}(u_1, \psi_I(u_1)) \end{pmatrix} \cdot f(u_1, \psi_I(u_1)) = 0$$

where the both parameterizations represent the fast variables $u_2$.

Consider the nonlinear reaction

$$2sp_1 \underset{1}{\overset{2}{\rightleftharpoons}} 2sp_2 \underset{90}{\overset{1000}{\rightleftharpoons}} 2sp_3 \xrightarrow{500} sp_4 \tag{4.21}$$

and take species $sp_1$ as the process variable. Then the QSSA–parameterization is given by

$$\psi_Q(u_1) = \begin{pmatrix} \sqrt{118/50059} \\ \sqrt{200/50059} \end{pmatrix} u_1. \tag{4.22}$$

The ILDM–parameterization cannot be given analytically for this example. But the result and the difference to the QSSA–parameterization is shown in figure 4.14.

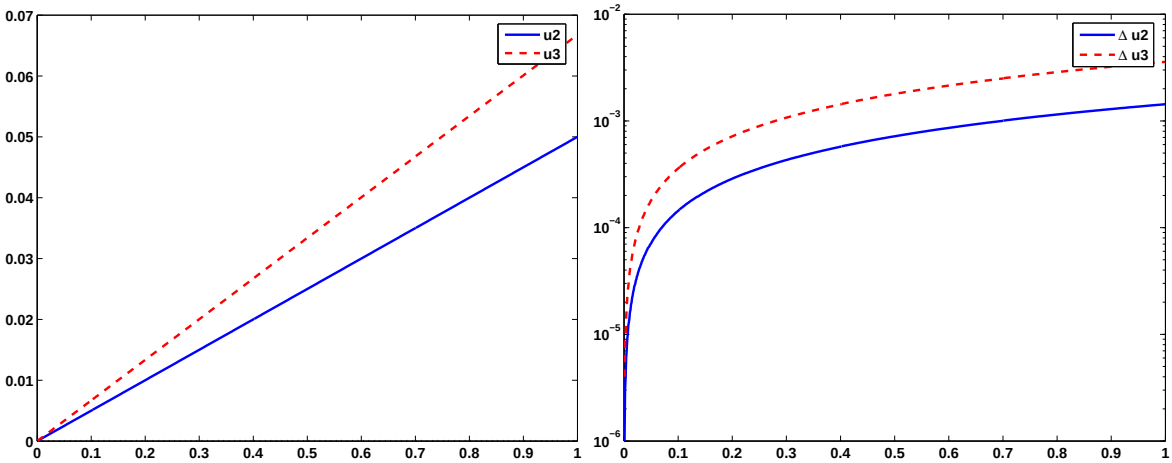


**Figure 4.14:** The left plot shows the value of the ILDM–parameterization $\psi_I$ for reaction (4.21). The right figure shows the difference between the QSSA– and the ILDM–parameterization.

### 4.4.2 Description of reduced problems

The direct application of the parameterizations reduces the problem

$$f(u) + g(u) = 0$$

to

$$\begin{aligned} f_1(u_1, \psi(u_1)) + g_1(u_1, \psi(u_1)) &= 0 \\ u_2 &= \psi(u_1). \end{aligned}$$

Clearly, these two formulations may lead to different solutions. They are equal, if the conditions of the following lemma are fulfilled.

**Lemma 4.4.1** *Let $\psi$ be the (QSSA– or ILDM–)parameterization for the function $f$ with $\psi(0) = 0$ and let $g$ be defined such that $\psi$ is the parameterization for $g$ as well. Then the reduced solution to*

$$f(u) + g(u) = 0$$

*equals the detailed solution.*

**Proof:** Let first $\psi$ denote the QSSA–parameterization. Then the reduced solution $u$ fulfills the condition $u_2 = \psi(u_1)$ and $f_2(u_1, u_2) = g_2(u_1, u_2) = 0$, because $\psi$ is the QSSA–parameterization for both $f$ and $g$. Hence also the detailed problem is solved by $u$.

Let now $\psi$ be parameterization of the ILDM–manifold. Then the reduced solution $u = (u_1, \psi(u_1))^T$ fulfills the relation $V_2(u_1, u_2)f(u_1, u_2)+V_2(u_1, u_2)g(u_1, u_2) = 0$, hence $V_{21}f_1+V_{22}f_2+V_{21}g_1+V_{22}g_2 = 0$, which leads with $f_1(u_1, u_2)+g_1(u_1, u_2) = 0$ and invertible $V_{22}(u_1, u_2)$ to $f_2(u_1, u_2)+g_2(u_1, u_2) = 0$. The reduced solution does therefore solve the detailed problem. ■

Examples, which fulfill the conditions of this lemma, are given here:

- The functions $g(u) = \alpha f(u)$ fulfill the conditions for all $\alpha \in \mathbb{R}$ for both the ILDM– and QSSA–parameterization.
- In case of QSSA, the condition $g_2(u) = \alpha f_2(u)$ is sufficient. This means especially that all functions $g$ with $g_2(u) = 0$ can be treated.
- A linear perturbation of the original problem of the form $g(u) = bu$, $b \in \mathbb{R}$ cannot be solved exactly by the reduction with ILDM as it was the case in the linear theory. The main difference to now is the fact, that in the linear case, the ILDM–parameterization was given by $\psi(u_1) = -V_{22}^{-1}V_{21}u_1$ and was therefore calculated directly of the fast left eigenvectors, which implies $V_2(u)u = 0$, if $u$ is on the manifold. This relation does not hold in the nonlinear case. Take for example the ODE

  $$\begin{aligned}\dot{u}_1 &= -u_1\\ \dot{u}_2 &= 1000(1/u_1 - u_2)\end{aligned}$$

  with the fast left eigenvector

  $$V_2 = \begin{pmatrix}\frac{1000}{999u_1^2} & 1\end{pmatrix}$$

  of the gradient and the parameterization of the ILDM

  $$\psi(u_1) = \frac{998}{999u_1}.$$

  Then clearly $V_2(u_1, \psi(u_1)) \cdot (u_1, \psi(u_1)) = 2/u_1 \neq 0$. A linear perturbation of the above form with exact reduction can therefore not exist.
- Taking a constant function $g(u) = b$ can only be successful, if either $b_2 = 0$ and the QSSA–method is considered, or if the left eigenvectors to the fast eigenvalues of the Jacobian of $f$ are independent of $u$.

### 4.4.3 Reducing ordinary and partial differential equations

Claiming the exact reduction of the implicit Euler method for ordinary differential equations means that the equation

$$u(t + k) - kf(u(t + k)) = u(t)$$

is to be reduced exactly in each time step. This means at least that there is a vector $r \in \mathbb{R}^n$ such that the equation

$$f(u) + bu = r$$

with $b \in \mathbb{R}$ is reduced exactly. But this is impossible, as the following observation shows: Let $V_2(u)$ denote the fast left eigenvectors of the gradient of $f$. Then $V_2(u)(bu + r) = 0$ for all $u$ is the necessary condition to fulfill the conditions of lemma 4.4.1, hence especially for $u = 0$. This implies that $V_2(u)r$ has to be zero, but this is only possible, if $V_2$ does not depend on $u$, which is in general not the case.

This means unfortunately that the ODE $\dot{u} = f(u)$ cannot be reduced as nicely as in the linear case. In order to see the systematic error produced by the application of the ILDM method to an ODE, consider the problem

$$\dot{u} = f(u)$$
$$u(0) = \begin{pmatrix} 0.9 \\ \psi(0.9) \end{pmatrix}$$

with $f$ denoting the source term of reaction (6.5) in the following section. The difference between the detailed and reduced solution can be seen in figure 4.15. The figures show

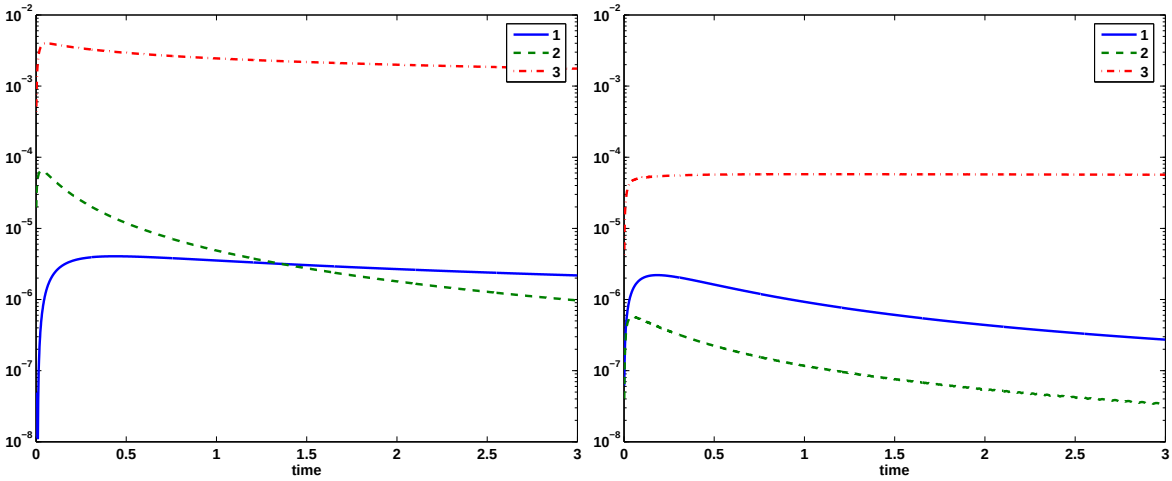


**Figure 4.15:** The difference between the detailed and reduced solution for the ode $\dot{u} = f(u)$ with source term (6.5). In the left figure, the reduction depends on QSSA, for the right figure, the ILDM method is applied.

clearly that systematic errors occur. They also show that the ILDM method performs much better than the application of the QSSA.

The same observation can be done by equations of the form $-u'' = f(u)$, where even linear differential operators contradict the exact reduction with ILDM, unless $f$ is linear. A detailed analysis of the produced errors will follow in the upcoming sections.

### 4.4.4 Reduction of general equations

As an example for a general disturbance $g$, which does not fulfill the conditions of the above lemma, consider the equation

$$f(u) + \alpha B u = \beta b \tag{4.23}$$

with $f$ denoting the source term of reaction (4.21) and

$$B = \begin{pmatrix} -2 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{pmatrix}, \qquad b = \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix}.$$

The error depending on the constants $\alpha$ and $\beta$ is shown in figure 4.16. It shows that a bigger

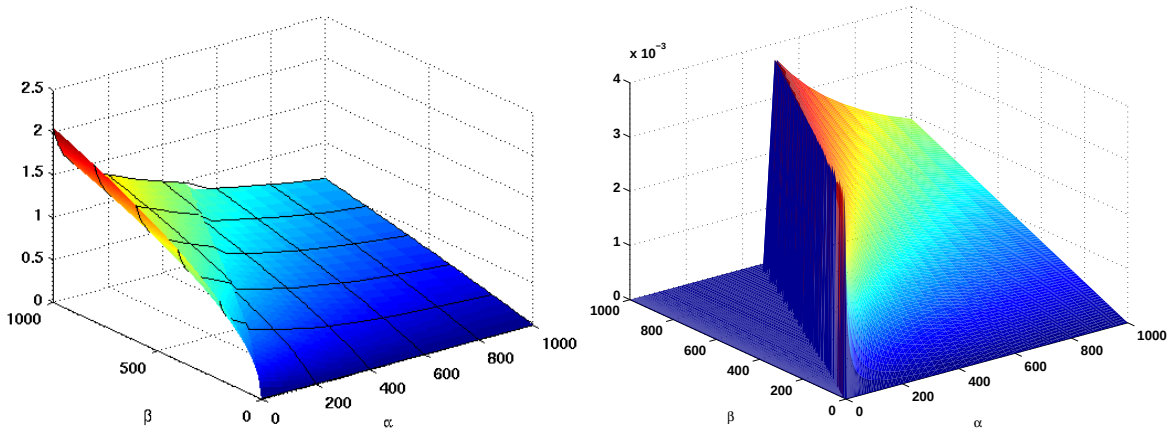


**Figure 4.16:** The left figure shows the error produced by the direct application of the ILDM–parameterization $\psi_I$ to equation (4.23). The right figure shows the difference of the two reduced solutions, if the QSSA– and the ILDM–parameterization is applied.

influence of the operator $B$ leads to a better performance, whereas the error increases with increasing right hand side $b$, similar as in the linear case. The figure does also suggest that the difference between the reduction methods QSSA and ILDM is rather small.

### 4.4.5 A PDE example

Consider the partial differential equation

$$-\nu u''(x) = f(u(x)), \quad x \in \Omega := ]0,1[ \tag{4.24}$$

with Dirichlet boundary conditions, where the function $f$ describes the source term of the nonlinear reaction (4.21). The solution and the error produced by the reduction process can be seen in figure 4.17.

Similar as in the linear case (cp. table 4.1), the performance of the reduction is almost independent of the grid size. In the above example with $\nu = 1$, the difference between the

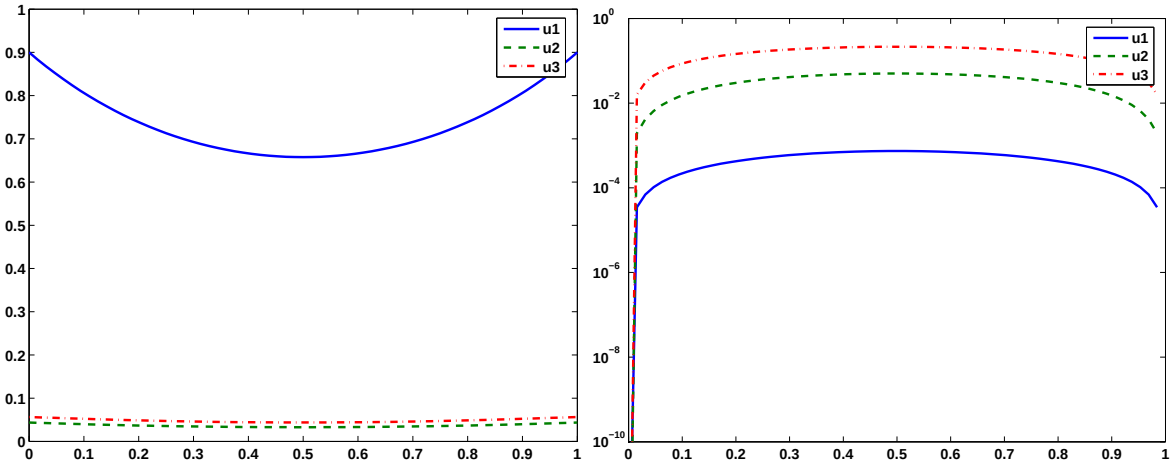


**Figure 4.17:** The solution of the PDE (4.24) with $\nu = 1$ is shown in the left figure. The right figure shows the systematic error produced by the reduction with QSSA.

detailed and the reduced solution is $\|u_{det} - u_{red}\|_\infty \approx 2 \cdot 10^{-3}$ for QSSA and $3 \cdot 10^{-3}$ for ILDM. Note that unlike in the linear case, the application of the ILDM–table leads to systematic errors for equation (4.24).

Clearly, the performance depends strongly on the spectral gap in the reaction mechanism and the viscosities. This dependence is shown in table 4.7 for the reaction mechanisms

$$2sp_1 \underset{1}{\overset{2}{\rightleftharpoons}} sp_2 \underset{k \cdot 90}{\overset{k \cdot 1000}{\rightleftharpoons}} sp_3 \xrightarrow{k \cdot 500} sp_4$$

with different values for $k$. The viscosity for the process variable $u_1$ is taken to be $\nu_1 = 10^{-2}$ in all cases. The table depicts clearly that the performance is best, if the viscosities for the fast variables equals the viscosity of the process variables. It also shows that an increasing $k$ and therefore an increasing spectral gap leads to better approximations.

## 4.5 Nonlinear equations including conservation laws

Now reaction mechanisms are considered, which produce a center manifold. The difference to the previous section is therefore mainly the conservation constant and the replacement of part of the equations by the equations, which force the solution to be on the center manifold. This effect leads to some differences compared to the results in the previous section, which can probably best be seen by comparing figures 4.16 and 4.20.

The existence of the conservation laws does unfortunately not improve the results on exactly reduced equations from the previous section. ODEs and PDEs can only be solved by applying reduction methods, if systematic errors are accepted.

| $k$ | $\nu_2 = \nu_3 = 1$ | $\nu_2 = \nu_3 = 10^{-2}$ | $\nu_2 = \nu_3 = 10^{-4}$ |
|---|---|---|---|
| 1 | $\Delta u_1 = 5e-4$<br>$\Delta u_2 = 2e-2$<br>$res_2 = 6e-1$ | $\Delta u_1 = 2e-9$<br>$\Delta u_2 = 4e-8$<br>$res_2 = 3e-7$ | $\Delta u_1 = 9e-6$<br>$\Delta u_2 = 2e-3$<br>$res_2 = 6e-3$ |
| 10 | $\Delta u_1 = 2e-5$<br>$\Delta u_2 = 5e-3$<br>$res_2 = 2e-1$ | $\Delta u_1 = 1e-10$<br>$\Delta u_2 = 4e-9$<br>$res_2 = 7e-8$ | $\Delta u_1 = 3e-7$<br>$\Delta u_2 = 2e-4$<br>$res_2 = 2e-3$ |
| 100 | $\Delta u_1 = 7e-7$<br>$\Delta u_2 = 8e-4$<br>$res_2 = 6e-2$ | $\Delta u_1 = 2e-11$<br>$\Delta u_2 = 5e-9$<br>$res_2 = 2e-7$ | $\Delta u_1 = 9e-9$<br>$\Delta u_2 = 2e-5$<br>$res_2 = 6e-4$ |

**Table 4.7:** The table shows the performance of the reduction provided by the ILDM method in dependence of the spectral gap in the gradient of the source term and the viscosities for the fast variables. The first two numbers in each block $\Delta u_1$ and $\Delta u_2$ denote the difference of the slow and fast variables between the detailed and reduced solution $\|u_{det} - u_{red}\|_\infty$. The number $res_2$ denotes the detailed residual of the fast variables of the reduced solution. The viscosity for the process variables is always taken to be $\nu_1 = 10^{-2}$.

### 4.5.1 The parameterization of the manifolds

Consider a system of nonlinear ODEs $\dot{u} = f(u)$ splitted into

$$\begin{aligned} \dot{u_1} &= f_1(u_1, u_2, u_3) \\ \dot{u_2} &= f_2(u_1, u_2, u_3) \\ \dot{u_3} &= f_3(u_1, u_2, u_3) \end{aligned} \tag{4.25}$$

with the matrices $V(u)$ and $\Lambda(u)$ such that

$$V(u)\nabla f(u) = \Lambda(u)V(u).$$

The ODEs are supposed to include the conservation laws

$$V_1^0 f_1(u) + V_2^0 f_2(u) + V_3^0 f_3(u) = 0$$

for all possible $u$, where $V_i^0$ are independent of the state vector $u$. This means that the term $\begin{pmatrix} V_1^0 & V_2^0 & V_3^0 \end{pmatrix} u = c$ is constant in time and therefore defined by the initial conditions

$$c := \begin{pmatrix} V_1^0 & V_2^0 & V_3^0 \end{pmatrix} u(0).$$

The rows of the matrix $\begin{pmatrix} V_1^0 & V_2^0 & V_3^0 \end{pmatrix}$ do clearly span a subspace of the left eigenspace to the eigenvalue $\lambda = 0$ of $Df(u)$.

Let the system of ODEs be ordered such that $u_1$ describes the process variables, $u_2$ the fast variables and $u_3$ the variables, which might be replaced by the conservation laws and leads to the DAE

$$\begin{aligned} \dot{u_1} &= f_1(u_1, u_2, u_3) \\ \dot{u_2} &= f_2(u_1, u_2, u_3) \\ c &= \begin{pmatrix} V_1^0 & V_2^0 & V_3^0 \end{pmatrix} \cdot u, \end{aligned}$$

which is equivalent to the original ODE.

The QSSA–parameterization $\psi_Q$ to this problem fulfills the equations

$$\begin{aligned} f_2(u_1, \psi_Q(u_1)) &= 0 \\ \begin{pmatrix} V_1^0 & V_2^0 & V_3^0 \end{pmatrix} \begin{pmatrix} u_1 \\ \psi_Q(u_1) \end{pmatrix} &= c. \end{aligned}$$

The ILDM–parameterization $\psi_I$ solves the equations

$$\begin{aligned} \begin{pmatrix} V_{21}(u_1, \psi_I(u_1)) & V_{22}(u_1, \psi_I(u_1)) & V_{23}(u_1, \psi_I(u_1)) \end{pmatrix} \cdot f(u_1, \psi_I(u_1)) &= 0 \\ \begin{pmatrix} V_1^0 & V_2^0 & V_3^0 \end{pmatrix} \begin{pmatrix} u_1 \\ \psi_I(u_1) \end{pmatrix} &= c, \end{aligned}$$

where the both parameterizations represent the variables $u_2$ and $u_3$.

Consider the nonlinear reaction given by (3.6) and take species $sp_1$ as the process variable. Then the QSSA–parameterization is given by

$$\psi_Q(u_1) = {}^1\!/_{911} \begin{pmatrix} -90 + 90u_1 + \sqrt{90090 - 180180u_1 + 91912u_1^2} \\ 1001 - 1001u_1 - \sqrt{90090 - 180180u_1 + 91912u_1^2} \end{pmatrix}. \tag{4.26}$$

Note that the value for $u_3$ is negative for $u_1 = 1$. In order not to obtain negative mass fractions, the negative values might be set to zero, if $u_2$ is also changed such that the conservation laws are not violated. The ILDM–parameterization cannot be given analytically for this example. But the result and the difference to the QSSA–parameterization is shown in figure 4.18.

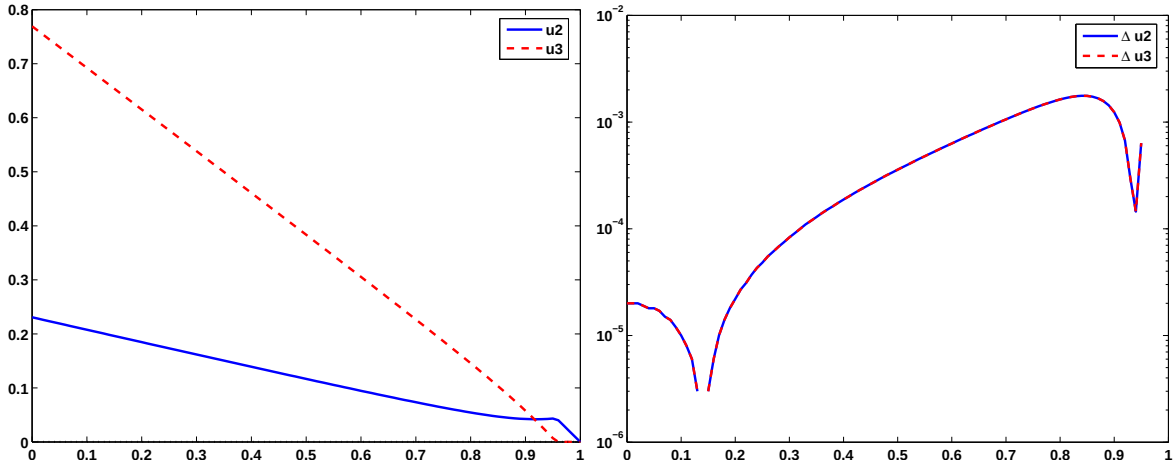


**Figure 4.18:** The left plot shows the value of the ILDM–parameterization $\psi$ for reaction (3.6) including conservation laws. Compare with figure 3.3 for a parameterization, where conservation laws are not considered. The right figure shows the difference between the QSSA– and the ILDM–parameterization.

### 4.5.2 Description of reduced problems

The direct application of the parameterizations reduces the problem

$$f(u) + g(u) = 0 \tag{4.27}$$

to

$$\begin{aligned} f_1(u_1, \psi(u_1)) + g_1(u_1, \psi(u_1)) &= 0 \\ u_2 &= \psi(u_1). \end{aligned}$$

These two formulations are equivalent, if the conditions of the following lemma are fulfilled.

**Lemma 4.5.1** *Let $\psi$ denote the parameterization of the manifold for the function $f$. If this function $\psi$ is also the parameterization for the function $g$, then the reduced solution to*

$$f(u) + g(u) = 0$$

*equals the detailed solution.*

**Proof:** Let first $\psi$ denote the QSSA–parameterization. Then the reduced solution $u = (u_1, u_2)$ of $f(u) + g(u) = 0$ is on the QSSA–manifold and fulfills therefore the relations $f_2(u) = g_2(u) = 0$ and $V^0(u)f(u) = V^0(u)g(u) = 0$, because $g$ has the same conservation laws as $f$. This leads together with $f_1(u) + g_1(u) = 0$ to $f_3(u) + g_3(u) = 0$, which completes the proof for QSSA.

In case of ILDM, we have $V_2(u)f(u) = V_2(u)g(u) = 0$ for the reduced solution $u$, which leads to $V_{22}(u)(f_2(u) + g_2(u)) + V_{23}(u)(f_3(u) + g_3(u)) = 0$. The conservation laws lead to the relation $V_2^0(u)(f_2(u) + g_2(u)) + V_3^0(u)(f_3(u) + g_3(u)) = 0$. These two equations result in $f_2(u) + g_2(u) = 0$ and $f_3(u) + g_3(u) = 0$. ■

One example of $g$, which fulfills the conditions of the above lemma, is the function $g = \alpha f$ with $\alpha \in \mathbb{R}$. Due to the treatment of the center manifold in the parameterization $\psi$, this example cannot even be broadened for the QSSA–method as in the case, where no conservation laws hold.

### 4.5.3 Reduction of ordinary and partial differential equations

Since conservation laws cannot improve the performance of the reduction methods with respect to exactness and ODEs and PDEs cannot be reduced exactly for nonlinear equations without conservation laws, these type of equations cannot be reduced without systematic errors in the current case either.

In order to see the effect of the reduction methods to ODEs, consider the problem

$$\begin{aligned} \dot{u} &= f(u) \\ u(0) &= \begin{pmatrix} 0.9 \\ \psi(0.9) \end{pmatrix} \end{aligned}$$

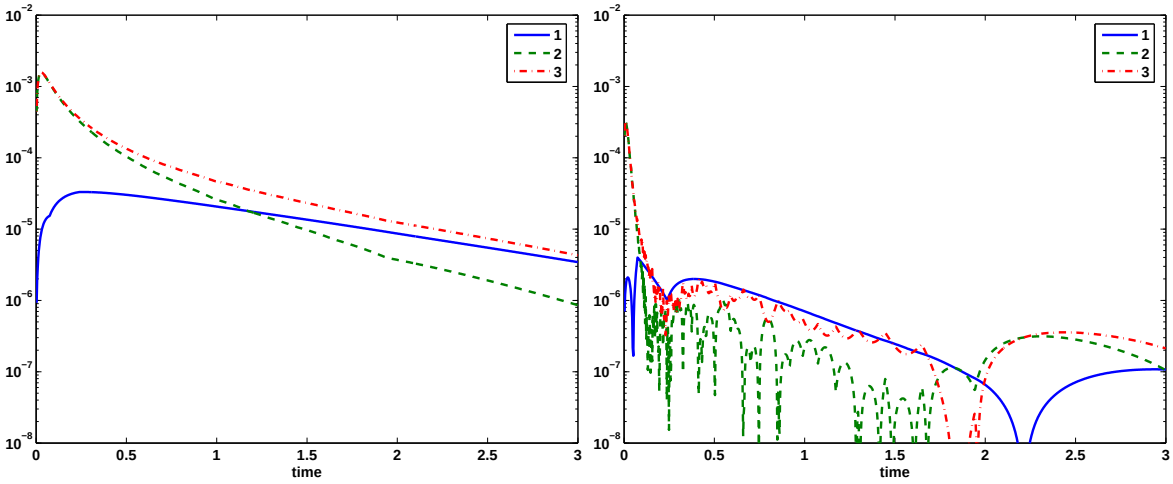


**Figure 4.19:** The difference between the detailed and reduced solution for the ode $\dot{u} = f(u)$ with source term (3.6). In the left figure, the reduction depends on QSSA, for the right figure, the ILDM–method is used.

with $f$ describing the source term (3.6). The differences between the detailed and reduced solutions can be seen in figure 4.19. Clearly, systematic errors occur with the application of the reduction method. But still, the ILDM–method performs much better than the reduction with the QSSA–manifold.

### 4.5.4 Reduction of general equations

For general disturbances $g$, the reduced formulation introduces clearly systematic errors. As an example, consider the equation

$$f(u) + \alpha B u = \beta b \tag{4.28}$$

with $f$ denoting the source term of reaction (3.6) and

$$B = \begin{pmatrix} -2 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{pmatrix}, \qquad b = \begin{pmatrix} -1 \\ -1 \\ 2 \end{pmatrix}.$$

The right hand side $b$ is chosen such that the conservation laws are fulfilled, if $\alpha = 0$. The error depending on the constants $\alpha$ and $\beta$ is shown in figure 4.20. It shows that a bigger influence of the operator $B$ leads to a better performance, whereas the error increases with increasing right hand side $b$, similar as in the linear case. The figure does also suggest that the difference between the reduction methods QSSA and ILDM is rather small.

### 4.5.5 A PDE example

As an example, consider the reaction–diffusion equation

$$-\nu u''(x) = f(u(x)), \quad x \in \Omega := ]0, 1[ \tag{4.29}$$

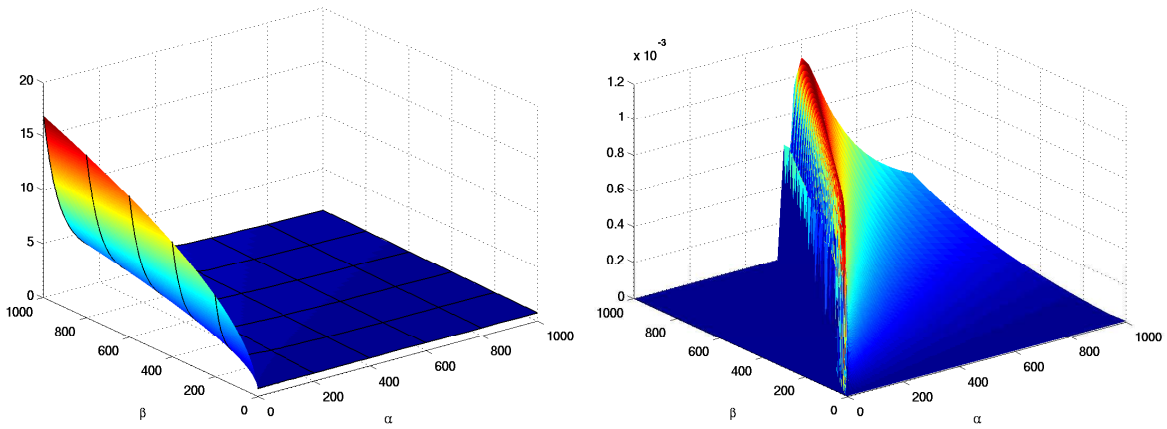


**Figure 4.20:** The left figure shows the error produced by the direct application of the ILDM–parameterization $\psi_I$ to equation (4.28). The right figure shows the difference of the two reduced solutions, if the QSSA– and the ILDM–parameterization is applied.

with Dirichlet boundary conditions. Here, the diffusion is modeled by the Laplacian, but more advanced models are also possible, think for example of Fick's law [17]. The function $f$ describes the source term of the nonlinear reaction (3.6) by

$$f(u) = \begin{pmatrix} -4u_1^2 + 2u_2^2 \\ 4u_1^2 - 2002u_2^2 + 180u_3^2 \\ 2000u_2^2 - 180u_3^2 \end{pmatrix}.$$

The domain $\Omega$ shall be discretized by

$$\Omega_h := \{0, h, ..., (N-1)h, Nh\}$$

with $Nh = 1$. This leads to the algebraic equations

$$f_h(u_h) + B_h u_h = b_h$$

with the state vector

$$u_h = \begin{pmatrix} u_h^1 \\ \vdots \\ u_h^N \end{pmatrix},$$

where $u_h^i$ describes the mass fractions of the three species at the corresponding grid point. If the trapezoid rule is applied as the integration strategy, this leads to

$$f_h(u_h) = \begin{pmatrix} u_h^1 \\ -1/h f(u_h^2) \\ \vdots \\ -1/h f(u_h^{N-1}) \\ u_h^N \end{pmatrix} \quad \text{and} \quad B_h = \begin{pmatrix} 0 & 0 & \cdots & \cdots & 0 \\ -1/h\mathbb{1} & 2/h\mathbb{1} & -1/h\mathbb{1} & & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & & -1/h\mathbb{1} & 2/h\mathbb{1} & -1/h\mathbb{1} \\ 0 & \cdots & \cdots & 0 & 0 \end{pmatrix}.$$

The right hand side $b_h$ is basically zero, only the first and last three entries are given by the boundary conditions for the differential equations. Here, the boundaries $(1, 0, 0)^T$ are taken for $x = 0$ and $x = 1$, because $\psi(1) = (0, 0)^T$.

The solution to this problem is visualized in figure 4.21.

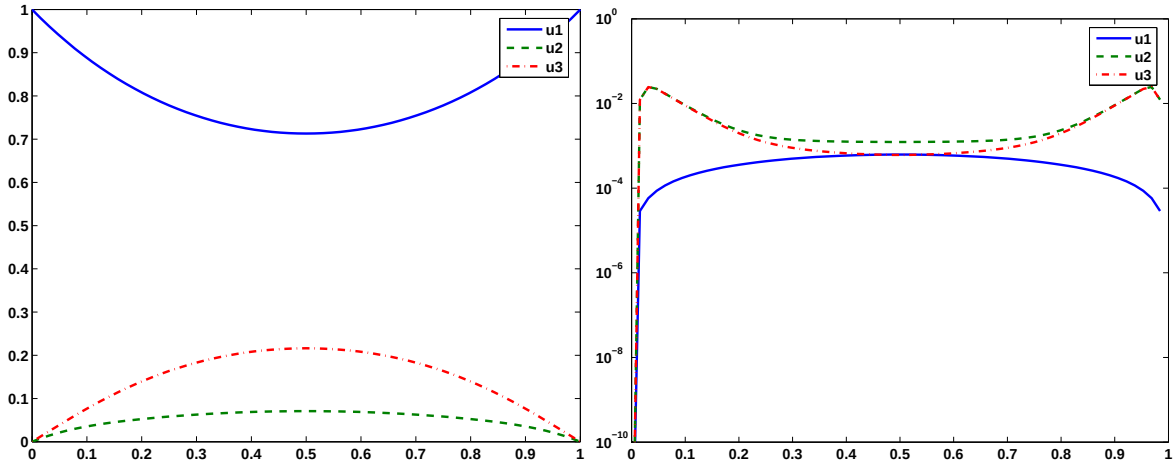


**Figure 4.21:** The solution to the problem $-u''(x) = f(u(x))$ with the nonlinear reaction (3.6) and Dirichlet boundary conditions according to the table $\psi$. The right figure shows the error produced by the reduction mechanism on a grid with 65 nodes.

# 5 Preconditioning with QSSA and ILDM

Due to the systematic errors, which are produced by the application of the QSSA– and ILDM–tables to general linear problems, an algorithm is required, which takes advantage of the relative low costs of the application of the reduction methods, but which allows to reduce the systematic error by spending little more computational effort. For that reason, preconditioning strategies with QSSA and ILDM are investigated in this chapter.

The first two sections will deal with various possibilities to precondition linear iterative solvers with and without center manifolds. In both cases, it will be shown that the reduction methods cannot be used successfully as preconditioners, because the iterations will not converge. In case of the pure application of the tables, even analytical proofs are presented.

The conjecture is that the performance of the preconditioning can be increased by an operator splitting. The chemical part shall be considered separately from the operators resulting from physical phenomena and discretized differential operators. Therefore a third section is presented, which treats a preconditioning technique with time steps. Within the time stepping, the Strang splitting technique will applied to separate the chemical source term, which can on its own perfectly be preconditioned by the reduction methods. But a detailed analysis of the stationary points lead to the conclusion that this reduced Strang splitting technique does not lead to convergence either.

The obtained results are clearly also valid for problems, where the ILDM–method leads to exact results. This means that situations can be created, where the reduction of the problem with ILDM does not lead to systematic errors, but the application of ILDM as preconditioner prevents the algorithm from convergence.

## 5.1 The solution process

In order to explain the application of the reduction methods QSSA and ILDM as preconditioners, consider the system of linear equations

$$(A + B)u = b$$

with $A$ denoting the source term of a linear reaction and $B$ an arbitrary disturbance, such that $A + B$ is invertible.

Then the most simple algorithm to obtain the solution iteratively is the Richardson iteration

$$u^{n+1} = u^n + b - (A + B)u^n,$$

which converges only for special choices of $A$ and $B$. An improvement can be achieved, if an approximation $P$ of the matrix $(A+B)^{-1}$ is known. Then the iteration

$$u^{n+1} = u^n + P(b - (A+B)u^n)$$

converges for all initial choices $u^0$, if and only if the spectral radius of the system matrix

$$S := \mathbb{1} - P(A+B)$$

is less than one, see e.g. [13, Ch. 8.1].

The most commonly known preconditioners for full matrices are probably $P = D^{-1}$ and $P = (D+L)^{-1}$ ($D$ describes the diagonal and $L$ the lower triangular part of $A+B$), which lead to the Jacobi and the Gauss–Seidel iteration. For sparse matrices, the incomplete LU–factorization or the incomplete Cholesky factorization lead to reasonable convergence rates. These preconditioners are therefore used for the smoother in the multi–level solver in Gascoigne [3].

In order to use the reduction methods QSSA and ILDM as preconditioners, reformulate the iteration to

$$u^{n+1} = u^n + h,$$

where $h$ is the reduced solution of the linear system

$$(A+B)h = b - (A+b)u^n,$$

which is obtained by the application of the QSSA or ILDM method.

## 5.2 Linear equations without conservation laws

In this section three possibilities to apply the reduction methods and their variations as preconditioners to linear equations is investigated. In the first and main part, the eigenvalues of the system matrix $\mathbb{1} - P(A+B)$ will be proven to be zero and one. It will be shown that the iteration can therefore only converge in the first iteration step and conditions for the initial guess will be given, such that convergence occurs at all. The first part will investigate the variations of the reduction methods, where $A_{22}$ or $A_{22}+B_{22}$ is allowed to be inverted. Finally, a more realistic example will be considered, namely a linear reaction–diffusion equation in one dimension of space.

The main results of this section are:

- Applying the reduction methods as preconditioner can only lead to convergence, if the initial value is chosen appropriately. Then convergence occurs in the first iteration step, but the method is unstable with respect to the initial guess.
- Applying $A_{22}^{-1}$ on top of the reduction methods leads to divergence for realistic examples.
- Applying $(A_{22} + B_{22})^{-1}$ in addition to QSSA or ILDM leads to nice results, but the strategy is too expensive for practical purposes.

### 5.2.1 Preconditioning with QSSA and ILDM

Recall from section 4.2.2 that the approximated solution to the problem $(A + B)u = b$ is given by

$$u = \begin{pmatrix} A_{11} + B_{11} & A_{12} + B_{12} \\ Y_{21} & Y_{22} \end{pmatrix}^{-1} \begin{pmatrix} b_1 \\ 0 \end{pmatrix}$$

with $Y_{2i} = A_{2i}$ for the QSSA method and $Y_{2i} = \Lambda_2 V_{22}$ for the application of the ILDM method.

The preconditioner $P$ can therefore be given explicitely by

$$P = \begin{pmatrix} A_{11} + B_{11} & A_{12} + B_{12} \\ Y_{21} & Y_{22} \end{pmatrix}^{-1} \begin{pmatrix} \mathbb{1} & 0 \\ 0 & 0 \end{pmatrix}. \tag{5.1}$$

Its performance can be characterized by the spectral radius of the system matrix, which is calculated in the following lemma.

**Lemma 5.2.1** *Let $A$ and $B$ be block matrices as above and let $P$ be given by* (5.1)*. Then the rows of the matrices*

$$V_1 = \begin{pmatrix} Y_{21} & Y_{22} \end{pmatrix} \quad \text{and} \quad V_2 = \begin{pmatrix} A_{11} + B_{11} & A_{12} + B_{12} \end{pmatrix}$$

*denote the left eigenvectors of the matrix $\mathbb{1} - P(A+B)$ to the eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 0$.*

**Proof:** Obviously

$$\begin{pmatrix} A_{11} + B_{11} & A_{12} + B_{12} \\ Y_{21} & Y_{22} \end{pmatrix} P = \begin{pmatrix} \mathbb{1} & 0 \\ 0 & 0 \end{pmatrix}$$

and

$$\begin{pmatrix} A_{11} + B_{11} & A_{12} + B_{12} \\ Y_{21} & Y_{22} \end{pmatrix} P(A + B) = \begin{pmatrix} A_{11} + B_{11} & A_{12} + B_{12} \\ 0 & 0 \end{pmatrix}.$$

The proof is completed by treating the second equation row-wise. ∎

This lemma shows that the Richardson iteration remains constant after the first step: Let

$u^n$ denote the $n$–th iterate for an arbitrary initial value $u^0$. Then $u^n$ fulfills the equation

$$\begin{aligned}
&\begin{pmatrix} A_{11}+B_{11} & A_{12}+B_{12} \\ Y_{21} & Y_{22} \end{pmatrix} u^n \\
&= \begin{pmatrix} A_{11}+B_{11} & A_{12}+B_{12} \\ Y_{21} & Y_{22} \end{pmatrix} ((\mathbb{1} - P(A+B))u^{n-1} + Pb) \\
&= \begin{pmatrix} A_{11}+B_{11} & A_{12}+B_{12} \\ Y_{21} & Y_{22} \end{pmatrix} (\mathbb{1} - P(A+B))^n u^0 \\
&+ \underbrace{\sum_{k=1}^{n-1} \begin{pmatrix} A_{11}+B_{11} & A_{12}+B_{12} \\ Y_{21} & Y_{22} \end{pmatrix} (\mathbb{1} - P(A+B))^k Pb}_{=0 \text{ by lemma } 5.2.1} + \begin{pmatrix} A_{11}+B_{11} & A_{12}+B_{12} \\ Y_{21} & Y_{22} \end{pmatrix} Pb \\
&= \begin{pmatrix} A_{11}+B_{11} & A_{12}+B_{12} \\ Y_{21} & Y_{22} \end{pmatrix} (\mathbb{1} - P(A+B))u^0 + \begin{pmatrix} A_{11}+B_{11} & A_{12}+B_{12} \\ Y_{21} & Y_{22} \end{pmatrix} Pb \\
&= \begin{pmatrix} A_{11}+B_{11} & A_{12}+B_{12} \\ Y_{21} & Y_{22} \end{pmatrix} u^1.
\end{aligned}$$

Due to the invertibility of $\begin{pmatrix} A_{11}+B_{11} & A_{12}+B_{12} \\ Y_{21} & Y_{22} \end{pmatrix}$, the iterate $u^n$ equals the first iterate $u^1$ for all $n \geq 1$.

This means that the performance of the Richardson iteration depends highly on the first iteration step and therefore on the initial value $u^0$, which is to be chosen, such that the Richardson iteration converges in the first step. This is technically possible. If the fast part of $u^0$ is chosen to be

$$\begin{aligned}
u_2^0 = (A_{22}+B_{22} - (B_{21}P_{11} + B_{22}P_{21})(A_{12}+B_{12}))^{-1} \\
\big[\beta b_2 - (A_{21}+B_{21} - (B_{21}P_{11}+B_{22}P_{21})(A_{11}+B_{11}))u_1^0 \\
- \beta \begin{pmatrix} B_{21} & B_{22} \end{pmatrix} Pb\big], \quad (5.2)
\end{aligned}$$

then, the existence of $u_2^0$ provided, the Richardson iteration converges, if QSSA is used for the preconditioning. The proof is obtained by the verification of the equation

$$(A + \alpha B)u^1 = \beta b,$$

which is purely technical and therefore omitted in this thesis.

A similar formulation for the fast part of $u^0$ can be given, if the preconditioner $P$ is created by the ILDM–parameterization. This formulation is of even higher technical difficulties, because the difference of $V_{2i}$ and $A_{2i} + B_{2i}$ is also required.

This observation shows that the parameterization can hardly be applied as a preconditioner. First of all, the calculation of $u^0$ involves the time consuming inversion of a matrix of type $A_{22}+B_{22}$. And secondly, errors in the computation of the initial value will be inherited to the following iterates, such that convergence occurs only theoretically. In practical applications, the iterates will be bounded away from the exact solution.

The fact, that the eigenvalues of the system matrix of the Richardson iteration are $\lambda_1 = 0$ and $\lambda_2 = 1$ is tempting to apply the parameterizations together with additional damping as a preconditioner in order to circumvent the expensive calculation of the fast part of the initial guess. But a closer look to the proof of lemma 5.2.1 shows, that the preconditioning matrix $P_\omega$ as a result of the damped preconditioner $P$ leads to the eigenvalues $\lambda_1 = 1-\omega$ and $\lambda_2 = 1$. This means that the damping leads to an even worse performance of the preconditioner.

Note that lemma 5.2.1 is also valid for matrices $A$ and $B$ with $AB = BA$. This means that the ILDM method leads under certain circumstances to an exact solver, but the same method applied as preconditioner is rather useless.

### 5.2.2 Preconditioning with modified parameterizations

#### Allow the inversion of $A_{22}$

If the problem $(A+B)u = b$ is solved with the modified reduction methods as in 4.2.6, where the additional inversion of the block $A_{22}$ was allowed in each evaluation of $\psi$, the obtained solution reads

$$u = \begin{pmatrix} A_{11} + B_{11} & A_{12} + B_{12} \\ A_{21} & A_{22} \end{pmatrix}^{-1} b$$

or

$$u = \begin{pmatrix} A_{11} + B_{11} & A_{12} + B_{12} \\ \Lambda_2 V_{21} & \Lambda_2 V_{22} \end{pmatrix}^{-1} \begin{pmatrix} b_1 \\ V_{21} b_1 + V_{22} b_2 \end{pmatrix}$$

depending on the reduction method. The application of the QSSA method leads therefore to the preconditioner

$$P = \begin{pmatrix} A_{11} + B_{11} & A_{12} + B_{12} \\ A_{21} & A_{22} \end{pmatrix}^{-1}, \tag{5.3}$$

the ILDM preconditioner reads

$$P = \begin{pmatrix} A_{11} + B_{11} & A_{12} + B_{12} \\ \Lambda_2 V_{21} & \Lambda_2 V_{22} \end{pmatrix}^{-1} \begin{pmatrix} \mathbb{1} & 0 \\ V_{21} & V_{22} \end{pmatrix}.$$

Consider the Richardson iteration for the problem $(A + \alpha B)u = b$ with $A$ and $B$ as in (4.8), then the spectral radius of $S$ grows with $\alpha$ for both QSSA and ILDM:

$$\alpha = 1 \implies \sigma(S) = 0.002$$
$$\alpha = 1000 \implies \sigma(S) = 2.0$$

For both reduction methods, $\lambda_2$ is obviously greater than one for big values of $\alpha$. This means that convergence is only possible for small influences of the perturbation $B$.

**Allow the inversion of $A_{22} + B_{22}$**

The possibility to invert the matrix block $A_{22} + B_{22}$ in addition to the evaluation of the tables leads to the solution

$$u = \underbrace{\begin{pmatrix} A_{11} + B_{11} - (A_{12} + B_{12})Y_{22}^{-1}Y_{21} & 0 \\ A_{21} + B_{21} & A_{22} + B_{22} \end{pmatrix}^{-1}}_{=:P} b,$$

of the problem $(A + B)u = b$. Let $Y_{2i} = A_{2i}$ for the application of QSSA and $Y_{2i} = \Lambda_2 V_{2i}$, if the ILDM method is applied.

Even though the above described solution process performs only slightly better than the pure application of the table $\psi$ (compare figures 4.6 and 4.3), the application of this technique as preconditioner is much more successful. As before, consider the system matrix of the Richardson iteration $S = \mathbb{1} - P(A + \alpha B)$ with $P$ as above. Then the spectral radius of $S$ remains small even for large $\alpha$:

$$\begin{aligned} \alpha = 1 &\implies \sigma(S) = 0.0059 \\ \alpha = 1000 &\implies \sigma(S) = 0.2311 \\ \alpha = 1\,000\,000 &\implies \sigma(S) = 0.3324 \end{aligned}$$

The spectral radius seems to be bounded, the iteration is therefore a contraction.

### 5.2.3 A PDE example

Reconsider the reaction–diffusion equation given by (4.5). The above introduced preconditioners shall now be applied to the system of linear equations

$$(A_h + B_h)u = b_h$$

resulting from the discretization of this PDE. The spectral radius of the system matrices will be calculated in order to decide, whether $B_h$ is of a form, which allows the application of the pure or modified reduction methods.

The application of the QSSA method leads to the preconditioner

$$P = \begin{pmatrix} A_{h,11} + B_{h,11} & A_{h,12} + B_{h,12} \\ A_{h,21} & A_{h,22} \end{pmatrix}^{-1} \begin{pmatrix} \mathbb{1} & 0 \\ 0 & 0 \end{pmatrix}.$$

The system matrix $S$ has the eigenvalues shown in table 5.1. Independently of the grid size, the spectral radius of $S$ is $\sigma(S) = 1$. Taking into account that $\lambda = 0$ and $\lambda = 1$ are the only eigenvalues, leads to the stagnation of the iteration after the first step. This means that the convergence of this iteration depends only on the initial guess for the linear iteration. How the initial guess is to be chosen in order to have convergence, can be seen in equation (5.2). Note that this method is instable with respect to errors in the calculation of the initial guess.

| # points | $h$ | QSSA | ILDM |
|---|---|---|---|
| 5 | 0.25 | 0; 1 | 0; 1 |
| 9 | 0.125 | 0; 1 | 0; 1 |
| 17 | 0.0625 | 0; 1 | 0; 1 |
| 33 | 0.03125 | 0; 1 | 0; 1 |
| 65 | 0.015625 | 0; 1 | 0; 1 |
| 129 | 0.0078125 | 0; 1 | 0; 1 |

**Table 5.1:** The table shows the eigenvalues of the system matrix $S = \mathbb{1} - P(A_h + B_h)$.

#### Allow the inversion of $A_{22}$

If in addition to the evaluation of the tables $\psi$ the inversion of the block $A_{h,22}$ is allowed, the spectral radius of the system matrix $S = \mathbb{1} - P(A_h + B_h)$ behaves as shown in table 5.2 for both QSSA and ILDM. Since the eigenvalues of $S$ are clearly greater than one even on the

| # points | $h$ | QSSA | ILDM |
|---|---|---|---|
| 5 | 0.25 | 0.99 | 0.99 |
| 9 | 0.125 | 0.99 | 0.99 |
| 17 | 0.0625 | 2.35 | 2.35 |
| 33 | 0.03125 | 9.46 | 9.46 |
| 65 | 0.015625 | 37.91 | 37.91 |
| 129 | 0.0078125 | 151.7 | 151.7 |

**Table 5.2:** The table shows the spectral radius of the system matrix $S = \mathbb{1} - P(A_h + B_h)$, if the renewed tables $\phi_Q$ and $\phi_I$ are used as preconditioners. The values for the spectral radii with QSSA and ILDM have a difference of order $10^{-6}$.

small grids, the modified tables cannot be applied as preconditioners. The application leads due to the high spectral radius to divergence on all levels. This is especially remarkable, because the spectral radius is already too big on the coarsest grid, so combined preconditioners consisting of multilevel techniques and tables, cannot work either.

#### Allow the inversion of $A_{22} + B_{22}$

Let now the parameterization $\psi$ in addition with the inversion of $A_{22} + B_{22}$ be applied as a preconditioner. The updates for the defect correction iteration are then calculated by

$$\begin{aligned}(A_{h,11} + B_{h,11} - (A_{h,12} + \alpha B_{h,12})\psi)\delta u_1 &= res_1 \\ (A_{h,22} + B_{h,22})\delta u_2 &= res_2 - (A_{h,21} + B_{h,21})\delta u_1.\end{aligned}$$

The performance of this preconditioner is again described with the spectral radius of the matrix $S = \mathbb{1} - P(A_h + B_h)$. It can be seen in table 5.3. Since the spectral radius is clearly less than one, this method can be applied as preconditioner for the considered reaction–diffusion equation. But according to table 4.4, this possibility is almost as expensive as the detailed solution without reduction mechanisms and will therefore not be considered any further.

| # points | $h$ | QSSA | ILDM |
|---|---|---|---|
| 5 | 0.25 | $3.1e-6$ | $3.4e-6$ |
| 17 | 0.0625 | $3.1e-6$ | $3.4e-6$ |
| 65 | 0.015625 | $3.1e-6$ | $3.4e-6$ |

**Table 5.3:** The spectral radius of the system matrix $S$, if in addition to the application of the parameterization $\psi$ the matrix $A_{h,22} + B_{h,22}$ is allowed to be inverted.

## 5.3 Linear equations including conservation laws

The main difference between the treatment of equations with and without conservation laws is the conservation constant $c$. The existence of $c$ prevents the iteration from convergence, unless $c = 0$. Even though this in general not the case, there is still a practical application of this situation: If the linear problems of the Newton method are considered with the linearized original tables (details will be discussed in chapter 6), then clearly the conservation constant is treated to be zero. This means that the Newton update does not change the conservation laws of the nonlinear iterates. For that reason, the same analysis as in section 5.2 is performed.

The main results are:

- A condition for the initial value is presented such that convergence occurs in the first iteration step. But the method is unstable with respect to errors in the calculation of this initial guess.
- Applying the inverse of the fast part of the chemical matrix in addition to the reduction mechanisms does not lead to convergence for practical problems.
- Taking instead the inverse of the fast part of the chemical and physical matrix leads to nice performance, but the solution strategy is too expensive for realistic problems.

### 5.3.1 Preconditioning with QSSA and ILDM

Recall from section 4.3.5 that the reduced solution to the problem $(A + B)u = b$ is given by

$$u = \begin{pmatrix} A_{11} + B_{11} & A_{12} + B_{12} & A_{13} + B_{13} \\ Y_{21} & Y_{22} & Y_{23} \\ V_1^0 & V_2^0 & V_3^0 \end{pmatrix}^{-1} \begin{pmatrix} b_1 \\ 0 \\ c \end{pmatrix}$$

with the conservation constant $c$, which is in general unequal to zero. The preconditioner $P$ is therefore an affine linear mapping given by

$$P(d) = \begin{pmatrix} A_{11} + B_{11} & A_{12} + B_{12} & A_{13} + B_{13} \\ Y_{21} & Y_{22} & Y_{23} \\ V_1^0 & V_2^0 & V_3^0 \end{pmatrix}^{-1} \left( \begin{pmatrix} \mathbb{1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} d + \begin{pmatrix} 0 \\ 0 \\ c \end{pmatrix} \right).$$

This means that the Richardson update $\delta u = P(b - (A + B)u^n)$ is unequal to zero, even if the defect $d = b - (A + B)u$ equals zero. This leads to the divergence of the iteration.

For completeness of the analysis, let now the conservation constant $c$ be zero. Then the update $\delta u$ depends linearly on the defect $d_1$ and the preconditioner can be written in the form

$$P = \begin{pmatrix} A_{11}+\alpha B_{11} & A_{12}+\alpha B_{12} & A_{13}+\alpha B_{13} \\ Y_{21} & Y_{22} & Y_{23} \\ V_1^0 & V_2^0 & V_3^0 \end{pmatrix}^{-1} \begin{pmatrix} \mathbb{1} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{5.4}$$

Applying $P$ and the block matrices

$$\bar{A} := \begin{pmatrix} A_{11} & \begin{pmatrix} A_{12} & A_{13} \end{pmatrix} \\ \begin{pmatrix} Y_{21} \\ V_1^0 \end{pmatrix} & \begin{pmatrix} Y_{22} & Y_{23} \\ V_2^0 & V_3^0 \end{pmatrix} \end{pmatrix} \quad \text{and} \quad \bar{B} := \begin{pmatrix} B_{11} & \begin{pmatrix} B_{12} & B_{13} \end{pmatrix} \\ \begin{pmatrix} B_{21} \\ B_{31} \end{pmatrix} & \begin{pmatrix} B_{22} & B_{23} \\ B_{32} & B_{33} \end{pmatrix} \end{pmatrix}$$

to lemma 5.2.1 shows that the preconditioned Richardson iteration remains constant after the first step. In order to have convergence, the fast part of the initial value has to be

$$\begin{aligned} \begin{pmatrix} u_2^0 \\ u_3^0 \end{pmatrix} = &\left(\bar{A}_{22}+\bar{B}_{22}-(\bar{B}_{21}P_{11}+\bar{B}_{22}P_{21})(\bar{A}_{12}+\bar{B}_{12})\right)^{-1} \\ &\left[b_2-(\bar{A}_{21}+\bar{B}_{21}-(\bar{B}_{21}P_{11}+\bar{B}_{22}P_{21})(\bar{A}_{11}+\bar{B}_{11}))u_1^0-\begin{pmatrix} \bar{B}_{21} & \bar{B}_{22} \end{pmatrix} Pb\right], \end{aligned} \tag{5.5}$$

compare with (5.2). This makes the calculation of the initial value unfortunately very expensive, because a matrix similar to $\bar{A}_{22}+\bar{B}_{22}$ has to be inverted in order to have convergence. Clearly, the calculation of the fast part of the initial guess for the Richardson iteration is very high. The numerical costs are even higher, if the ILDM method is applied, because the left eigenvectors of the matrix $A$ are required. In addition, the iteration is unstable with respect to errors in the computation of the initial value. These errors will be inherited to all following iterates.

### 5.3.2 Preconditioning with modified parameterizations

#### Allow the inversion of the chemical fast part

In section 4.3.5 three possibilities for the evaluation of the tables were introduced, given that the fast part of the matrix $A$ was allowed to be inverted. These possibilities read

$$\phi(u_1) = \psi(u_1) + \begin{pmatrix} A_{22} & A_{23} \\ V_2^0 & V_3^0 \end{pmatrix}^{-1} \begin{pmatrix} rhs_1 \\ rhs_2 \end{pmatrix}$$

with $rhs$ being one of the vectors $(b_2, 0)$, $(0, -c)$ or $(b_2, -c)$.

A necessary condition for the successful application of the parameterizations as preconditioners is that the Richardson update $\delta u$ is zero, if the current iterate $u^n$ solves the equation $(A+\alpha B)u^n = \beta b$ exactly. This is unfortunately not the case for the first parameterization with $rhs_2 = 0$, because the conservation constant $c$ included in $\psi$ is not erased.

The second parameterization can be applied as a preconditioner, because the conservation constant was set to $c = 0$. This leads to the situation of the application of the original $\psi$ given that the conservation constant $c$ is zero.

The application of the parameterization obtained by the right hand side $rhs = (b_2, -c)$ leads to the preconditioner

$$P = \begin{pmatrix} A_{11} + B_{11} & A_{12} + B_{12} & A_{13} + B_{13} \\ A_{21} & A_{22} & A_{23} \\ V_1^0 & V_2^0 & V_3^0 \end{pmatrix}^{-1} \begin{pmatrix} \mathbb{1} & 0 & 0 \\ 0 & \mathbb{1} & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

if the underlying parameterization is created with QSSA. Change the values of $A_{2i}$ to $\Lambda_2 V_{2i}$ for the formulation with the ILDM method.

As an example, consider again the matrix $A + \alpha B$ with the definition of (4.19). Then the spectral radius of the system matrix for the Richardson iteration $S = \mathbb{1} - PA$ is

$$\begin{aligned} \sigma(S) &= 1.003 \quad (\sigma(S) = 1.003) \quad \text{for} \quad \alpha = 1 \\ \sigma(S) &= 3.643 \quad (\sigma(S) = 3.716) \quad \text{for} \quad \alpha = 1000 \end{aligned}$$

for the QSSA– (ILDM–)parameterization. The spectral radius is in both cases clearly greater than one and increases for increasing $\alpha$. This means that the preconditioned Richardson iteration cannot converge.

#### Allow the inversion of the chemical and physical fast part

Even though the evaluation of the tables is only used for the calculation of the slow variables and therefore for the update of the slow variables (compare section 4.3.5), if preconditioning is considered, the conservation constant $c$ prevents the update from being zero, even if the defect is zero. This means that this technique can only be applied as a preconditioner, if the conservation constant $c$ is zero. Otherwise a non–existing defect of the slow equation leads still to an update of the slow variables, which contradicts the convergence.

In order to investigate the applicability of the above introduced solution technique as a preconditioner, assume that the table $\psi$ was created with conservation constant $c = 0$. Then the spectral radius $\sigma(S)$ of the system matrix $S = \mathbb{1} - P(A + \alpha B)$ with the preconditioner $P$ as given by the solution process behaves nicely. It remains small even for large values of $\alpha$:

$$\begin{aligned} \alpha = 1 &\implies \sigma(S) = 0.1508 \\ \alpha = 1000 &\implies \sigma(S) = 0.2848 \\ \alpha = 1\,000\,000 &\implies \sigma(S) = 0.3591 \end{aligned}$$

This indicates that this preconditioner is applicable for a wide range of problems.

### 5.3.3 A PDE example

Reconsider the 1d reaction–diffusion equation given by 4.16 with the linear source term given by the matrix

$$A = \begin{pmatrix} -2 & 1 & 0 \\ 2 & -1001 & 90 \\ 0 & 1000 & -90 \end{pmatrix},$$

which describes the reaction (4.16). In order to use the ILDM– and QSSA–parameterization as preconditioners, the conservation constant has to be equal to zero. This is clearly not the case in this example, because the sum over all mass fractions has to be equal to one, the conservation constant is therefore $c = 1$. This means that the parameterizations cannot be applied as preconditioner to the above system.

If nonlinear reactions are considered with the Newton method, the linearized problem for the calculation of the Newton update has basically the same structure as the considered linear PDE 4.16. Then clearly $c = 0$ and preconditioning can theoretically be performed, but the eigenvalues of the corresponding system matrix are zero and one. This contradicts the convergence.

An eigenvalue analysis of the system matrix $S = \mathbb{1} - P(A_h + B_h)$ is therefore performed only for the preconditioner resulting from the modified parameterizations, where the conservation constant is erased. The spectral radius can be seen in table 5.4. Clearly, $\sigma(S)$ is greater

| # points | $h$ | QSSA | ILDM |
|---|---|---|---|
| 5 | 0.25 | 1.26 | 1.26 |
| 9 | 0.125 | 1.15 | 1.15 |
| 17 | 0.0625 | 1.12 | 1.12 |
| 33 | 0.03125 | 1.15 | 1.15 |
| 65 | 0.015625 | 1.25 | 1.25 |
| 129 | 0.0078125 | 1.48 | 1.48 |

**Table 5.4:** The table shows the spectral radius of the system matrix of the Richardson iteration, if the modified parameterizations $\phi_Q$ and $\phi_I$ are used as preconditioners. Even though the spectral radii seem to be the same for QSSA and ILDM, they differ with the size of order $10^{-6}$.

than one on all grids. Even though the spectral radius shrinks on the first refinements, the value of $\sigma$ increases afterwards, which contradicts the convergence of the iteration.

If the inversion of the chemical and physical fast part is allowed, the spectral radius in table 5.5 is clearly less than one and seems to be less than one for much finer grids. The

| # points | $h$ | QSSA | ILDM |
|---|---|---|---|
| 5 | 0.25 | 0.0087 | 0.0087 |
| 17 | 0.0625 | 0.0083 | 0.0083 |
| 65 | 0.0078125 | 0.0083 | 0.0083 |

**Table 5.5:** The table shows the spectral radius of the system matrix $S$, if the chemical and physical fast part is allowed to be inverted in addition to the application of the tables.

preconditioned iteration will therefore converge fastly.

## 5.4 Preconditioning with time steps

In the previous chapters, the application of the reduction methods QSSA and ILDM as preconditioners leaded to big problems due to the existence of the disturbance $B$. The idea is now to split the operator $A + B$, and treat the effect of $A$ and $B$ to the current iterate separately. The advantage is clearly that the reduction methods can now be applied to the part of the algorithm, which treats only the operator $A$, and apply a detailed solving strategy, if $B$ is considered.

One splitting possibility is provided by the algorithm of Strang [47], which acts on ODEs. The solution to the problem

$$(A + B)h = res := b - (A + B)u^n,$$

which is approximated by the preconditioner, is therefore obtained by solving the ODE

$$\dot{h} = -(A + B)h + res$$

and accepting $h(t)$ for a certain time $t$ as steady state and therefore as approximation for the above equation.

For that reason, the Strang splitting is investigated in this section. First, the method is presented in a very basic form, then the reduction methods will be applied to the chemical part of the iteration. The main result of this section is, that the obtained preconditioner is an approximation of the preconditioners obtained directly from QSSA and ILDM. The Strang splitting can therefore not be applied as preconditioner.

### 5.4.1 Strang splitting

One solving technique for ODEs of the form

$$\dot{u} = Au - e + Bu - f$$

was introduced for nonlinear problems by Strang [47] and is based on the splitting of the right hand side into $Au - e$ and $Bu - f$. Here, $Au$ is supposed to denote the linear chemical source term without conservation laws, $B$ shall describe an arbitrary second operator, possibly describing physical phenomena.

Let then $M_{\Delta t}$ and $N_{\Delta t}$ denote the operators, which describe a numerical time step from $t$ to $t + \Delta t$ by $u(t + \Delta t) = M_{\Delta t}u(t)$ and $v(t + \Delta t) = M_{\Delta t}v(t)$ for the problems

$$\dot{u} = Au - e \quad \text{and} \quad \dot{v} = Bv - f$$

respectively. Then the operators

$$S_{\Delta t} = M_{\Delta t/2} N_{\Delta t} M_{\Delta t/2}$$

and

$$T_{\Delta t} = N_{\Delta t/2} M_{\Delta t} N_{\Delta t/2}$$

do also describe one time-step and are of order one, if $M_{\Delta t}$ and $N_{\Delta t}$ are of order one. The same result holds for second order operators, as shown in [47].

The formulation of $S$ and $T$ means that one time step is divided into three sub-steps. Assume for simplicity that both $M$ and $N$ are explicit operators of first order, and are therefore of the form

$$M_{\Delta t}u = u + \Delta t(Au - e) \quad \text{and} \quad N_{\Delta t}v = v + \Delta t(Bv - f).$$

Then

$$\begin{aligned}
u(t_{n+1/4}) &= u(t_n) + \frac{\Delta t}{2}(Au(t_n) - e) \\
u(t_{n+3/4}) &= u(t_{n+1/4}) + \Delta t(Bu(t_{n+1/4}) + f) \\
&= u(t_n) + \frac{\Delta t}{2}Au(t_n) + \Delta t Bu(t_n) + \frac{\Delta t}{2}e + \Delta t f + O(\Delta t^2) \\
u(t_{n+1}) &= u(t_{n+3/4}) + \frac{\Delta t}{2}(Au(t_{n+3/4}) - e) + O(\Delta t^2) \\
&= u(t_n) + \Delta t(A + B)u(t_n) - \Delta t(e + f) + O(\Delta t^2).
\end{aligned}$$

Clearly the solution process is independent of the way, how the constant vector $e + f$ is splitted.

For general ODEs, the operators $S_{\Delta t}$ and $T_{\Delta t}$ may both be taken for the solution process. This is different, if one of the operators $A$ or $B$ is stiff. Then [24] showed that it is advantageous to have the stiff solver as an outer solver. Let for example $A$ be stiff and $M_{\Delta t}$ the corresponding solution operator, then $S_{\Delta t}$ leads to better results than $T_{\Delta t}$.

One drawback of this splitting technique is the treatment of the stationary point. The stationary point of the ODE will never be reached exactly by a splitted algorithm. The solution does not even remain constant, if the theoretical stationary point is used as initial value, because it changes under the first step $M_{\Delta t}$, which considers only a part of the total right hand side.

### 5.4.2 Reduced Strang splitting

The pure splitting technique introduced by Strang has the possibility to reduce the stiffness of the equations by choosing the splitted operators $A + B$ appropriately, but the computational costs for the solution process of the ODE are not reduced as efficiently as with ILDM. But the clear separation of the chemical and physical part offers a great possibility to apply chemical reduction mechanisms, and to take advantage of both the splitting and the reduction by the parameterizations.

Since the reduced chemical part is assumed not to be stiff anymore, two splitting techniques arise: The outer iteration in one time step can either be the chemical or the physical part.

**Chemical part as outer iteration.**

Let first the chemical part $A$ describe the outer iteration $M_{\Delta t}$ of the Strang splitting. Then one time step is calculated by

$$u_1(t_{n+1}) = u_1(t_n) + \Delta t\Big(A_{11}u_1(t_n) + 1/2 A_{12}\left(\psi(u_1(t_n)) + \psi(u_1(t_{n+3/4}))\right) + B_{11}u_1(t_n) + B_{12}\psi(u_1(t_{n+1/4}))\Big) - \Delta t(e_1 + f_1) + O(\Delta t^2)$$

for the fast and

$$u_2(t_{n+1}) = \psi(u_1(t_{n+1}))$$

for the slow variables.

The above formulation offers the possibility to treat the original tables $\psi_Q$ and $\psi_I$ on the one side, but to consider also the modified parameterizations $\phi_Q$ and $\phi_I$ on the other side:

- The original parameterizations $\psi(u_1)$ lead to the formalisms

  $$\begin{aligned} u_1(t_{n+1}) = u_1(t_n) &+ \Delta t(A_{11} + B_{11})u_1(t_n) \\ &- \Delta t\left(\begin{pmatrix} A_{12} + B_{12} & A_{13} + B_{13}\end{pmatrix}\begin{pmatrix} Y_{22} & Y_{23} \\ V_2^0 & V_3^0\end{pmatrix}^{-1}\begin{pmatrix} Y_{21} \\ V_1^0\end{pmatrix}u_1(t_n)\right) \\ &+ \Delta t\left(\begin{pmatrix} A_{12} + B_{12} & A_{13} + B_{13}\end{pmatrix}\begin{pmatrix} Y_{22} & Y_{23} \\ V_2^0 & V_3^0\end{pmatrix}^{-1}\begin{pmatrix} 0 \\ c\end{pmatrix} - e_1 - f_1\right) + O(\Delta t^2) \end{aligned}$$

  with $Y_{2i} = A_{2i}$ for QSSA and $Y_{2i} = \Lambda_2 V_{2i}$ for ILDM respectively. The new iterate $u_1(t_{n+1})$ is independent of the fast part of the constant $e + f$, therefore $e + f$ may be treated with the chemical or the physical part. The stationary point is then

  $$u(t_\infty) = \begin{pmatrix} A_{11} + B_{11} & A_{12} + B_{12} & A_{13} + B_{13} \\ Y_{21} & Y_{22} & Y_{23} \\ V_1^0 & V_2^0 & V_3^0\end{pmatrix}^{-1}\begin{pmatrix} e_1 + f_1 \\ 0 \\ c\end{pmatrix}$$

  again with $Y_{2i}$ replaced according to the reduction method.

- In order to investigate the influence of the renewed parameterizations to the iteration, take the first possibility of section 5.3.2, where the fast part of the right hand side $b_2$ and the conservation constant $c$ was considered. This leads to the formulation

  $$\begin{aligned} u_1(t_{n+1}) = u_1(t_n) &+ \Delta t(A_{11} + B_{11})u_1(t_n) \\ &- \Delta t\left(\begin{pmatrix} A_{12} + B_{12} & A_{13} + B_{13}\end{pmatrix}\begin{pmatrix} Y_{22} & Y_{23} \\ V_2^0 & V_3^0\end{pmatrix}^{-1}\begin{pmatrix} Y_{21} \\ V_1^0\end{pmatrix}u_1(t_n)\right) \\ &+ \Delta t\left(\begin{pmatrix} A_{12} + B_{12} & A_{13} + B_{13}\end{pmatrix}\begin{pmatrix} Y_{22} & Y_{23} \\ V_2^0 & V_3^0\end{pmatrix}^{-1}\begin{pmatrix} e_2 \\ c\end{pmatrix} - e_1 - f_1\right) + O(\Delta t^2) \end{aligned}$$

This possibility introduces the fast part $e_2$ to the calculation of $u_1(t_{n+1})$, but neglects the influence of $f_2$. Therefore it seems to be advantageous to set $f$ to zero and to have the total vector $e + f$ be treated in the chemical part. Here, the stationary point is

$$u(t_\infty) = \begin{pmatrix} A_{11} + B_{11} & A_{12} + B_{12} & A_{13} + B_{13} \\ Y_{21} & Y_{22} & Y_{23} \\ V_1^0 & V_2^0 & V_3^0 \end{pmatrix}^{-1} \begin{pmatrix} e_1 + f_1 \\ e_2 \\ c \end{pmatrix}$$

with $Y_{2i}$ chosen according to the type of the parameterization.

The preconditioning by doing several time steps is therefore in both cases an approximation to the preconditioners introduced in sections 5.2.1 and 5.2.2, which were not applicable to practical problems, and an improvement of the performance cannot be expected.

**Chemical part as inner iteration.**

Let now the inner time step be driven by the reduced chemical reaction. Then the calculation of one time step reads for the slow variables

$$\begin{aligned} u_1(t_{n+1}) = u_1(t_n) &+ \Delta t(A_{11} + B_{11})u_1(t_n) \\ &+ \Delta t(A_{12}\psi(u_1(t_{n+1/4})) + 1/2 B_{12}(u_2(t_n) + \psi(u_1(t_{n+3/4}))) - (e_1 + f_1)) + O(\Delta t^2) \end{aligned}$$

and for the fast variables

$$u_2(t_{n+1}) = \psi(u_1(t_{n+3/4})) + \Delta t/2(B_{21}u_1(t_n) + B_{22}\psi(u_1(t_{n+3/4})) - f_2) + O(\Delta t^2).$$

In this formulation the new fast iterate does not only depend on the slow iterate via evaluating the parameterizations, but also on the intermediate values of $u_1$ and $u_2$ after the first and second splitted step.

Three interesting facts can be drawn out of this formulation:

- The Richardson defect $e + f$ must be treated in the chemical part. The constant $f_2$ has to be equal to zero, otherwise the solution will not become stationary, if both $A$ and $B$ are of block diagonal form and no conservation laws occur. If this is the case, the original parameterization $\psi(u_1)$ is zero for all $u_1$, therefore $u_2(t_{n+1})$ is given by

  $$u_2(t_{n+1}) = -\Delta t/2 f_2,$$

  which contradicts the existence of a stationary point.

- The modified parameterizations cannot be used. By the same argument as before, the fast part of the new iterate reads

  $$u_2(t_{n+1}) = A_{22}^{-1} e_2 - \Delta t/2 B_{22} A_{22}^{-1} e_2,$$

  if the renewed QSSA–parameterization $\phi_Q$ is applied and the matrices $A$ and $B$ are both of block diagonal form. This means that the solution will never become stationary. A similar analysis for the renewed ILDM–parameterization $\phi_I$ leads to the same result.

- For simplicity of the notations, assume that no conservation laws hold. Then the existence of a stationary point implies $u_2(\infty) = \psi(u_1(\infty))$ by the second equation, hence the relation

$$u_2(\infty) = -Y_{22}^{-1} Y_{21} u_1(\infty)$$

holds. By the first equation, the existence of a steady state implies clearly the equality

$$\begin{pmatrix} A_{11} + B_{11} - A_{12}Y_{22}^{-1}Y_{21} - 1/2 B_{12}Y_{22}^{-1}Y_{21} & 1/2 B_{12} \\ B_{21} - B_{22}Y_{22}^{-1}Y_{21} & 0 \end{pmatrix} \begin{pmatrix} u_1(\infty) \\ u_2(\infty) \end{pmatrix} = \begin{pmatrix} e_1 + f_1 \\ 0 \end{pmatrix}.$$

Both equations together lead to the steady state

$$u(t_\infty) = \begin{pmatrix} A_{11} + B_{11} & A_{12} + B_{12} \\ Y_{21} & Y_{22} \end{pmatrix}^{-1} \begin{pmatrix} e_1 + f_1 \\ 0 \end{pmatrix},$$

where the blocks $Y_{2i}$ are to be chosen according to the type of parameterization.

The steady state

$$u(t_\infty) = \begin{pmatrix} A_{11} + B_{11} & A_{12} + B_{12} & A_{13} + B_{13} \\ Y_{21} & Y_{22} & Y_{23} \\ V_1^0 & V_2^0 & V_3^0 \end{pmatrix}^{-1} \begin{pmatrix} e_1 + f_1 \\ 0 \\ c \end{pmatrix}$$

is obtained by a similar analysis, if a center manifold is existent.

As before, the time steps lead to a steady state, which is equivalent to the formulation with QSSA in section 5.2.2. This means that the time steps as preconditioner are an approximation to a preconditioner, which cannot be applied successfully.

# 6 Quasi–Newton methods with QSSA and ILDM

For nonlinear reaction mechanisms, the application of the reduction methods QSSA and ILDM as solver for the linear problem arising from the Newton iteration shall be investigated. The resulting quasi–Newton methods have the advantage that the convergence of the iterations implies the convergence to the exact solution. This chapter will therefore be concerned with the convergence properties of the various applications of the reduction methods.

The main problem hereby is that the tables $\psi$ are provided to reduce the nonlinear problem $\dot{u} = f(u)$. They have to be linearized in order to be valid for problems of the form $\dot{v} = Df(u)v$. This linearization process contradicts the convergence of the quasi–Newton method even for the relatively simple problem

$$f(u) = 0,$$

where $f$ denotes only the chemical source term. Conditions to $f$ and $\psi$ will be given, such that the quasi–Newton iteration does still lead convergence, but these conditions are unfortunately irrelevant for problems in praxis. Modifications of the application of the tables will neither lead to satisfying results, if more realistic problems $f(u) + g(u) = 0$ are considered.

## 6.1 Convergence aspects for inexact Newton methods

Inexact Newton methods for finding zeros of functions $f : \mathbb{R}^n \to \mathbb{R}^n$ are variations of the exact Newton's method [33], which reads

$$\begin{aligned} Df(u^n)\delta u &= -f(u^n) \\ u^{n+1} &= u^n + \delta u. \end{aligned}$$

This algorithm is due to the creation of the derivative of $f$ and due to the solution process for the linear equation very expensive. Therefore various simplifications were investigated in order to reduce the numerical costs of the iteration. The most commonly known strategy is to keep the derivative $f'$ for a few iterations and save therefore the effort of the assembling of $f'$.

A second possibility is to replace the linear equation leading to the Newton update $\delta u$ by

$$\|Df(u^n)\delta u + f(u^n)\| \leq \theta \|f(u^n)\|, \quad \theta \in [0, 1], \tag{6.1}$$

where the resulting $\delta u$ is clearly not uniquely determined. For details see for example [15, 16]. It can be shown that the Newton method converges, if an $M > 0$ exists such that $\delta u$ fulfilling the above condition, satisfies $\delta u < M$ in all iteration steps. For the proof, see [6].

The Newton update might also be calculated by

$$\widetilde{Df}\delta u = -f(u^n), \tag{6.2}$$

where the matrix $\widetilde{Df}$ is in some sense close to $Df(u^n)$, but much cheaper to be inverted. These methods are called quasi–Newton methods, see for example [23].

In this chapter, the Newton updates will be calculated by (6.2) with $\widetilde{Df}$ given by the reduction methods QSSA and ILDM. A criteria for the convergence can then be the spectral radius

$$\sigma(\mathbb{1} - Df(u^n)(\widetilde{Df})^{-1}).$$

If the spectral radius is zero, then (6.1) is by definition fulfilled and the quasi–Newton method leads to the same performance as the detailed Newton iteration. The conjecture is now that a small spectral radius leads at least to comparable results and that a large spectral radius in an indication for poor convergence or even divergence.

## 6.2 Nonlinear equations without conservation laws

This section investigates the possibilities for the application of the reduction methods as part of the linear solver within the Newton method. It starts on a very basic level, where the Newton algorithm for the problem $f(u) = 0$ is considered, where $f$ represents only the non–disturbed source term of a chemical reaction. Then disturbances will be allowed and the application of pure tables and modified tables will be tested on an algebraic level. The results of this section are substantiated by considering a 1d reaction–diffusion equation.

The main results of this section are:

- The linearization $\nabla\psi$ can be used as reduction table for $\dot{v} = Df(u_1, \psi(u_1)) \cdot v$.
- Even for the simple case $f(u) = 0$, the table $\nabla\psi$ can only be used in very special cases, namely, if $\psi$ is linear and the initial guess for the Newton iteration is already on the manifold given by $\psi$.
- Modified tables are only of value, if $A_{22} + B_{22}$ is allowed to be inverted. But then the algorithm is very expensive.

### 6.2.1 The linearization of the pure source term

In the Newton iteration for the problem

$$f(u) = 0,$$

a linear equation of the form

$$\nabla f(u^n)\delta u = -f(u^n)$$

has to be solved to obtain the Newton update $\delta u$ for the current iterate $u$. In order to apply the reduction methods to this linear problem, the original table $\psi$ shall be transformed to be applied to the linearized problem

$$\begin{aligned} \dot{v} &= \nabla f(u^n)v \\ v(0) &= v_0 \end{aligned}$$

for the current iterate $u^n$. The following lemma shows that the tables $\psi$ can be applied to this linearized problem as well, if $u^n$ is on the manifold defined by $\psi$.

**Lemma 6.2.1** *Let $F$ be in $C^1(\mathbb{R}^m)$ and $\psi \in C^1(\mathbb{R}^{m_1}, \mathbb{R}^{m_2})$ such that*

$$F_2(u_1, \psi(u_1)) = 0.$$

*Then*

$$\left[\nabla F(u_1, \psi(u_1)) \begin{pmatrix} v_1 \\ \nabla\psi(u_1)v_1 \end{pmatrix}\right]_2 = 0 \tag{6.3}$$

*for all $u_1, v_1 \in \mathbb{R}^{n_1}$.*

**Proof:**

$$\begin{aligned} \left[\nabla F(u_1, \psi(u_1)) \begin{pmatrix} v_1 \\ \nabla\psi(u_1)v_1 \end{pmatrix}\right]_2 &= \left(\frac{\partial F_2}{\partial u_1} + \frac{\partial F_2}{\partial u_2}\nabla\psi(u_1)\right) v_1 \\ &= \frac{d}{du_1} F_2(u_1, \psi(u_1))v_1 = 0, \end{aligned}$$

because $F_2(u_1, \psi(u_1))$ is constant in $u_1$ by assumption. ∎

This lemma shows that the QSSA–table for the ODE

$$\dot{v} = \nabla f(u_1, \psi(u_1))v$$

is defined by the function

$$\phi_Q(v_1) = \nabla\psi_Q(u_1)v_1$$

by applying the function $f$ to the above lemma. A closer look to the proof gives also a relation between the gradient of $\psi$ and the source term $f$:

$$\nabla\psi_Q(u_1)v_1 = -\left(\frac{\partial f_2}{\partial u_2}(u_1, \psi_Q(u_1))\right)^{-1} \frac{\partial f_2}{\partial u_1}(u_1, \psi_Q(u_1))v_1. \tag{6.4}$$

Lemma 6.2.1 can also be applied for ILDM–tables, if $f$ is chosen such that the matrix $V$ containing the left eigenvectors is independent of $u$. Let then $F$ be defined such that $F_1(u) = f_1(u)$ and

$$F_2(u) = \begin{pmatrix} V_{11} & V_{12} \end{pmatrix} f(u),$$

and let the ILDM–table $\psi_I$ for problem (4.20) map the slow variables $u_1$ to the fast variables $u_2$. Then the ILDM–table $\phi_I$ for the linearized problem is given by

$$\phi_I(v_1) = \nabla\psi_I(u_1) \cdot v_1$$

for all possible $u_1$.

With these parameterizations $\phi = \nabla\psi$, the problem

$$Df(u_1^n, \psi(u_1^n))\delta u = -f(u_1^n, \psi(u_1^n))$$

can now be exactly reduced. This means that the Newton iteration with reduced linear problem has the same convergence properties as the exact Newton iteration, if all iterates $u^n$ are on the manifold given by $\psi$. This is the case, if

- the starting value $u^0$ fulfills the relation $u_2^0 = \psi(u_1^0)$. Note that the iteration is instable with respect to errors in the evaluation of $\psi$.
- the table $\psi$ is linear.

The question, which arises out of this observation, is of course, how the violation of these conditions effects the Newton iteration. For that reason, two examples are considered:

- The nonlinear reaction mechanism (4.21) has a linear QSSA–table, see (4.22). The Newton iteration with reduced linear solver and initial guess $u^0 = (1, 1, 1)$ performs nicely only for the process variables. The residual for the fast variables remains almost constant. For details, see table 6.1, where also the spectral radius of the matrix $\mathbb{1} - (\widetilde{Df}) \cdot Df$ can be found. $\widetilde{Df}$ describes the approximative inverse of the Jacobian of $f$.
- The reaction mechanism given by

$$2sp_1 \xrightarrow{2} sp_2 \xrightarrow{1000} sp_3 \xrightarrow{500} sp_4 \tag{6.5}$$

  has the source term

$$g(u) = \begin{pmatrix} -4u_1^2 \\ 4u_1^2 - 1000u_2 \\ 1000u_2 - 500u_3 \end{pmatrix} \tag{6.6}$$

  for the first three species. The corresponding QSSA–table is clearly

$$\begin{pmatrix} u_2 \\ u_3 \end{pmatrix} = \psi_Q(u_1) = \begin{pmatrix} 1/250 \\ 1/125 \end{pmatrix} u_1^2.$$

  Take now the reduced Newton iteration in order to solve the equation $g(u) = 0$ with the initial value $u^0 = (1, \psi_Q(1))$, such that the fast part of the initial residual is zero. Table 6.1 shows, how the iteration performs.

| | $f(u)=0$ | | | $g(u)=0$ | | |
|---|---|---|---|---|---|---|
| iteration | $\|res_1\|$ | $\|res_2\|$ | $\sigma$ | $\|res_1\|$ | $\|res_2\|$ | $\sigma$ |
| 0 | $2.00e+00$ | $1.99e+03$ | 1.00 | $4.00e+00$ | $0.00e+00$ | 1.00 |
| 1 | $2.62e-01$ | $1.95e+03$ | 1.00 | $3.36e+00$ | $1.34e+00$ | 1.00 |
| 2 | $8.29e-03$ | $1.94e+03$ | 1.00 | $3.22e+00$ | $1.62e+00$ | 1.00 |
| 3 | $9.43e-06$ | $1.94e+03$ | 1.00 | $3.22e+00$ | $1.62e+00$ | 1.00 |
| 4 | $1.23e-11$ | $1.94e+03$ | 1.00 | $3.22e+00$ | $1.62e+00$ | 1.00 |
| 5 | $2.22e-16$ | $1.94e+03$ | 1.00 | $3.22e+00$ | $1.62e+00$ | 1.00 |
| 6 | $2.22e-16$ | $1.94e+03$ | 1.00 | $3.22e+00$ | $1.62e+00$ | 1.00 |

**Table 6.1:** The performance of the Newton iteration with reduced linear solver for the problems $f(u)=0$ (left) and $g(u)=0$ (right). The initial guess is for the left table chosen not to be on the manifold, the initial guess for the right table fulfills the relation $u_2^0=\psi(u_1^0)$. The function $f$ denotes the source term of (4.21), $g$ is given by (6.6).

### Modified linearized tables

The above investigations demand a different treatment of the tables, because the fast part of the residual seems to be necessary to be considered. Let the linear equation for the Newton update be written in the form

$$A\delta u = res.$$

Then the treatment of the fast residual $res_2$ for the slow Newton update offers two possibilities:

1. Apply $A_{22}^{-1}$ to both the slow and the fast variables:

$$\begin{aligned}(A_{11}+A_{12}\nabla\psi)\delta u_1 &= res_1 - A_{12}A_{22}^{-1}res_2\\ \delta u_2 &= \nabla\psi\delta u_1 + A_{22}^{-1}res_2\end{aligned}$$

2. Accept the pure table for the slow variables and do not use the tables in the calculation of the fast Newton update:

$$\begin{aligned}(A_{11}+A_{12}\nabla\psi)\delta u_1 &= res_1\\ A_{22}\delta u_2 &= res_2 - A_{21}\delta u_1\end{aligned}$$

Note that the first and third possibility are closely related: If the source term $f$ is linear, they are equivalent, because then $\nabla\psi = -A_{22}^{-1}A_{21}$ holds. The development of the residuals in the Newton iteration for the reaction mechanism (4.21) can be seen in table 6.2. Here, the table $\psi$ represents a linear function, but the condition $u_2^0=\psi(u_1^0)$ for the initial value is violated. Clearly, the residual of the fast part decreases now, such that convergence occurs. The consideration of the fast part of the residual in the linear solver leads even to the same convergence properties as the detailed Newton solver.

| | 1. possibility | | | 2. possibility | | |
|---|---|---|---|---|---|---|
| iteration | $\|res_1\|$ | $\|res_2\|$ | $\sigma$ | $\|res_1\|$ | $\|res_2\|$ | $\sigma$ |
| 0 | $4.00e+00$ | $0.00e+00$ | 0.00 | $4.00e+00$ | $0.00e+00$ | 0.00 |
| 1 | $1.00e+00$ | $1.00e+00$ | 0.00 | $1.00e+00$ | $1.00e+00$ | 0.00 |
| 2 | $2.50e-01$ | $2.50e-01$ | 0.00 | $2.50e-01$ | $2.50e-01$ | 0.00 |
| 3 | $6.25e-02$ | $6.25e-02$ | 0.00 | $6.25e-02$ | $6.25e-02$ | 0.00 |
| 4 | $1.56e-02$ | $1.56e-02$ | 0.00 | $1.56e-02$ | $1.56e-02$ | 0.00 |
| 5 | $3.91e-03$ | $3.91e-03$ | 0.00 | $3.91e-03$ | $3.91e-03$ | 0.00 |
| 6 | $9.77e-04$ | $9.77e-04$ | 0.00 | $9.77e-04$ | $9.77e-04$ | 0.00 |

**Table 6.2:** The table shows the evolution of the residuals in the Newton iteration for the problem $g(u) = 0$ with $g$ denoting the source term of the reaction mechanism (6.5). The initial value is chosen to be $u^0 = (1, \psi(1))$.

### 6.2.2 The linearization of disturbed equations

Instead of the purely chemical equation $f(u) = 0$ in the previous section, the disturbed equation

$$f(u) + g(u) = 0$$

will be now considered and the applicability of the table $\psi$ for the linear problem arising from the Newton's method investigated. This investigation will be performed for three different possibilities:

1. Use only $\nabla\psi$ for the reduction. This is the cheapest method, but numerical differentiation of the table has to be done anyway.
2. Use $\nabla\psi$ for the reduction and allow the inversion of the matrix $\frac{\partial f_2}{\partial u_2}(u_1^n, u_2^n)$. The inversion of this matrix leads of course to higher numerical expenses.
3. Finally, allow the inversion of $\frac{\partial(f_2+g_2)}{\partial u_2}(u_1^n, u_2^n)$. This is the numerically most expensive possibility, if $f$ and $g$ represent discretized operators from a partial differential equation.

In the following, these possibilities will be presented with the example

$$f(u) + \alpha \begin{pmatrix} -2 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{pmatrix} u = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \tag{6.7}$$

with $\alpha = 1$ and $\alpha = 1000$.

#### Application of the tables without inversion

If the Newton update is calculated without the inversion of any of the matrix blocks, the formula for $\delta u$ reads

$$\left(\frac{\partial(f_1+g_1)}{\partial u_1} + \frac{\partial(f_1+g_1)}{\partial u_2}\nabla\psi\right)\delta u_1 = res_1$$
$$\delta u_2 = \nabla\psi\,\delta u_1,$$

where functions and derivatives are evaluated at the point $(u_1^n, u_2^n)$. Obviously, the fast part of the residual is not considered, therefore this possibility will not lead to convergence, see table 6.3. In the above example, the residual decreases fastly in the first few steps, but

| | $\alpha = 1$ | | | $\alpha = 1000$ | | |
|---|---|---|---|---|---|---|
| iteration | $\|res_1\|$ | $\|res_2\|$ | $\sigma$ | $\|res_1\|$ | $\|res_2\|$ | $\sigma$ |
| 0 | $4.95e+00$ | $9.69e-01$ | 1.00 | $1.95e+03$ | $9.69e+02$ | 1.00 |
| 1 | $9.89e-01$ | $4.87e-01$ | 1.00 | $3.97e+00$ | $2.47e+00$ | 1.00 |
| 2 | $1.10e-01$ | $3.26e-01$ | 1.00 | $1.66e-05$ | $4.97e-01$ | 1.00 |
| 3 | $2.24e-03$ | $3.03e-01$ | 1.00 | $2.67e-16$ | $4.97e-01$ | 1.00 |
| 4 | $1.01e-06$ | $3.03e-01$ | 1.00 | $4.51e-17$ | $4.97e-01$ | 1.00 |
| 5 | $2.05e-13$ | $3.03e-01$ | 1.00 | $4.51e-17$ | $4.97e-01$ | 1.00 |

**Table 6.3:** The table shows the evolution of the residuals in the Newton iteration for problem (6.7). The initial value is chosen to be $u^0 = (1, \psi(1))$.

remains constant for the fast variables afterwards.

#### Application of the tables with inversion of the chemical part

The inversion of the matrix $\frac{\partial f_2}{\partial u_2}(u_1^n, u_2^n)$ gives the possibility to calculate the Newton update by

$$\left(\frac{\partial(f_1+g_1)}{\partial u_1} + \frac{\partial(f_1+g_1)}{\partial u_2}\nabla\psi\right)\delta u_1 = res_1 - \frac{\partial(f_1+g_1)}{\partial u_2}\left(\frac{\partial f_2}{\partial u_2}(u_1^n, u_2^n)\right)^{-1} res_2$$

$$\delta u_2 = \nabla\psi\,\delta u_1 + \left(\frac{\partial f_2}{\partial u_2}(u_1^n, u_2^n)\right)^{-1} res_2.$$

Here, the fast part of the residual $res_2$ does have an effect on the Newton iterates, therefore convergence might occur. Table 6.4 shows clearly that this conjecture is true for small values

| | $\alpha = 1$ | | | $\alpha = 1000$ | | |
|---|---|---|---|---|---|---|
| iteration | $\|res_1\|$ | $\|res_2\|$ | $\sigma$ | $\|res_1\|$ | $\|res_2\|$ | $\sigma$ |
| 0 | $4.95e+00$ | $9.69e-01$ | 0.20 | $1.95e+03$ | $9.69e+02$ | 1.00 |
| 1 | $9.87e-01$ | $6.28e-02$ | 0.33 | $1.79e+03$ | $7.47e+02$ | 1.00 |
| 2 | $1.09e-01$ | $1.69e-02$ | 0.42 | $5.39e+02$ | $8.52e+02$ | 1.00 |
| 3 | $2.22e-03$ | $3.21e-03$ | 0.44 | $2.75e+02$ | $8.52e+02$ | 1.00 |
| 4 | $1.02e-06$ | $7.66e-05$ | 0.44 | $2.75e+02$ | $8.52e+02$ | 1.00 |

**Table 6.4:** The table shows the evolution of the residuals in the Newton iteration with reduced linear solver and modified tables for problem (6.7). The initial value is chosen to be $u^0 = (1, \psi(1))$.

of $\alpha$. If the influence of the function $g$ is too big, the iteration does not converge.

#### Application of the tables with inversion of the chemical and physical part

The formulation of the equations for the Newton update with inversion of the block matrix $\frac{\partial(f_2+g_2)}{\partial u_2}(u_1^n, u_2^n)$ offers three possibilities. Calculate the fast Newton update with

$$\delta u_2 = \left(\frac{\partial(f_2+g_2)}{\partial u_2}\right)^{-1}(res_2 - \frac{\partial(f_2+g_2)}{\partial u_1}\delta u_1),$$

given that the slow update $\delta u_1$ was calculated by the application of $\nabla\psi$. This strategy leads to convergence for both values of $\alpha$, see table 6.5.

| | $\alpha = 1$ | | | $\alpha = 1000$ | | |
|---|---|---|---|---|---|---|
| iteration | $\|res_1\|$ | $\|res_2\|$ | $\sigma$ | $\|res_1\|$ | $\|res_2\|$ | $\sigma$ |
| 0 | $4.95e+00$ | $9.69e-01$ | 0.00 | $1.95e+03$ | $9.69e+02$ | 0.29 |
| 1 | $9.86e-01$ | $2.27e-01$ | 0.00 | $2.52e+00$ | $3.13e-01$ | 0.32 |
| 2 | $1.07e-01$ | $3.85e-02$ | 0.00 | $6.79e-01$ | $1.36e-03$ | 0.32 |
| 3 | $1.65e-03$ | $1.41e-03$ | 0.00 | $2.15e-01$ | $1.39e-04$ | 0.32 |
| 4 | $2.02e-05$ | $5.41e-07$ | 0.00 | $6.81e-02$ | $1.40e-05$ | 0.32 |
| 5 | $1.87e-08$ | $1.35e-11$ | 0.00 | $2.16e-02$ | $1.40e-06$ | 0.32 |
| 6 | $9.34e-12$ | $8.89e-17$ | 0.00 | $6.82e-03$ | $1.40e-07$ | 0.32 |

**Table 6.5:** The table shows the evolution of the residuals in the Newton iteration with reduced linear solver and detailed linear solver for the fast Newton update. The considered problem is given by (6.7), the initial value is chosen to be $u^0 = (1, \psi(1))$.

### 6.2.3 A more advanced example

Let now the table $\psi$ be applied to the linear problems arising from the Newton's method. First, the pure application of the tables will be investigated, the additional inversion shall be allowed.

#### Application of the tables without inversion.

As before, the application of the tables does not consider the fast part of the residual in the calculation for the Newton update, if no additional inversion of the matrices is allowed. This contradicted the convergence of the Newton iteration in the simple examples in the previous section and does also prevent the iteration from convergence in the above PDE example. A typical phenomenon is the decreasing residual in the first few Newton steps, but also the constance afterwards. Let for example $\nu$ be given by

$$\nu = \begin{pmatrix} 10^{-4} & 0 & 0 \\ 0 & 10^{-5} & 0 \\ 0 & 0 & 10^{-6} \end{pmatrix}$$

and $\Omega$ be discretized with 129 nodes. Then the residual decreases from 1.87 to $1.06 \cdot 10^{-3}$ in 7 Newton steps, but does not decrease any further. The grid size does not change this behavior significantly.

**Application of the tables with inversion of the chemical part.**

The treatment of the fast part of the source term's gradient has a negative effect on the convergence properties of the Newton iteration. In general, the iteration diverges, which is shown in four configurations of problem (4.24).

- Let the boundaries be $u(0) = u(1) = (0.9, \psi(0.9))$ and the viscosities $\nu_i = 1$. The Newton iteration with detailed linear solver converges then after 3 steps independently of the grid size, whereas the iteration with reduced linear solver diverges. On all grids, the residual decreases in the first few steps and increases afterwards. For the residuals on a grid with 33 nodes, see table 6.6. The values for the residual are not surprising, if the spectral radius $\sigma(S) = \sigma(\mathbb{1} - P(A + B))$ is known.

- Take now the viscosities to be $\nu_i = 0.1$ with boundaries as above. Then the Newton iteration with detailed linear solver converges after 4 steps on each grid. If the linear problem is reduced, the iteration diverges on all grids. The evolution of the residuals and the spectral radius of the matrix $\mathbb{1} - P(A + B)$ is shown in table 6.6.

- Now boundaries are considered, which are not on the manifold described by $\psi$. Let therefore $u(0) = u(1) = (0.5, 0.3, 0.2)$ and the viscosities $\nu_i = 1$. With detailed linear solver, the Newton iteration needs 6 steps to reach convergence. The reduced linear solver leads to divergence.

- Finally, the viscosities $\nu_i = 0.1$ are treated with the new boundaries. The Newton's method with detailed linear solver needs 7 steps to converge on every grid and the reduced linear solution leads again to divergence. The residuals of the iteration steps can be seen in table 6.6.

| | 1. configuration | | 2. configuration | | 3. configuration | | 4. configuration | |
|---|---|---|---|---|---|---|---|---|
| step | residual | $\sigma(S)$ | residual | $\sigma(S)$ | residual | $\sigma(S)$ | residual | $\sigma(S)$ |
| 0 | 0.56 | $7.7e-8$ | 0.56 | $2.5e-8$ | 37.8 | 11.1 | 37.8 | 1.11 |
| 1 | 0.31 | $9.3e-4$ | 8.61 | 7.34 | 22.7 | 76.8 | 915.4 | 0.77 |
| 2 | 0.39 | 1308 | 8.80 | 30.5 | 22.6 | 60.5 | 915.8 | 0.76 |
| 3 | 5.30 | 141.8 | 9.91 | 55.6 | 22.4 | 92.5 | 916.3 | 0.75 |
| 4 | 5.27 | 372.6 | 10.08 | 49.3 | 27.1 | 482.0 | 916.7 | 1.89 |
| 5 | 5.12 | $5.0e+4$ | 10.31 | 2.78 | 29.0 | 152.9 | 917.1 | 2.08 |
| 6 | 292.91 | — | 10.37 | — | 41.2 | — | 917.5 | — |

**Table 6.6:** The table shows the residuals of the Newton iteration with reduced linear solver including the inversion of the chemical fast part. The boundaries and viscosities are as described in the four configurations.

**Application of the tables with inversion of the chemical and physical part.**

Even though the purely algebraic problem in the previous section suggested that the Newton iteration with inversion of the chemical and physical gradient might diverge, the convergence properties in the reaction–diffusion example are very nice. The number of Newton steps required to obtain a residual of order $10^{-6}$ are shown in table 6.7. This table suggests that

| Boundaries: $u_1 = 0.9$, $(u_2, u_3) = \psi(0.9)$ | | | |
|---|---|---|---|
| viscosities | 33 nodes | 65 nodes | 129 nodes |
| $\nu_1 = \nu_2 = \nu_3 = 1$ | 3 (3) | 3 (3) | 3 (3) |
| $\nu_1 = \nu_2 = \nu_3 = 10^{-2}$ | 6 (6) | 6 (6) | 6 (6) |
| $\nu_1 = 10^{-2}$, $\nu_2 = \nu_3 = 10^{-4}$ | 6 (6) | 6 (6) | 6 (6) |
| $\nu_1 = 10^{-2}$, $\nu_2 = \nu_3 = 1$ | 6 (6) | 6 (6) | 6 (6) |
| Boundaries: $u_1 = 0.5$, $u_2 = 0.3$, $u_3 = 0.2$ | | | |
| viscosities | 33 nodes | 65 nodes | 129 nodes |
| $\nu_1 = \nu_2 = \nu_3 = 1$ | 6 (6) | 6 (6) | 6 (6) |
| $\nu_1 = \nu_2 = \nu_3 = 10^{-2}$ | 9 (9) | 9 (9) | 9 (9) |
| $\nu_1 = 10^{-2}$, $\nu_2 = \nu_3 = 10^{-4}$ | 9 (9) | 9 (9) | 9 (9) |
| $\nu_1 = 10^{-2}$, $\nu_2 = \nu_3 = 1$ | 7 (7) | 7 (7) | 7 (7) |

**Table 6.7:** The table shows the number of Newton steps required for the reaction–diffusion equation (4.24) for different viscosities on different grids. Here, the linear problem is reduced with $\nabla\psi$ and the inversion of the chemical and physical matrix. The number in brackets is the number of required detailed Newton steps.

the size of the grid has no influence on the convergence properties, neither do the viscosities $\nu_2$ and $\nu_3$ of the fast variables. The number of required Newton steps seems to depend only on the viscosity $\nu_1$ of the process variables and the chosen boundary conditions. But these choices do not change the performance of the Newton iteration, if the linear part is reduced.

An interesting side effect of this method is that the convergence of the Newton's method is independent of the choice of the boundary values, whereas the errors of the direct reduction of the nonlinear equations depends highly on the choice of the boundaries.

### 6.2.4 Numerical costs

The computation times for the solution process are studied with the problem

$$-\Delta u = f(u) \tag{6.8}$$

on the domain $\Omega = [0,1] \times [0,1]$ with Dirichlet boundaries. The function $f$ denotes the source term of reactions of the type 6.5, where intermediate species are added between $sp_3$ and $sp_4$ in order to investigate the computation time for bigger reaction mechanisms.

The problem shall be discretized by the finite element method on a grid with 1024 cells, the resulting nonlinear algebraic equations are solved by the Newton's method. The linear problems are solved by a multigrid strategy, the smoothing is performed by few preconditioned Richardson steps, where the ILU is taken as a preconditioner.

The computation times depending on the number of fast variables can be seen in table 6.8 for the detailed and reduced solver. It depicts clearly that the computation costs for the

| | detailed | | | | reduced | | | |
|---|---|---|---|---|---|---|---|---|
| $n_2$ | Newton | total | matrix | ilu | Newton | total | matrix | ilu |
| 2 | 4 | 9 | 2.8 | 0.7 | 2 | 68 | 45 | 0.24 |
| 4 | 2 | 81 | 7.47 | 64 | 2 | 73 | 49 | 0.2 |
| 9 | 2 | 1031 | 45 | 961 | 2 | 97 | 66 | 0.2 |
| 14 | 2 | 4956 | 140 | 4762 | 2 | 130 | 91 | 0.21 |

**Table 6.8:** The computational costs for the nonlinear problem (6.8). Depending on the number of fast variables $n_2$, the required number of Newton steps are plotted as well as the computation costs in seconds. The computation time is splitted into the total time, the time for the evaluation of the matrix and the computation of the ILU.

detailed solver grow faster than with $n^3$, whereas the costs for the reduced solver grow less than linearly in the number of species. Most of the computational costs for the linear solver has to be spent for the computation of the ILU. Since only one process variable is to be computed, the computation of the ILU is rather cheap in the reduced case.

The computation time for the Newton method, where the inversion of $A_{22} + B_{22}$ is allowed in the linear solver, can be seen in table 6.9. Except for the smallest reaction system, the

| $n_2$ | Newton | total | $\text{matrix}_1$ | $\text{ilu}_1$ | $\text{matrix}_2$ | $\text{ilu}_2$ |
|---|---|---|---|---|---|---|
| 2 | 5 | 204 | 101 | 0.5 | 9.0 | 5.3 |
| 4 | 2 | 116 | 41.4 | 0.2 | 8.0 | 28.0 |
| 9 | 2 | 840 | 57 | 0.2 | 43.0 | 657 |
| 14 | 2 | 4070 | 80.0 | 0.2 | 136 | 3752 |

**Table 6.9:** The computational costs for the nonlinear problem (6.8), if the inversion of $A_{22} + B_{22}$ is allowed in the linear solver. Depending on the number of fast variables $n_2$, the required number of Newton steps are plotted as well as the computation costs in seconds. The computation time is splitted into the total time, the time for the evaluation of the matrix and the computation of the ILU.

computation is almost as expensive as with the detailed solver.

## 6.3 Nonlinear equations including conservation laws

The difference between section 6.2 and this section is the existence of the center manifold. Here, this difference is not too important, because the linearization $\nabla\psi$ is taken as reduction method for the linear problem within the Newton method, and the conservation laws considered in $\nabla\psi$ have the constant $c = 0$. But still, differences to the previous section occur due to the replacement of the blocks $A_{3,i}$ by the left eigenvectors to the eigenvalue $\lambda = 0$. The effect of this difference is shown in this section.

The main results are comparable to the main results of the previous section.

### 6.3.1 The linearization

Similar as in section 6.2.1, the nonlinear table $\psi$ can under certain circumstances be used to reduce the linear problem arising from the Newton iteration. The table for the problem

$$Df(u_1, \psi(u_1))x = b$$

is given by

$$\phi(x_1) = \nabla\psi(u_1)x_1,$$

as can be seen in lemma 6.2.1, if $F$ is defined such that

$$F_1(u) = f_1(u) \quad \text{and} \quad F_2(u) = \begin{pmatrix} f_2(u) \\ V_1^0 u_1 + V_2^0 u_2 + V_3^0 u_3 - c \end{pmatrix}.$$

Since the formulation of $\phi$ depends on the derivative of $F_2$, the constant $c$ has clearly be equal to zero for the linearized problem. But this is not necessarily a restriction, especially, if the reduction method is used for the calculation of the Newton update. The Newton update is supposed not to change the conservation constant for the nonlinear problem, therefore

$$\begin{pmatrix} V_1^0 & V_2^0 & V_3^0 \end{pmatrix} \delta u = 0$$

has to hold.

This new table $\phi$ can now be applied to reduce the linear problem without systematic errors, if

- the starting value $u^0$ is on the manifold and fulfills the equation $u_2^0 = \psi(u_1^0)$,
- the table $\psi$ is affine linear.

The linearity of $\psi$ assures that the first and the following iterates are also on the manifold, if this is true for the initial guess $u^0$.

In order to see the effect of the violation of these conditions to the Newton iteration, consider the problem

$$f(u) = 0$$

with $f$ denoting the source term of reaction (3.6), which has the nonlinear QSSA–table (4.26). The initial value for the Newton iteration shall be chosen to be $u^0 = (0.5, \psi(0.5))$, which is on the manifold with the conservation constant $c = 1$. The behavior of the residuals in the iteration can be seen in table 6.10. The residual is reduced in the first iteration step, then the residual for the process variables decreases further, whereas the fast residual increases again. The whole iteration remains almost constant after a few iteration steps, eve though the residual is still remarkably big.

| iteration | $\|res_1\|$ | $\|res_2\|$ | $\sigma$ |
|---|---|---|---|
| 0 | $9.73e-01$ | $9.73e-01$ | 1.00 |
| 1 | $3.99e-01$ | $4.33e-01$ | 1.00 |
| 2 | $3.34e-01$ | $4.67e-01$ | 1.00 |
| 3 | $3.15e-01$ | $4.79e-01$ | 1.00 |
| 4 | $3.11e-01$ | $4.82e-01$ | 1.00 |
| 5 | $3.09e-01$ | $4.83e-01$ | 1.00 |
| 6 | $3.08e-01$ | $4.83e-01$ | 1.00 |

**Table 6.10:** The performance of the Newton iteration for the problem $f(u) = 0$ ($f$ denotes the source term of reaction (3.6)), if the linear problem is reduced on the basis of the nonlinear table $\psi$. The initial value is chosen to be $u^0 = (0.5, \psi(0.5))$.

#### Modified linearized tables

Fortunately, the pure application of the tables not the only possibility to apply the tables $\psi$ to the linear problem. Three more possibilities arise, if in addition to the calculation of $\nabla\psi$ the inversion of parts of the Jacobian $Df$ is allowed. Let the linear equation to be solved for the Newton update $\delta u$ be written as

$$A\,\delta u = res.$$

Then the three possibilities on top of the pure reduction read:

1. Accept the pure table for the slow variables and apply the inverse of the fast part only for the fast update:

$$\begin{aligned}(A_{11} + \begin{pmatrix} A_{12} & A_{13}\end{pmatrix} \nabla\psi)\delta u_1 &= res_1 \\ \begin{pmatrix}\delta u_2 \\ \delta u_3\end{pmatrix} &= \nabla\psi\delta u_1 + \begin{pmatrix} A_{22} & A_{23} \\ V_2^0 & V_3^0\end{pmatrix}^{-1} \begin{pmatrix} res_2 \\ 0\end{pmatrix}\end{aligned}$$

2. Accept the pure table for the slow variables and do not use the tables in the calculation of the fast Newton update:

$$\begin{aligned}(A_{11} + \begin{pmatrix} A_{12} & A_{13}\end{pmatrix} \nabla\psi)\delta u_1 &= res_1 \\ \begin{pmatrix} A_{22} & A_{23} \\ A_{32} & A_{33}\end{pmatrix} \begin{pmatrix}\delta u_2 \\ \delta u_3\end{pmatrix} &= \begin{pmatrix} res_2 \\ res_3\end{pmatrix} - \begin{pmatrix} A_{21} \\ A_{31}\end{pmatrix} \delta u_1\end{aligned}$$

   Here, the conservation laws are only treated for the slow Newton update.

The performance of the Newton iteration to the problem $f(u) = 0$ with the Newton updates calculated by the above reductions can be seen in table 6.11. The first iterations seem to show that convergence occurs only for the second possibility. Here, the convergence rate is even better than for the detailed Newton method. But the situation changes after a few more iteration steps, the strategy according to the first possibility converges now.

| | 1. possibility | | 2. possibility | | detailed Newton | |
|---|---|---|---|---|---|---|
| iteration | $\|res_1\|$ | $\|res_2\|$ | $\|res_1\|$ | $\|res_2\|$ | $\|res_1\|$ | $\|res_2\|$ |
| 0 | $4.00e+00$ | $4.00e+00$ | $4.00e+00$ | $4.00e+00$ | $4.00e+00$ | $4.00e+00$ |
| 1 | $3.35e+00$ | $2.40e+00$ | $3.35e+00$ | $2.40e+00$ | $3.35e+00$ | $2.40e+00$ |
| 2 | $2.97e+00$ | $2.18e+00$ | $2.60e+00$ | $1.94e+00$ | $2.81e+00$ | $2.15e+00$ |
| 3 | $2.63e+00$ | $2.33e+00$ | $1.78e+00$ | $1.84e+00$ | $1.93e+00$ | $2.17e+00$ |
| 4 | $2.42e+00$ | $2.47e+00$ | $4.26e-01$ | $3.18e-01$ | $4.56e-01$ | $5.73e-01$ |
| 5 | $2.28e+00$ | $2.55e+00$ | $8.84e-02$ | $1.43e-01$ | $1.82e-01$ | $1.29e-01$ |
| 6 | $2.15e+00$ | $2.63e+00$ | $1.11e-02$ | $8.27e-03$ | $2.84e-02$ | $4.11e-02$ |
| ⋮ | | ⋮ | | ⋮ | | ⋮ |
| 10 | $1.13e+00$ | $2.96e+00$ | $3.97e-15$ | $7.11e-15$ | $1.39e-17$ | $2.01e-14$ |
| 11 | $2.58e-01$ | $1.44e+00$ | $3.57e-15$ | $0.00e+00$ | $0.00e+00$ | $0.00e+00$ |
| 12 | $4.50e-02$ | $9.43e-02$ | $3.57e-15$ | $7.11e-15$ | $0.00e+00$ | $0.00e+00$ |
| 13 | $3.27e-03$ | $5.14e-03$ | $3.30e-15$ | $0.00e+00$ | $0.00e+00$ | $0.00e+00$ |
| 14 | $2.37e-05$ | $3.44e-05$ | $3.29e-15$ | $7.11e-15$ | $0.00e+00$ | $0.00e+00$ |

**Table 6.11:** The table shows the evolution of the residuals for the slow and fast variables for the problem $f(u) = 0$ with initial value $u^0 = (1, \psi(1))$. The Newton update is calculated with the reduction possibilities described above.

### 6.3.2 The linearization of disturbed equations

Instead of the equation $f(u) = 0$, a disturbed equation

$$f(u) + g(u) = 0$$

is now considered.

The applicability of the original table $\psi$ will be investigated for three different cases:

1. Use only $\nabla\psi$ for the reduction. This is numerically the cheapest method.
2. Use $\nabla\psi$ for the reduction and allow the inversion of the matrix $\frac{\partial f_2}{\partial u_2}(u_1^n, u_2^n)$. The inversion of this matrix leads of course to higher numerical expenses.
3. Finally, allow the inversion of $\frac{\partial (f_2+g_2)}{\partial u_2}(u_1^n, u_2^n)$. Even though this is of equal numerical expenses as the previous possibility, the treatment of $g_2$ introduces additional coupling, if $g$ denotes a discretized differential operator.

In the following, these possibilities will be presented with the example

$$f(u) + \alpha \begin{pmatrix} -2 & 1 & 0 \\ 1 & -2 & 1 \\ 0 & 1 & -2 \end{pmatrix} u = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \tag{6.9}$$

with $\alpha = 1$ and $\alpha = 1000$. The function $f$ will again be given by the source term of the reaction (3.6).

| | $\alpha = 1$ | | | $\alpha = 1000$ | | |
|---|---|---|---|---|---|---|
| iteration | $\|res_1\|$ | $\|res_2\|$ | $\sigma$ | $\|res_1\|$ | $\|res_2\|$ | $\sigma$ |
| 0 | $4.95e+00$ | $4.22e+00$ | 1.00 | $1.96e+03$ | $8.72e+02$ | 1.00 |
| 1 | $4.15e+00$ | $2.97e+00$ | 1.00 | $1.29e+03$ | $1.29e+03$ | 1.00 |
| 2 | $3.67e+00$ | $3.06e+00$ | 1.00 | $6.56e+02$ | $1.49e+03$ | 1.00 |
| 3 | $3.52e+00$ | $3.16e+00$ | 1.00 | $5.79e+02$ | $1.51e+03$ | 1.00 |
| 4 | $3.42e+00$ | $3.25e+00$ | 1.00 | $5.75e+02$ | $1.52e+03$ | 1.00 |
| 5 | $3.38e+00$ | $3.29e+00$ | 1.00 | $5.74e+02$ | $1.52e+03$ | 1.00 |

**Table 6.12:** The residuals in the first 5 Newton steps for equation (6.9) with initial condition $u^0 = (1, \psi(1))$.

#### Application of the tables without inversion

If the table $\psi$ is to be applied for the linear equation in the Newton iteration without any further inversion of matrix blocks, the Newton iteration reads

$$\left(\frac{\partial f_1}{\partial u_1} + \frac{\partial g_1}{\partial u_1} + \left(\frac{\partial f_1}{\partial u_2} + \frac{\partial g_1}{\partial u_2} \qquad \frac{\partial f_1}{\partial u_3} + \frac{\partial g_1}{\partial u_3}\right) \nabla\psi(u_1^n)\right) \delta u_1 = -f_1 - g_1$$

$$\begin{pmatrix} \delta u_2 \\ \delta u_3 \end{pmatrix} = \nabla\psi(u_1)\delta u_1$$

with $f$ and $g$ evaluated at the current nonlinear iterate $u^n = (u_1^n, u_2^n, u_3^n)^T$. Clearly, the residual for the fast variables is neglected in this formulation, which explains the unfortunate convergence behavior of the iteration: The residual decreases in the first steps and remains almost constant afterwards. For details, see table 6.12.

#### Application of the tables with inversion of the chemical part

Assume now that the fast part of $Df$ is easy to be inverted, hence equations of the form

$$\begin{pmatrix} \frac{\partial f_2}{\partial u_2} & \frac{\partial f_2}{\partial u_3} \\ V_2^0 & V_3^0 \end{pmatrix} \begin{pmatrix} x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_2 \\ 0 \end{pmatrix}$$

are allowed to be solved in each evaluation of the tables. This offers the possibility to compute the fast Newton update by

$$\begin{pmatrix} \delta u_2 \\ \delta u_3 \end{pmatrix} = \nabla\psi(u_1)\delta u_1 + \begin{pmatrix} \frac{\partial f_2}{\partial u_2} & \frac{\partial f_2}{\partial u_3} \\ V_2^0 & V_3^0 \end{pmatrix}^{-1} \begin{pmatrix} res_2 \\ 0 \end{pmatrix},$$

if $res_2$ denotes the fast part of the nonlinear residual.

The evolution of the residual in the solution process of equation (6.9) can be seen in table 6.13. For both values of $\alpha$, convergence is not achieved. Clearly, for $\alpha = 1000$, the iteration diverges much faster than for the smaller value $\alpha = 1$.

| | $\alpha = 1$ | | | $\alpha = 1000$ | | |
|---|---|---|---|---|---|---|
| iteration | $\|res_1\|$ | $\|res_2\|$ | $\sigma$ | $\|res_1\|$ | $\|res_2\|$ | $\sigma$ |
| 0 | $4.95e+00$ | $4.22e+00$ | 1.00 | $1.96e+03$ | $8.72e+02$ | 1.40 |
| 1 | $3.82e+00$ | $3.45e+00$ | 4.82 | $1.92e+03$ | $7.76e+02$ | 3.42 |
| 2 | $3.66e+00$ | $2.91e+00$ | 6.02 | $1.76e+03$ | $9.75e+02$ | 1.32 |
| 3 | $3.66e+00$ | $2.91e+00$ | 6.03 | $1.34e+03$ | $1.29e+03$ | 22.82 |
| 4 | $3.66e+00$ | $2.91e+00$ | 6.03 | $8.79e+02$ | $1.51e+03$ | 36.56 |
| 5 | $3.66e+00$ | $2.92e+00$ | 6.04 | $1.22e+00$ | $1.23e+03$ | 31.96 |
| 6 | $3.66e+00$ | $2.92e+00$ | 6.04 | $1.22e+00$ | $1.23e+03$ | 32.03 |

**Table 6.13:** The residuals in the first 6 Newton steps for equation (6.9) with initial condition $u^0 = (1, \psi(1))$. Here the two possibilities are considered, if the inversion of the fast part of the chemical matrix is allowed.

#### Application of the tables with inversion of the chemical and physical part

Finally, allow the computation of the solution to problems of the form

$$\begin{pmatrix} \frac{\partial f_2}{\partial u_2} + \frac{\partial g_2}{\partial u_2} & \frac{\partial f_2}{\partial u_3} + \frac{\partial g_2}{\partial u_3} \\ \frac{\partial f_3}{\partial u_2} + \frac{\partial g_3}{\partial u_2} & \frac{\partial f_3}{\partial u_3} + \frac{\partial g_3}{\partial u_3} \end{pmatrix} \begin{pmatrix} x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} b_2 \\ b_3 \end{pmatrix}$$

in every evaluation of the tables. This allows the computation of the Newton update for the slow variables by

$$\left( \frac{\partial f_1}{\partial u_1} + \frac{\partial g_1}{\partial u_1} + \left( \frac{\partial f_1}{\partial u_2} + \frac{\partial g_1}{\partial u_2} \right) \nabla \psi(u_1^n) \right) \delta u_1 = -f_1 - g_1$$

and to obtain the fast update by solving the above linear equation for $b = -f_2 - g_2$. The evolution of the residuals in the iteration are shown in detail in table 6.14. Clearly, the iteration converges for both values of $\alpha$.

### 6.3.3 A more advanced example

Reconsider the differential equation

$$-\nu u''(x) = f(u(x)), \quad x \in \Omega := ]0, 1[$$

with Dirichlet boundary conditions, where the function $f$ describes the source term of the nonlinear reaction (3.6).

#### Application of the tables without inversion

Let the table $\phi(v_1) = \nabla \psi(u_1) v_1$ reduce the linear problem

$$\nabla F(u_1^n, u_2^n) \delta u = b - F(u_1^n, u_2^n)$$

| | $\alpha = 1$ | | $\alpha = 1000$ | | $\sigma$ | |
|---|---|---|---|---|---|---|
| iteration | $\|res_1\|$ | $\|res_2\|$ | $\|res_1\|$ | $\|res_2\|$ | $\alpha = 1$ | $\alpha = 1000$ |
| 0 | $4.95e+00$ | $4.22e+00$ | $1.96e+03$ | $8.72e+02$ | 0.11 | 0.26 |
| 1 | $3.82e+00$ | $3.31e+00$ | $6.87e+01$ | $1.51e+01$ | 0.21 | 0.42 |
| 2 | $2.92e+00$ | $3.10e+00$ | $2.72e+01$ | $1.08e+00$ | 0.26 | 0.41 |
| 3 | $4.07e-01$ | $7.67e-01$ | $1.08e+01$ | $1.77e-01$ | 0.31 | 0.40 |
| 4 | $1.16e-01$ | $1.38e-01$ | $4.34e+00$ | $2.85e-02$ | 0.32 | 0.40 |
| 5 | $1.04e-02$ | $1.49e-02$ | $1.74e+00$ | $4.60e-03$ | 0.33 | 0.40 |
| 6 | $3.95e-03$ | $1.28e-04$ | $6.99e-01$ | $7.44e-04$ | 0.33 | 0.40 |
| 0 | $4.95e+00$ | $4.22e+00$ | $1.96e+03$ | $8.72e+02$ | | |
| 1 | $4.15e+00$ | $2.96e+00$ | $3.97e+00$ | $3.93e+00$ | | |
| 2 | $3.47e+00$ | $2.86e+00$ | $1.56e-05$ | $1.01e-03$ | | |
| 3 | $2.35e+00$ | $2.83e+00$ | $1.40e-13$ | $3.57e-10$ | | |
| 4 | $8.58e-01$ | $6.13e-01$ | $2.72e-16$ | $1.76e-16$ | | |
| 5 | $6.53e-02$ | $2.51e-01$ | $5.02e-17$ | $6.55e-17$ | | |
| 6 | $5.22e-04$ | $3.58e-03$ | $5.02e-17$ | $6.55e-17$ | | |

**Table 6.14:** The performance of the Newton iteration with the reduced linear solver described above. The initial condition to solve equation (6.9) is $u^0 = (1, \psi(1))$. The lower part of the table shows the evolution of the residuals with the detailed Newton iteration.

in the Newton's method. Then the reduced problem

$$\left( \frac{\partial F_1(u_1^n, u_2^n)}{\partial u_1} + \frac{\partial F_1(u_1^n, u_2^n)}{\partial u_2} \nabla \psi(u_1^n) \right) \delta u_1 = b_1 - F_1(u_1^n, u_2^n)$$

is to be solved. The Newton update for the fast variables is then obtained by $\delta u_2 = \phi(\delta u_1)$.

It is probably not surprising that this technique cannot lead to convergence of the iteration, because the update for the fast variables $\delta u_2$ is totally independent of the defect for the fast variables.

In the numerical example, the residual of the slow variables is quickly reduced to values of order $10^{-5}$ and less. This implies the smallness of the slow update and due to the linear dependence of the fast on the slow update, the smallness of the fast update, even though the fast part of the residual is big. For this reason, convergence cannot occur.

In the concrete example, the residual for the process variables is in the first three steps reduced from $\|res_1\| = 25$ to $\|res_2\| = 10^{-9}$, but the fast part of the residual $res_2$ increases from $\|res_2\| = 2.1$ to $\|res_2\| = 20.4$ and remains constant in the following iteration steps.

#### Application of the tables with inversion of the chemical part

The consideration of the fast residual for the slow Newton update by inverting the fast chemical block leaded to questionable results in the algebraic case. This is also the case

| | $\nu_i = 1$ | | $\nu_i = 0.1$ | | $\nu_i = 0.01$ | |
|---|---|---|---|---|---|---|
| step | res | $\sigma$ | res | $\sigma$ | res | $\sigma$ |
| 0 | $7.96e-1$ | 0.07 | $7.96e-1$ | 0.27 | $7.96e-1$ | 2231 |
| 1 | $7.66e-1$ | 0.10 | 1.33 | 0.47 | $1.72e+2$ | 269 |
| 2 | $7.77e-1$ | 0.10 | 1.11 | 0.93 | $1.70e+2$ | 629 |
| 3 | $7.98e-1$ | 0.10 | 1.11 | 0.88 | $1.67e+2$ | 2642 |
| 4 | $8.09e-1$ | 0.10 | 1.11 | 0.81 | $1.64e+2$ | 526 |
| 5 | $8.19e-1$ | 0.10 | 1.11 | 0.72 | $1.61e+2$ | 683 |

**Table 6.15:** The residuals and spectral radii in the Newton iteration for problem (4.24) on 33 nodes with boundaries $u(0) = u(1) = (0.9, \psi(0.9))$. As initial value constant mass fractions fulfilling the boundary conditions are chosen.

in the computation of solutions to the reaction–diffusion equation. The Newton iteration diverges in the considered examples, see table 6.15.

Even though the initial values are reasonable for the Newton iteration with detailed linear solver, they might still be too far from the solution for the iteration with reduced linear solver. For that reason, initial values are considered, which are obtained by the detailed Newton iteration after 5 and 6 steps. The performance of the iteration can be seen in the right table 6.15. Clearly, divergence occurs even with a starting point with residual of order $10^{-6}$.

#### Application of the tables with inversion of the chemical and physical part

Apply now the given table $\psi$ only for the calculation of the slow Newton update. The fast update will be calculated in detail by inverting the corresponding block of the physical and chemical matrix. The number of required Newton steps can be found in table 6.16. As in the case without conservation laws (cp. table 6.7), the performance of the Newton iteration seems to be (almost) independent of the grid size and depends in the detailed formulation only slightly on the viscosities. But the reduced Newton iteration performs much worse with smaller values of $\nu$.

In order to understand this phenomenon, consider the solution process of the equation

$$-10^{-3}\, u''(x) = f(u(x)), \quad x \in \Omega := ]0,1[$$
$$u(0) = u(1) = \begin{pmatrix} 0.9 \\ \psi(0.9) \end{pmatrix}$$

on a grid with 65 nodes. The solution to this reaction–diffusion equation can be seen in figure 6.1, also the difference between the detailed and reduced nonlinear problem is shown. With errors of order $10^{-4}$ for the fast variables and even $10^{-5}$ for the process variables, the reduced solution leads to a decent approximation of the total problem.

Consider now the reduction of the linear problem for the Newton update. Then the nonlinear residual decreases within the first six Newton steps from 0.57 to $5.9 \cdot 10^{-5}$, see table 6.17. Then, another 36 iterations are necessary in order to decrease the residual to $10^{-5}$, the

| Boundaries: $u_1 = 0.9$, $(u_2, u_3) = \psi(0.9)$ | | | |
|---|---|---|---|
| viscosities | 33 nodes | 65 nodes | 129 nodes |
| $\nu_i = 1$ | 5 (4) | 5 (5) | 5 (5) |
| $\nu_i = 10^{-2}$ | 16 (7) | 16 (7) | 16 (7) |
| $\nu_i = 10^{-3}$ | 93 (7) | 93 (7) | 93 (7) |
| $\nu_i = 10^{-4}$ | 527 (7) | 526 (7) | 528 (7) |
| Boundaries: $u_1 = 0.5$, $u_2 = 0.3$, $u_3 = 0.2$ | | | |
| viscosities | 33 nodes | 65 nodes | 129 nodes |
| $\nu_i = 1$ | 4 (4) | 4 (4) | 4 (4) |
| $\nu_i = 10^{-2}$ | 14 (4) | 14 (4) | 14 (4) |
| $\nu_i = 10^{-3}$ | 79 (4) | 79 (4) | 79 (4) |
| $\nu_i = 10^{-4}$ | 298 (4) | 302 (4) | 299 (4) |

**Table 6.16:** The table shows the number of Newton steps required for the reaction–diffusion equation (4.24) for different viscosities on different grids. Here, the linear problem is reduced with $\nabla\psi$ and the inversion of the chemical and physical matrix. The number in brackets is the number of required detailed Newton steps.

| Newton step | residual, detailed linear | residual, reduced linear | $\sigma(\mathbb{1} - P(A+B))$ |
|---|---|---|---|
| 0 | 0.567596 | 0.567596 | 0.6475 |
| 1 | 0.542231 | 0.538301 | 0.8450 |
| 2 | 0.338205 | 0.208876 | 0.9053 |
| 3 | 0.255162 | 0.201844 | 0.9370 |
| 4 | 0.031354 | 0.024962 | 0.9433 |
| 5 | 0.001073 | 0.001175 | 0.9457 |
| 6 | $1.41e-06$ | $5.91e-05$ | 0.9458 |
| 7 | $3.20e-11$ | $5.52e-05$ | — |

**Table 6.17:** The residuals of the first seven Newton steps for the iteration with reduced and detailed linear problem. The third column shows the spectral radius of the matrix $\mathbb{1} - P(A+B)$, where $P$ denotes the preconditioner created by $\nabla\psi$ and the inversion of the fast parts of the chemical ($A$) and physical ($B$) matrix.

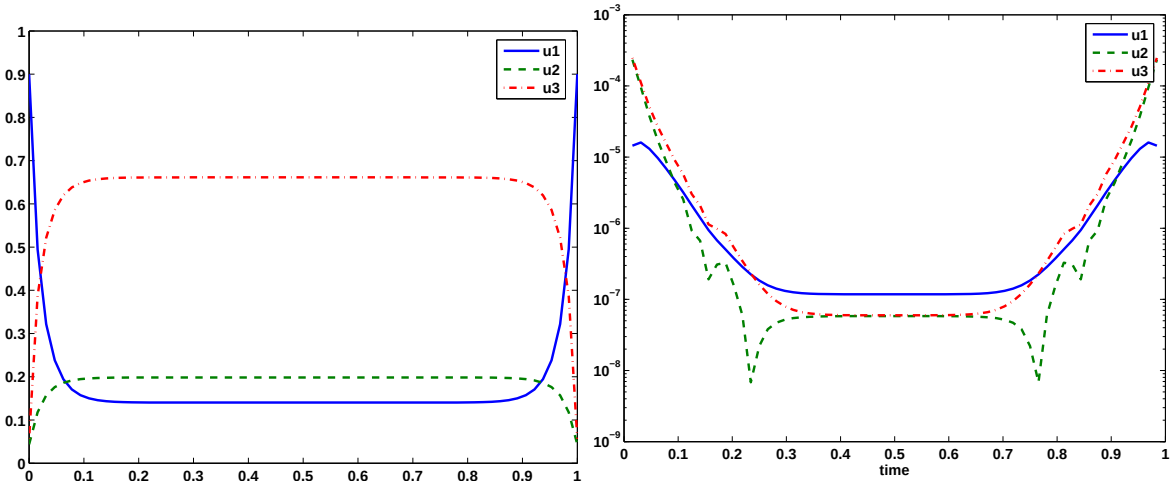


**Figure 6.1:** The solution to the reaction–diffusion equation with $\nu = 10^{-3}$ and the produced error, if the total problem is reduced with the QSSA–table $\psi_Q$.

residual of order $10^{-6}$ is reached after 81 iterations. This observation suggests clearly that the residual decreases fastly at the beginning and convergence problems arise at the end of the iteration. This behavior is substantiated by the eigenvalue analysis of the system matrix for the Richardson iteration. Clearly, the spectral radius tends to one, which leads to the poor convergence properties.

The quite decent performance in the first steps is is also supported by figure 6.2, which

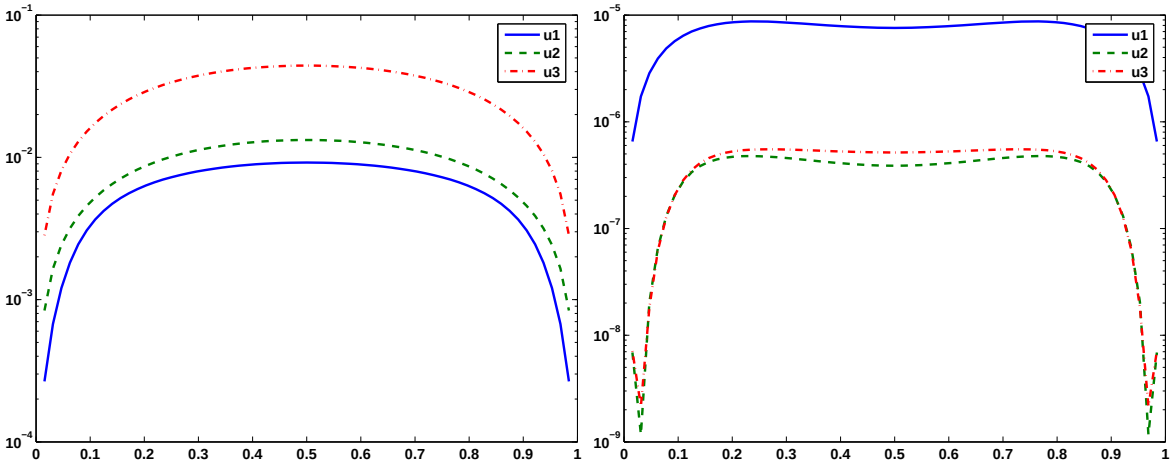


**Figure 6.2:** The difference between the detailed solution and the "solution", which is obtained by the reduced Newton iteration after 5 steps. The right figure shows the residual after 5 iterations.

shows the difference between the detailed solution and the iterate, which is obtained by the reduced Newton iteration after 5 steps. Also the residual after 5 iterations is plotted. Both figures suggest clearly that the 5–th iterate is already close to the detailed solution. The

differences are biggest in the center of the domain, where the diffusion term plays a minor role.

# 7 Adaptive Model Reduction

The previous chapter showed clearly that preconditioning with QSSA and ILDM cannot be performed for stationary problems. This chapter will therefore concentrate on instationary homogeneous processes, where the introduced reduction methods will be applied in order to reduce the computational costs. Clearly, the application of these methods introduces systematic errors in the time steps, which cannot be avoided. The goal of this chapter is therefore to find criteria, which indicate, if the next time step can be performed with ILDM or has to be calculated in detail in order to obtain errors with a certain tolerance.

The first section will concentrate on the information, which can be obtained by considering only the residual of the ODE. It will turn out that this information is not sufficient in order to find criteria, whether the forthcoming time step is to be calculated in detail or not. The section thereafter combines then the residual with the so called dual solution of the problem, a strategy, which proves to be expensive, but also very successful, if computation time plays a minor role. It can be expected that the method performs nicely, if the original problem is highly nonlinear, such that each dual time step is rather cheap compared to the original time step. The third section in this chapter will finally investigate the possibilities of applying the obtained dual solution also for estimating the discretization error. It will turn out to work nicely.

## 7.1 Solving strategy for ODEs

In this section, an ODE involving chemical combustion shall be solved. The goal is to do as many time steps with ILDM as possible and use a detailed solver only, if the produced error is too big. This procedure reduces the numerical expenses dramatically, if the reduction works nicely. And in the worst case, where all reduced time steps have to be redone with a detailed solver, only little more effort has to be spent than the pure detailed solver requires, see table 7.1. Here, the computation time for an ODE with $n$ variables can be seen. Two different linear reaction mechanisms are considered, where only one variable is considered to be a process variable in the first reaction. Two process variables exist for the second source term. The reaction formalism is similar as in (4.2) with additional species and adjusted reaction rates.

In general, there are three sources for big errors for the application of ILDM:

- The previous iterate is not on the manifold. This phenomenon occurs frequently at the beginning of the solution process, if the initial value is not on the manifold.
- External sources are not consistent with the reduction mechanisms.

| | 1 PV | | 2 PV | |
|---|---|---|---|---|
| $n$ | detailed | ILDM | detailed | ILDM |
| 20 | 3.0 | 1.5 | 2.8 | 2.2 |
| 40 | 21 | 3.8 | 20 | 5.7 |
| 60 | 67 | 7.3 | 66 | 9.6 |
| 80 | 167 | 12.5 | 154 | 14.9 |
| 100 | 309 | 18.7 | 313 | 24.7 |

**Table 7.1:** The computation time in seconds for an ODE with $n$ variables, where 1 and 2 variables are process variables respectively. Clearly, the numerical costs grow with $n^3$ for the detailed solver, whereas they grow with less than $n^2$, if ILDM is applied.

- The reduction mechanisms are not valid, for example due to a change in the spectrum of $\nabla f$.

Because of these phenomena, ODEs exist, where the solution process for one time step has to be switched from an ILDM–solver to a detailed solver and vice versa. Therefore criteria have to be found, which indicate that a change in the solution method is necessary.

In order to investigate the behavior of the detailed and reduced solution and to find possible criteria, the ODE

$$\begin{aligned} \dot{u} &= f(u) + b \\ u(0) &= u_0 \end{aligned}$$

shall be considered, where $f$ denotes the source term, for which the table $\psi$ was created. The vector $b$ shall be a (possibly time–dependent) external source term, which is not contained in $\psi$.

### 7.1.1 Switching from detailed to reduced solver

Let now the source term $f$ be given, such that the spectral gap remains almost constant and the set of physically fast variables does not change in time. The external source $b$ is set to zero, but the initial value $u_0$ is chosen not to be on the manifold given by either QSSA or ILDM.

The detailed solution of the ODE does therefore relax fastly onto the manifold and remains on it, until it eventually reaches the stationary point. This means mathematically

$$u_2(t) \approx \psi(u_1(t)) \quad \forall t \geq t_0,$$

which leads to a criterion, how the acceptance of an ILDM–solution can be detected, if the current time steps are solved in detail: Switch to ILDM, if

$$|u_2(t) - \psi(u_1(t))| < tol \cdot |u_2(t)|.$$

In order to see the relation between the solution and the manifold, consider the distance of the detailedly obtained trajectory to the manifold for the ODEs

$$\dot{u} = f(u) \qquad \dot{u} = g(u) \tag{7.1}$$
$$u(0) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \qquad u(0) = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

for the source term $f$ of the reaction (4.21). The function $g$ shall be defined by

$$g(u) = \begin{pmatrix} -u_1 \\ 1000(1/u_1 - u_2) \end{pmatrix}$$

Clearly, both initial values are not on the manifold. The value $|u_2(t) - \psi(u_1(t))|/|u_2(t)|$ can be seen in figure 7.1. The figure depicts clearly that the manifold created by ILDM

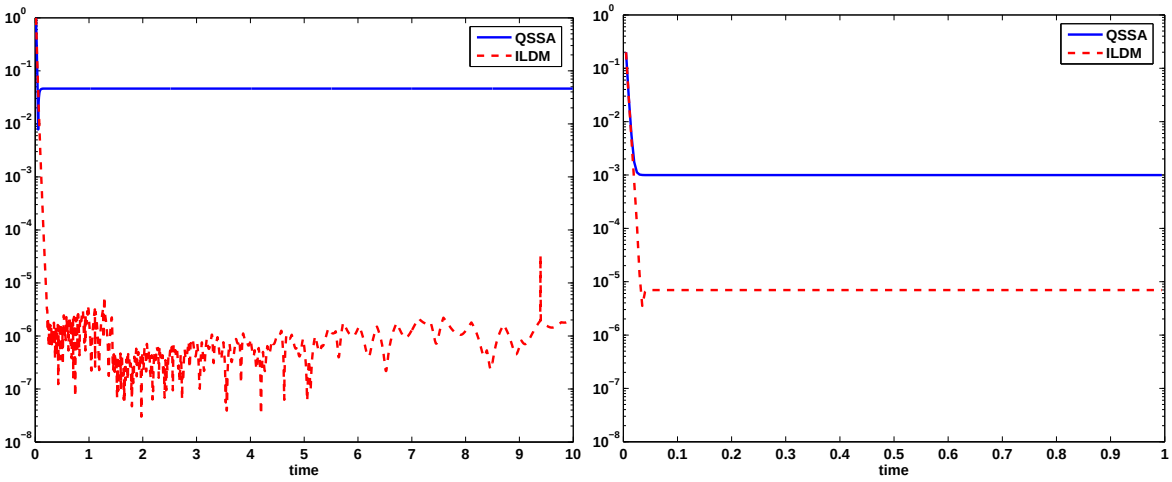


**Figure 7.1:** The relative difference of the fast part of the solution to the manifolds given by QSSA and ILDM. The left figure was created for the ODE $\dot{u} = f(u)$, the right figure for $\dot{u} = g(u)$ as in (7.1).

attracts the detailed solution much better than the manifold by QSSA. The tolerance *tol* has therefore to be much smaller for the treatment of ILDM, the figures suggest the values

$$tol_{QSSA} = 10^{-1} \qquad tol_{ILDM} = 10^{-5}$$

for the problems defined in (7.1). The errors produced by the switched strategy can be seen in figure 7.2 for the chemical ODE. For the ODE $\dot{u} = g(u)$, the error for the process variables is zero, because the source term for the slow variables is independent of the fast variables. The difference in the fast variables can be seen in figure 7.3.

Both figures (7.2 and 7.3) depict clearly that the introduced criteria can be applied to determine, whether the following time step shall still be calculated detailedly, or if the reduction method can be applied. The above suggested values for *tol* are suitable for the considered examples, but do in general depend on the given problem and its spectral gap.

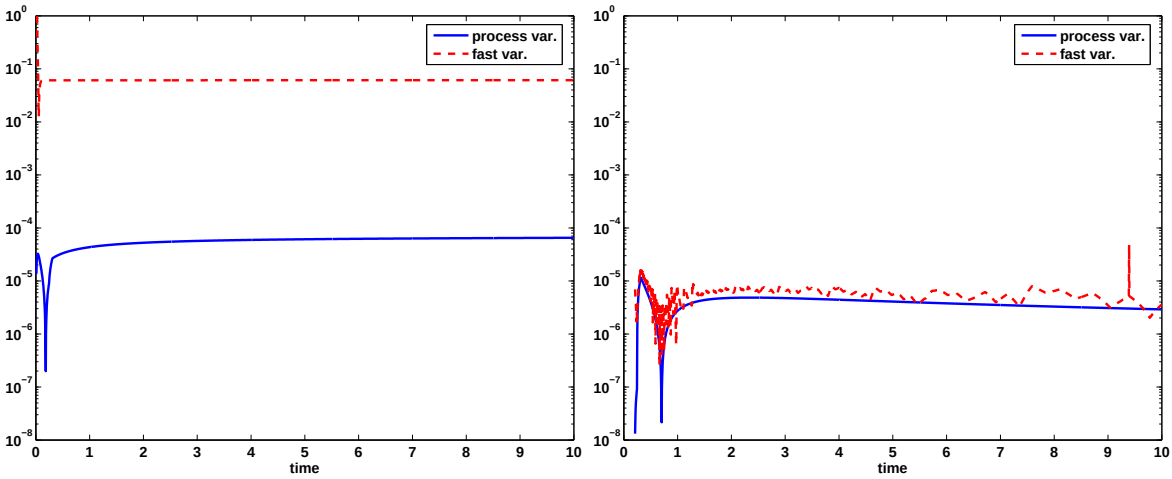


**Figure 7.2:** The relative difference between the switched and the detailed solution for the ODE $\dot{u} = f(u)$ as in (7.1). The left figure shows the reduced time steps with QSSA, the right figure with ILDM. In both cases, 2000 time steps were performed, the number of detailed time steps is 40 for ILDM and only 1 for QSSA.

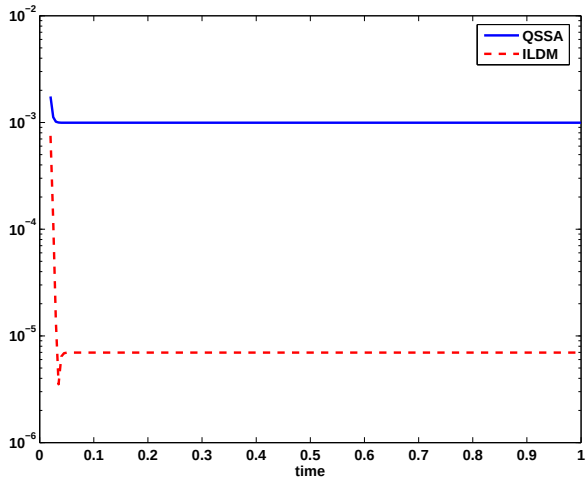


**Figure 7.3:** The relative difference between the fast variables in the detailed and switched solution process for the ODE $\dot{u} = g(u)$. For both reduction methods, 4 of the 2000 time steps were performed in detail, before the solution process switched to the reduced solver.

### 7.1.2 Switching from reduced to detailed solver

The more difficult task in the switched solution strategy is to find criteria to determine, whether the time step, which was calculated with ILDM, is still accurate enough. Otherwise this time step has to be repeated with a detailed solver. This means basically that the table is not valid anymore.

Clearly, the above criteria and therefore the distance between the current time step and the manifold cannot be used, because the current state is by definition on the manifold, if the last time step was calculated with the reduction methods.

In this section, three different possibilities for this criteria are investigated:

- Comparison of detailed and reduced solution.
- Validation of the table.
- Calculating the detailed residual of the reduced time step.

#### Compare the detailed with the reduced solution

The difference between the detailed and the reduced solution does clearly contain valuable information about the applicability of the reduction methods, but has the big disadvantage that the detailed solution is unknown. And even if the detailedly calculated state vector is known, then there is no need to proceed with the less accurate reduced state vector. So this paragraph is only of theoretical interest in order to obtain a feeling for the behavior of the solution, if the manifold is valid for the problem or not.

For that reason, compare the two ODEs

$$\begin{aligned} \dot{u_1} &= -u_1 & \dot{v_1} &= -v_1 \\ \dot{u_2} &= 1000(1/u_1 - u_2) & \dot{v_2} &= 1000(1 - v_1 v_2) \end{aligned} \tag{7.2}$$

Both systems have the same manifold $\psi(u_1) = 1/u_1$, if the QSSA–method is considered. For the left problem, this manifold is always valid, whereas it is only valid in certain regions for the problem on the right hand side. This can easily be seen by considering the gradients of the right hand side, which read

$$\nabla f(u) = \begin{pmatrix} -1 & 0 \\ -1000/u_1^2 & -1000 \end{pmatrix} \quad \text{and} \quad \nabla f(v) = \begin{pmatrix} -1 & 0 \\ -1000v_2 & -1000v_1 \end{pmatrix}.$$

The eigenvalues for the gradient of the left problem are clearly constant in $u$ and therefore also in time ($\lambda_1 = -1$, $\lambda_2 = -1000$), therefore the spectral gap does not change. The eigenvalues for the right hand side depend on the current state ($\lambda_1 = -1$, $\lambda_2 = -1000v_1$) and the spectral gap decreases linearly with decreasing $v_1$. This means that the manifold is only valid for the right problem, if $v_1$ is big enough.

This can be seen by performing a few time steps from a state on the manifold with reasonably small $u_1$. Let for example $u_1 = 10^{-5}$, therefore $u_2 = 10^5$. The next five time steps with

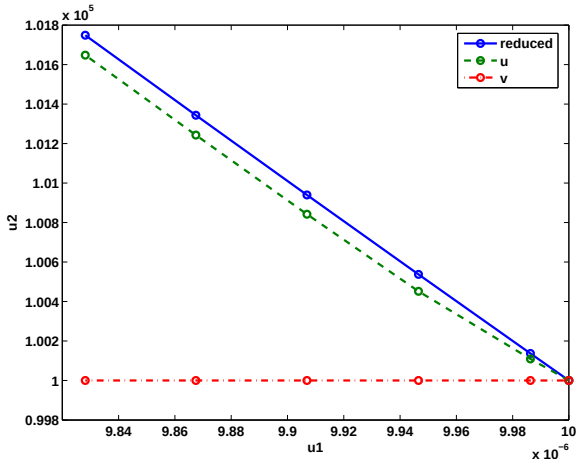


**Figure 7.4:** The phase diagram of the reduced solution for both problems described by (7.2), together with the time steps obtained from the reduced solver.

step size $k = 10^{-3}$ can be seen in figure 7.4. Clearly, the detailed time steps for the ODE on the left hand side with valid manifold removes only slightly from the reduced solution (and therefore from the manifold) and shows similar behavior as the reduced solution in the following. This is typical and shows the difference between the exactly calculated attracting manifold and its approximation calculated with QSSA. The problem, for which the manifold is not valid anymore, shows totally different behavior. Already in the first step, the current state removes strongly from the manifold and keeps removing in the following time steps. This is clearly a (very expensive) indicator that the reduced solution does not lead to accurate solutions anymore.

### Validate the table

As already seen before, the provided tables can only be valid, if a spectral gap exists and is reasonably big. The information about the spectral gap can therefore be used as an indicator for the applicability of the reduction methods. This leads to the idea to estimate the spectral gap in each time step and decide, whether the following time step is to be done in detail.

This procedure has two major drawbacks:

- The computational costs are very high. In order to calculate the spectral gap, information about the eigenvalues is required, which can be obtained by two different approaches. The first possibility is the execution of the QR–algorithm [49]. Due to the convergence properties, the eigenvalues right beside the spectral gap are fastly obtained, such that the algorithm does not have to be executed until convergence occurs for all eigenvalues.

  The second possibility is the inverse iteration [21, Ch. 7]. Assuming that the fastest slow eigenvalue and the slowest fast eigenvalue do not change much in one time step,

the corresponding eigenvalues of the previous time step can be used as an initial guess for the new inverse iteration. This will in general lead to quick convergence.

But still, even little QR–steps require the QR–factorization in each iteration step and the inverse iteration the solution of a linear system of the size of the detailed problem. This leads to similar numerical costs as one or even more Newton steps for the detailed problem, so the whole procedure will be almost as expensive as the detailed calculation of the upcoming time step.

- This criterion is only valid, if ODEs without disturbance are considered. An external source might easily lead to big differences between the detailed and reduced solution, even though the analysis of the spectral gap is not affected at all. This means that the switched solution process will continue to solve the problem with the reduction methods and lead to big errors.

Because of the second problem, this criteria can only be applied, if problems of the form $\dot{u} = f(u)$ with the table created for $f$ are considered. If this is the case, switch from the reduced solution to a detailed solver, if

$$\frac{\lambda_{n_1+1} - \lambda_{n_1}}{\lambda_{n_1}} < SG$$

for a given constant $SG \in \mathbb{R}$.

Consider again the ODE $\dot{u} = g(u)$ as defined on the right hand side of (7.2) with initial condition $u(0) = (1, 0)^T$. The evolution of the spectral gap is shown in figure 7.5 as well as the difference between the detailed and the switched solution for different values of $SG$. The

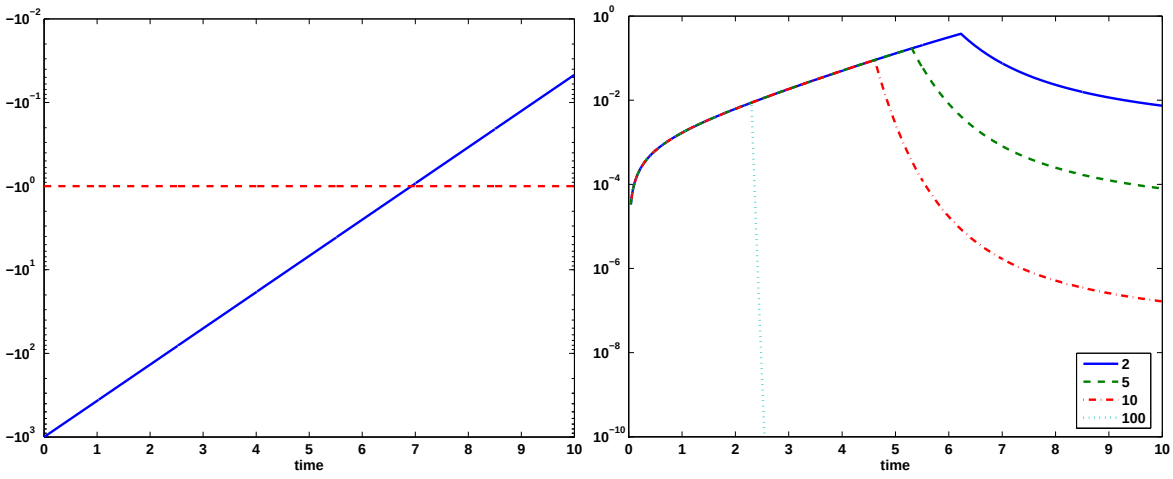


**Figure 7.5:** The left figure shows the evolution of the eigenvalues in time. The considered problem is the second problem of (7.2). The right figure shows the relative difference between the detailed and the switched solution for the fast variable, if $SG$ is chosen to be in $\{2, 5, 10, 100\}$.

error is of course bigger, if the switching occurs at later stage, but then the computational costs are also lower. From the 2000 performed time steps are 1545 steps calculated in detail

for $SG = 100$, whereas only 760 detailed steps are necessary for $SG = 2$. An interesting property of this example is that the error decreases after the switching.

Also chemical reaction systems can have a changing spectral gap. Consider for example the reactions

$$\begin{aligned} sp_1 &\underset{k_2}{\overset{k_1}{\rightleftharpoons}} sp_2 \\ sp_1 + sp_2 &\xrightarrow{k_3} sp_1 + sp_3 \\ sp_2 &\xleftarrow{k_4} sp_3 \\ sp_3 &\xrightarrow{k_5} sp_4, \end{aligned}$$

which differs from reaction (4.2) only in the second reaction, where now $sp_1$ forms a third body. One can also think of the reaction mechanism

$$sp_1 \underset{k_2}{\overset{k_1}{\rightleftharpoons}} sp_2 \underset{k_4}{\overset{k_3 \cdot u_1}{\rightleftharpoons}} sp_3 \xrightarrow{k_5} sp_4, \tag{7.3}$$

where the reaction rate for the reaction $sp_2 \longrightarrow sp_3$ depends on the mass fraction of species $sp_1$. These two formulations are equivalent. If only $sp_1$, $sp_2$ and $sp_3$ are calculated, this is a nonlinear reaction system without conservation laws. It has qualitatively the same properties as the above considered example.

### The detailed residual as switching indicator

If one considers the residual of an ODE to a solution $u$, the term

$$\rho(u) = f(u) - \dot{u}$$

is treated. From the numerical point of view, this is of course impossible, because the exact derivative of $u$ is in general unknown. One possibility to approximate the residual is given by

$$\rho(u^{n+1}) = f(u^{n+1}) - 1/k(u^{n+1} - u^n)$$

with $u^n = u(t_n)$ and $u^{n+1} = u(t_n + k)$. This is basically the algebraic residual of an implicit Euler step.

An analysis of the examples (7.1) and (7.2) will show that the switching criteria cannot be based directly on the residual. In order to see, why this proposition is valid, consider first the residual of the fast variables if QSSA is used as reduction method. Then the residual of the fast variables reads

$$\rho_2(u_1, \psi(u_1)) = f_2(u_1, \psi(u_1)) - \dot{u_2} = -\dot{u_2},$$

because $f_2(u_1, \psi(u_1)) = 0$ by definition.

Now four problems are considered:

- The ODE defined by the source term of reaction (4.21). The reduction methods perform nicely, the relative difference between the reduced and detailed solution is less than 0.1%.
- The second ODE is defined by the reaction (7.3). Here, the application of the reduced solver leads to errors of size 100%.
- The problem defined by
$$\dot{u} = \begin{pmatrix} -u_1 \\ 1000(1/u_1 - u_2) \end{pmatrix}$$
can perfectly be reduced, the produced error is again less than 0.1%.
- The last problem is obtained by the above defined ODE with the second equation changed to
$$\dot{u_2} = 1000(1 - u_1 u_2).$$
The QSSA–table of the last two problems is clearly the same, but here, the reduction produces errors of more than 100%.

For all four problems, the detailed residuals of the reduced solution can be seen in figure 7.6. Clearly, the residuals for the first and second problem show similar properties. They are of

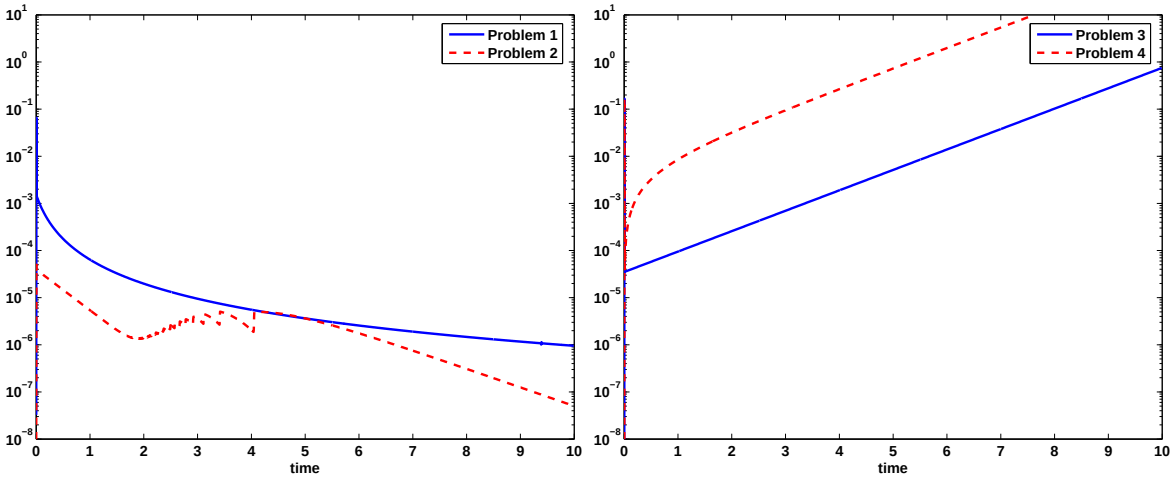


**Figure 7.6:** The relative residuals for the four previously defined ODEs. For problems 1 and 3, the reduction methods produce only small errors, whereas the reduced solvers lead to big errors for problems 2 and 4.

the same size and are generally decreasing. Also the reductions of problem three and four lead to similar behavior of the big and increasing residuals. But the reduction methods show only nice behavior for the first and third problem, which indicates, that the fast residual on its own cannot be used to distinguish good and bad performance of the reduction methods.

In order to see that the slow residual can neither be used as an indicator for the performance of the reduction, reconsider the third and the fourth problem. Since the source term for the process variable is independent of the fast variables, the reduced solver leads to the same

solution for the slow variables as the detailed solver. This means that the slow part of the residual is zero (up to rounding errors). Again, one of these problems can be successfully reduced, whereas the reduction mechanisms produce a big error for the other ODE.

## 7.2 A posteriori error control with dual solutions

As shown in the last section, information about the residual is not sufficient in order to obtain criteria, when the reduced solution is to be recalculated with a detailed solver. One possibility to obtain a relation between the residual and the produced error, is the consideration of dual problems. They provide information, how the residual effects the error of the solution. This strategy was invented for Galerkin methods with adaptive step sizes in order to find an optimal discretization. For details, see for example [4] and [9]. In this section, this method will be explored for the application of the reduction methods.

### 7.2.1 A basic introduction into dual problems

Assume that the solution to the ODE

$$\begin{aligned} \dot{u} &= f(u) \\ u(0) &= u_0 \end{aligned}$$

is to be calculated on the interval $I = [0, T]$, which is equivalent to the weak formulation

$$\begin{aligned} \int_I (u' - f(u), \varphi) dt = 0 \quad \forall \varphi \in C(I) \\ u(0) = u_0, \end{aligned}$$

if $u$ is in $C^1(I)$. A second equivalent formulation is

$$\begin{aligned} \sum_{n=1}^{N} \left( \int_{I_n} (u' - f(u), \varphi) dt + ([u]_{n-1}, \lim_{t \downarrow t_n} \varphi_{n-1}) \right) = 0 \quad \forall \varphi \in V(I) \\ u(0) = u_0, \end{aligned} \tag{7.4}$$

where $V(I) = \{v : I \to \mathbb{R}^d | v_{|I_n} \in C_c^1(I_n)\}$ and $[u]_n = \lim_{t \downarrow t_n} u - \lim_{t \uparrow t_n} u$. Taking the finite dimensional space $S^{(r)}(I) = \{v \in V(I) | v_{|I_n} \in P_r(I_n)\}$ of all piecewise polynomials of degree $r$ or less instead of $V(I)$ in the above formulation, leads to discontinuous Galerkin methods. If $r = 1$, this method is even equivalent to the implicit Euler algorithm.

Solving (7.4) with $S^{(r)}(I)$ instead of $V(I)$ leads to an approximation $u_h$ of the exact solution $u$. The residual

$$\rho(u_h)(\varphi) = \sum_{n=1}^{N} \left( \int_{I_n} (f(u_h) - u_h', \varphi) dt - ([u_h]_{n-1}, \lim_{t \downarrow t_n} \varphi_{n-1}) \right)$$

is therefore unequal to zero.

In order to obtain a relation between the residual and the actual difference between $u$ and $u_h$, dual problems are considered. Assume therefore that the error $J(e(T)) = (j, e(T))$ is of interest. Think for example of $j = (\delta_{i,k})_i$, then $J(e(T))$ describes the error $e_k(T) = x_k(T) - u_k(T)$ of the $k$–th variable at time $T$. Let further the matrix $B$ be defined by

$$Be := \int_0^1 f'(u + se)e \, ds$$

with $e := u - u_h$ and let $z$ be the solution of the dual problem

$$\begin{aligned} \dot{z} + B^* z &= 0 \\ z(T) &= j. \end{aligned}$$

Here, $z$ is called the dual solution to the original problem $\dot{u} = f(u)$. Then the error identity

$$(j, e(T)) = \rho(u_h)(z)$$

holds, and describes the effect of the residual to the error. The values of the $\rho(u_h)(z)$ on the intervals $I_n$ can therefore be used as an indicator, where the detailed calculation has to be performed, such that the mixed solution process has an optimal effect to the error. Clearly, the error identity can only be valid, if $z$ was calculated exactly. The value of $\rho(u_h)(z_h)$ equals zero, if $z_h$ is an approximation of $z$ with $z_h$ in $S^{(r)}(I)$.

Note that not only errors at the final time $T$ can be calculated. Change the dual problem to

$$\begin{aligned} \dot{z} + B^* z &= j \\ z(T) &= 0, \end{aligned}$$

if the error $e(T) = \int_0^T (j, x(t) - u(t)) \, dt$ is to be calculated.

In order to estimate the error produced by the reduction methods, the discretization error can be neglected, because only the difference between the discretized solution and the discretized reduced solution is of interest. For a given reduced solution $u_{h,red}$, the dual solution $z_h$ is to be calculated and the term $\rho_h(u_{h,red})(z_h)$ estimates the systematic error produced by QSSA or ILDM.

This method has two bottle–necks:

- For the creation of the dual problem, the exact solution is required, which is for obvious reasons impossible. The matrix $B$ has therefore to be approximated by only the approximating solution $u_{red}$ and might be calculated to be $B = f'(u_{red})$.
- The calculation of the dual solution $z$ is very expensive, even though the problem is only linear. The linearity means that the solution process is as expensive as the detailed solution process of the original problem $\dot{x} = f(x)$, if it is assumed that only one Newton step is required in each time step. The adaptive strategy can therefore only be time saving, if
  - the dual problem can be reduced, or

– the original problem is highly nonlinear, such that many Newton steps are required in each time step.

The algorithm to calculate the solution to an ODE with a switched strategy is therefore:

1. calculate the reduced solution
2. estimate the error by solving the dual solution, which was created with the just obtained reduced solution
3. calculate the switched solution
4. estimate the error by the renewed dual solution. The dual problem changes, because the switched solution leads to a different approximation of $B$.
5. repeat steps 3. and 4. until the desired tolerance is reached.

### 7.2.2 The reduced creation of the dual problem

The effect of the non–existing exact solution for the creation of the matrix $B$ and therefore of the dual problem shall here be investigated with two examples. The first ODE is given by

$$\dot{u} = f(u) + b(t) \tag{7.5}$$

with initial conditions on the manifold and the source term $f$ of reaction (4.21). The time interval of interest is $[0, 3]$ and the disturbance $b$ acts on the fast variables only for $1 < t < 2$. With $b = 0$, the ODE is nicely approximated by the application of the tables, the existence of $b$ ensures therefore that the tables are not valid in the time region, where $b(t) \neq 0$. The reduced solver leads to an approximation, which differs from the detailed solution at time $t = 3$ by $2 \cdot 10^{-2}$ for the slow and $10^{-3}$ for the fast variables. The averaged error is $6 \cdot 10^{-2}$ and $2 \cdot 10^{-1}$ respectively.

The second problem is the ODE

$$\dot{u} = \begin{pmatrix} -u_1 \\ 1000(u_1^3 - u_1^4 u_2) \end{pmatrix} \tag{7.6}$$

for $t \in [0, 1]$. The initial value is chosen to be on the manifold. For this problem, the spectral gap changes in time. The application of the tables leads therefore to decent results in the first time steps, but its accuracy decreases highly for $t \to 1$. The errors produced by the application of the reduction methods are $10^{-1}$ and $9 \cdot 10^2$ for the slow and fast variables at time $t = 1$. The averaged error is $2 \cdot 10^{-2}$ and $5 \cdot 10^1$ respectively. The solution to these problems can be seen in figure 7.7. The right figure shows clearly, that the produced errors are also relatively very big.

In tables 7.2 and 7.3, the effects of the approximated creation of the dual problem can be seen. Table 7.2 suggests clearly that the creation of the dual problem with the approximated solution leads to similar results as the exact dual solution. It also shows that there is basically no difference, if the fast or the slow variables are controlled. The situation for the second problem (7.6) is a bit different. Here, the controlling of the fast variables leads of course to

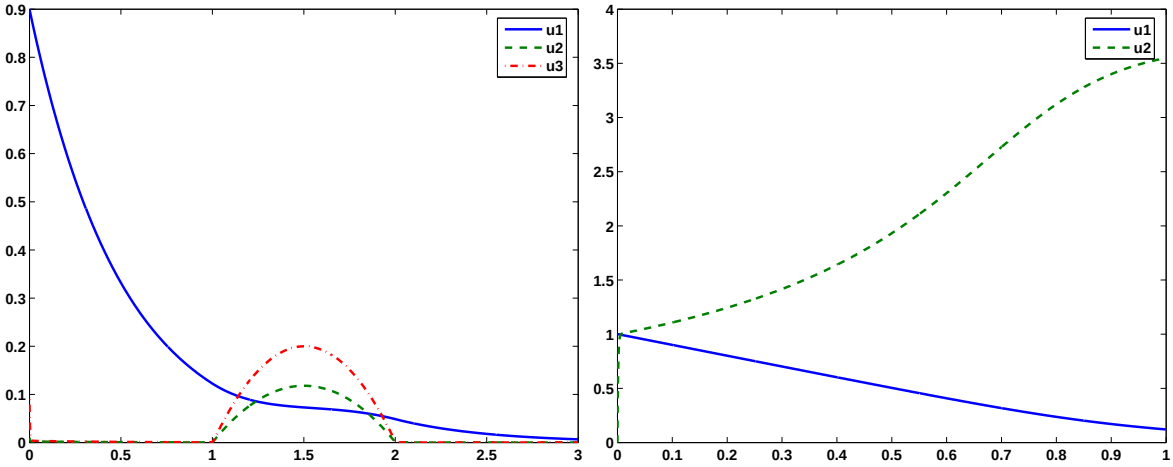


**Figure 7.7:** The solutions to the problems (7.5) (left) and (7.6) (right). The values of the solution at the final time is $u(3) = (6 \cdot 10^{-3}, 2 \cdot 10^{-5}, 3 \cdot 10^{-5})$ in the left figure.

| | | Final time | | Average | |
|---|---|---|---|---|---|
| control | measure | dual exact | dual approx. | dual exact | dual approx. |
| slow variables | # detailed | 469 | 472 | 484 | 481 |
| | error slow | $2e-3$ | $2e-3$ | $5e-3$ | $6e-3$ |
| | error fast | $1e-4$ | $1e-4$ | $3e-2$ | $3e-2$ |
| fast variables | # detailed | 469 | 481 | 538 | 482 |
| | error slow | $2e-3$ | $2e-3$ | $3e-3$ | $6e-3$ |
| | error fast | $1e-4$ | $1e-4$ | $2e-2$ | $3e-2$ |

**Table 7.2:** The behavior of problem (7.5), if the dual problem is either exactly or approximatively created. The table is splitted into controlling the error of the slow (fast) variables at time $t = 3$ and controlling the error averaged over the time interval $[0, 1]$. The total number of iteration steps is 2000.

an improvement of their accuracy, but the accuracy of the slow variables is not improved, see table 7.3. If the controlled variables are the process variables, the error for the fast variables is also dramatically reduced, but the solution process is much too expensive. Fortunately, this observation holds for both the exactly and approximatively created dual problem.

It can therefore be concluded that the approximative creation of the dual problem leads to similar performances as the creation of $B$ with the exact solution of the original problem.

### 7.2.3 The reduction of the dual problem

A major drawback of the idea of solving dual problems is the numerical expenses for the calculation of the backward solution $z$. Since the forward solution $u$ was calculated with a reduced solver, the dual solver might easily be even more expensive, if the dual problem

| | | Final time | | Average | |
|---|---|---|---|---|---|
| control | measure | dual exact | dual approx. | dual exact | dual approx. |
| slow variables | # detailed | 1027 | 1652 | 1445 | 1425 |
| | error slow | $1e-3$ | $1e-4$ | $7e-4$ | $8e-4$ |
| | error fast | $3e-3$ | $3e-4$ | $5e-3$ | $1e-1$ |
| fast variables | # detailed | 20 | 20 | 391 | 420 |
| | error slow | $1e-1$ | $1e-1$ | $1e-2$ | $1e-2$ |
| | error fast | 87 | 87 | $5e-1$ | $5e-1$ |

**Table 7.3:** The behavior of problem (7.6), if the dual problem is either exactly or approximatively created. Note that the nice behavior at the control of the slow variables is very expensive, because the number of detailed steps is high. In addition, three to four adaptive cycles had to be performed, such that the detailed solution process is in fact much cheaper.

is solved in detail. This leads directly to the question, whether the dual problem can be reduced as well.

Clearly, if the dual problem is reduced, it has to be reduced with the same table as the forward problem, because the creation of a new table just for the dual problem is for obvious reasons much too expensive. The application of the derivative of the function $\psi$ created for the solution process of $u$ is therefore to be investigated for the applicability for the dual problem.

Consider therefore the two dual problems

$$\begin{cases} \dot{z} + B^* z = 0 \\ z(T) = j \end{cases} \qquad \text{and} \qquad \begin{cases} \dot{z} + B^* z = j \\ z(T) = 0 \end{cases}$$

related to the control of the error at final time $t = T$ and the averaged error. The vector $j$ equals, given that the error of the $k$–th variable is to be controlled, the $k$–th unit vector.

#### Error control of fast variables

Assume now that one of the fast variables is to be controlled. The reduced dual problems read then

$$\begin{cases} \dot{z}_1 + ((B^*)_{11} + (B^*)_{12}\nabla\psi(u_1))z_1 = 0 \\ z_1(T) = 0 \end{cases}$$

for both detailed dual problems, because $j_1 = 0$. Clearly, the equation is linear in $z_1$, the unique solution for both problems is therefore $z_1(t) = 0$ due to the right hand sides and initial conditions. The total solution is

$$z(t) = 0,$$

because the reducing function $\nabla\psi$ provides a linear relation between $z_2(t)$ and $z_1(t)$. The residual $\rho$ tested with $z$ is clearly also zero and does therefore not contain any information about the relation between the residual and the actually produced error. It can be concluded

that the dual problem has to be solved with a detailed solver, if a fast variable is to be controlled.

### Error control of process variables

The situation is a bit more optimistic, if one of the slow variables is controlled. The dual problem for the control of the error at time $t = T$ reads then

$$\begin{cases} \dot{z}_1 + ((B^*)_{11} + (B^*)_{12}\nabla\psi(u_1))z_1 = 0 \\ z_1(T) = j_1, \end{cases}$$

which has a solution different from zero. The question is now, under which circumstances the table $\psi$ for the problem $\dot{u} = f(u)$ can be applied as reduction method for its dual problem.

The question will be stated more precisely by considering example (7.6). The detailedly calculated dual solution for the second problem is shown in figure 7.8. It clearly differs

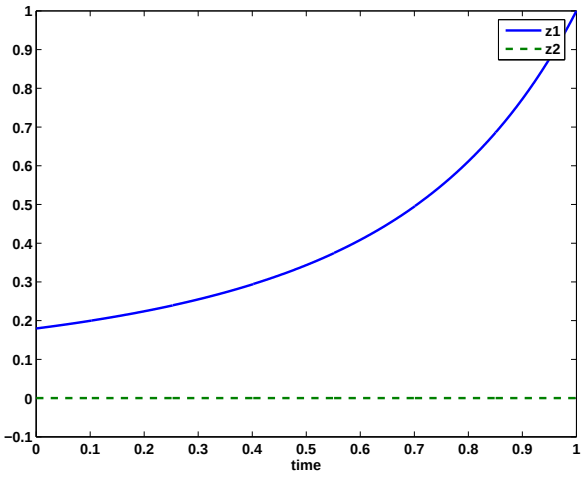


**Figure 7.8:** The dual solution to problem (7.5).

highly from the reduced dual solution, which is calculated to be

$$z(t) = \begin{pmatrix} e^{999(1-t)} \\ -e^{999-997t} \end{pmatrix}.$$

This can easily be seen by reducing the original dual problem

$$\dot{z} + \begin{pmatrix} -1 & 3000u_1^2 - 4000u_1^3u_2 \\ 0 & -1000u_1^4 \end{pmatrix} z = 0$$

to

$$\dot{z}_1 - z_1 + \underbrace{(3000u_1^2 - 4000u_1^3u_2)\left(-\frac{1}{u_1^2}\right)}_{=1000,\text{ if } u_2 = 1/u_1} z_1 = 0.$$

Note that the fast part of the exact dual solution is

$$z_2(t) = 0.$$

This difference has a strong effect on the error estimator: The residual for the first equation is clearly zero up to rounding errors, because the first part of the source term is independent of the second variable $u_2$. The estimated error depends therefore only on the term $\rho_2 \cdot z_2$. With $z_2(t) = 0$ for the detailedly calculated dual solution, the error is clearly estimated to be zero, whereas the reduced dual solution leads to a strong overestimation and therefore to detailed time steps in the next solution cycle.

Even though the combination of the reduced dual solver with few detailed time steps for the original problem might be advantageous from the computation cost of view, the above analytic results are somehow misleading. The reason is the large values of the reduced dual solution. Even though the dual solution can analytically be given in this example, it cannot be obtained numerically, simply because $e^{999}$ is larger than the biggest possible double value. The question about the applicability of the reductions to the dual problem is therefore to be split into two questions:

- Can the reduced dual problem be solved on a computer?
- Given that the reduced solution can technically be calculated, how accurate is the estimated error?

**The first question.** The first question can at least be partly answered by considering the products of a chemical reaction: If a variable describes a product of a reaction and is considered to be fast, then the reduced dual problem cannot be solved numerically.

In order to substantiate this statement, consider a linear reaction with three species $sp_1$, $sp_2$ and $sp_3$, where $sp_3$ is supposed to be the product. The reaction can therefore be described by the ODE

$$\dot{u} = \begin{pmatrix} -a_{11} & a_{12} & a_{13} \\ a_{21} & -a_{22} & a_{23} \\ a_{11} - a_{21} & -a_{12} + a_{22} & -a_{13} - a_{23} \end{pmatrix} u$$

with the coefficients $a_{ij} > 0$. Moreover, since $sp_3$ is the product, the coefficient $a_{13} + a_{23}$ is much smaller than $a_{11} - a_{21}$ and $-a_{12} + a_{22}$. This implies that at least one of the coefficients $a_{11}$ and $a_{22}$ is big, whereas the other coefficients are comparably small. For a concrete example, consider reaction (4.16). Since the ODE contains the conservation law $\sum_i u_i = 1$, the reaction can be modeled by replacing one equation by the conservation law. Here, the replacement of the second and third equation will be investigated in order to see the effect, if the product is still in the reaction formalism or not. For both cases, $sp_1$ will be considered to be the process variable (which implies the smallness of $a_{11}$, therefore $a_{22}$ is the biggest coefficient in the above ODE).

Equivalent formulations of the above dynamical system are

$$\begin{pmatrix} \dot{u}_1 \\ \dot{u}_2 \end{pmatrix} = \begin{pmatrix} -a_{11} - a_{13} & a_{12} - a_{13} \\ a_{21} - a_{23} & -a_{22} - a_{23} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} + \begin{pmatrix} a_{13} \\ a_{23} \end{pmatrix}$$

and

$$\begin{pmatrix}\dot{u}_1\\ \dot{u}_3\end{pmatrix} = \begin{pmatrix}-a_{11}-a_{12} & -a_{12}+a_{13}\\ a_{11}+a_{12}-a_{21}-a_{22} & a_{12}-a_{13}-a_{22}-a_{23}\end{pmatrix}\begin{pmatrix}u_1\\ u_3\end{pmatrix} + \begin{pmatrix}a_{12}\\ a_{22}-a_{12}\end{pmatrix},$$

if the missing variable is computed by the conservation laws within the post–processing. The corresponding reduced dual problems read

$$\dot{z}_1 + \left(\frac{(a_{21}-a_{23})^2}{a_{22}+a_{23}} - a_{11} - a_{13}\right) z_1 = 0, \qquad z_1(T) = 1$$

and

$$\dot{z}_1 + \left(\frac{(a_{11}+a_{21}-a_{21}-a_{22})^2}{a_{13}+a_{22}+a_{23}-a_{12}} - a_{11} - a_{12}\right) z_1 = 0, \qquad z_1(T) = 1.$$

Clearly, the factor in front of $z_1$ is small for the first dual problem and large for the second dual problem, because only $a_{22}$ is big. At least in the second case, this factor is even positive, which leads to the very large values of $z_1(0)$. For reaction (4.16), these factors are approximately 5 and 914 for the first and second equation respectively, so $z_1(0) \approx e^5$ for the first dual problem and $z_1(0) \approx e^{914}$ for the second problem, if $T = 1$. Clearly, even for this simple problem, the reduced dual solution cannot be calculated numerically, if the product $sp_3$ is modeled, because $e^{914}$ is much bigger than the biggest possible double value.

**The second question.** Recall from lemma 6.2.1 that the linearization of the function $\psi$ can be applied to problems of the form $\dot{y} + By = 0$, if $B$ was created with a nonlinear state $u$ on the manifold. Since the last section showed that the dual problem can be created with the reduced solution, this condition is perfectly fulfilled. But does the applicability of the linearized table $\nabla\psi$ to problem $\dot{y} + By = 0$ imply the applicability of the same reduction method to $\dot{z} + B^*z = 0$? This is clearly true, if $B = B^*$, which is at least for realistic problems almost never the case. For realistic problems, the reutilization of $\nabla\psi$ will lead to less accurate solutions, the less symmetric the matrix $B$ is.

A criteria for the grade of symmetry of $B$ is for example the minimal angle between the eigenvectors. $B$ is clearly symmetric, if the eigenvectors are pairwise orthogonal. Moreover, the minimal angle between the eigenvectors depends continuously on the symmetry of the matrix $B$: If the symmetric matrix $B$ is disturbed by a nonsymmetric small $\Upsilon$, then the maximal scalar product of the eigenvectors depends clearly continuously on the eigenvectors, which depend continuously on $\Upsilon$. This can be proven by the implicit function theorem applied to the function

$$F(A, X) = F\left(A, \begin{pmatrix}\lambda\\ x\end{pmatrix}\right) = \begin{pmatrix}Ax - \lambda x\\ \sum_i x_i - 1\end{pmatrix}.$$

In order to see the effect of the missing symmetry, consider the problem

$$\dot{u} = \begin{pmatrix}-1 & a_{12}\\ a_{21} & -100\end{pmatrix} u, \qquad u(0) = \begin{pmatrix}1\\ 0\end{pmatrix} \tag{7.7}$$

and compare the error produced by the reduction methods at time $t = 1$ with its estimations for varying $a_{12}$ and $a_{21}$. The result can be seen in tables 7.4 and 7.5, where either $a_{12}$ or $a_{21}$

| $a_{21}$ | $\lvert u(1) - u_{red}(1)\rvert$ | $\eta$ | $\eta_{red}$ |
|---|---|---|---|
| 1 | $3.72e-05$ | $3.69e-05$ | $3.70e-05$ |
| 11 | $4.03e-04$ | $4.03e-04$ | $7.64e-03$ |
| 21 | $7.43e-04$ | $7.54e-04$ | $2.15e-01$ |
| 31 | $1.05e-03$ | $1.07e-03$ | $3.10e+01$ |
| 41 | $1.34e-03$ | $1.34e-03$ | $3.28e+04$ |
| 51 | $1.76e-03$ | $1.52e-03$ | $2.52e+08$ |

**Table 7.4:** The error and its detailed and reduced estimates for problem (7.7) with $a_{12} = 1$.

| $a_{12}$ | $\lvert u(1) - u_{red}(1)\rvert$ | $\eta$ | $\eta_{red}$ |
|---|---|---|---|
| 1 | $3.72e-05$ | $3.69e-05$ | $3.70e-05$ |
| 11 | $3.90e-02$ | $3.99e-02$ | $3.93e-02$ |
| 21 | $8.21e-02$ | $8.39e-02$ | $8.27e-02$ |
| 31 | $1.30e-01$ | $1.32e-01$ | $1.31e-01$ |
| 41 | $1.82e-01$ | $1.86e-01$ | $1.84e-01$ |
| 51 | $2.39e-01$ | $2.45e-01$ | $2.42e-01$ |

**Table 7.5:** The error and its detailed and reduced estimates for problem (7.7) with $a_{21} = 1$.

is set to one respectively. The estimate obtained with the reduced dual solution is clearly far beyond acceptable for large values of $a_{21}$, but is still very reasonable, if the matrix entry $a_{12}$ becomes large.

The explanation for this phenomenon is rather simple: If $a_{21}$ is zero (or at least small), then the matrix defining the dual problem reads

$$A^* = \begin{pmatrix} a_{11} & 0 \\ a_{12} & a_{22} \end{pmatrix}$$

and the influence of the table is neglected in the calculation of the slow dual solution because of the zero in the upper right part of $A^*$. The table influences therefore only the fast variables of the dual solution. They are approximated to be decreasing or even to be zero from the very beginning, because the decreasing of the fast variables was a necessary requirement for the numerical solvability of the dual problem. If now the exactly calculated fast dual variables are decreasing, the approximation by the tables is at least qualitatively correct.

A more interesting and nonlinear example is given by (7.5), where the gradient

$$\nabla f(u) = \begin{pmatrix} -8u_1 & 4u_2 & 0 \\ 8u_1 & -4004u_2 & 360u_3 \\ 0 & 4000u_2 & -2360u_3 \end{pmatrix}$$

is clearly nonsymmetric, but the coupling from the fast to the slow variables $(8u_1, 0)^T$ is at least for certain values of $u_1$ and $u_2$ small compared to the entries in $A_{22}$. The reduced dual solution for this problem shows similar properties as the detailed dual solution, see figure 7.9. The small differences in the dual solutions does of course lead to differences in the estimation of the errors. The exact error for the initial values $u(0) = (0.9, \psi(0.9))$ at time

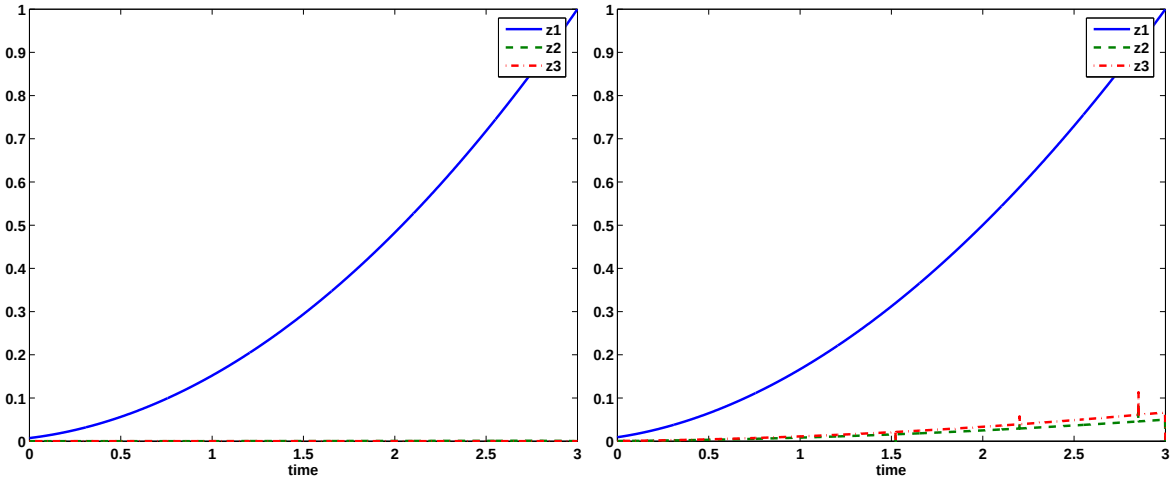


**Figure 7.9:** The detailed (left) and reduced (right) dual solution to problem (7.5).

$t = 3$ is $1.9 \cdot 10^{-2}$ and is estimated to be $2.5 \cdot 10^{-2}$ with the detailed and 1.1 with the reduced dual solution. But the effect to the adaptive application of the reduction methods is similar. The detailed dual solution leads to detailed time steps in the time interval $[1.27, 1.94]$, the reduced dual solution to the slightly bigger interval $[1.24, 1.94]$, where a detailed solver is to be applied.

The experiences with the considered problems showed clearly that the dual problem cannot always be reduced with the tables created for the original problem. In case of controlling the fast variables, the dual solution is zero and does therefore not contain any information about the produced error. If the process variables are to be controlled, the situation is different, but only more optimistic, if the fast variables are decreasing and the coupling of the fast variables to the slow equations is small. For practical problems this will hardly be the case, the application of the reduction methods to the dual problem is therefore not recommended.

This observation is probably surprising to the readers, who are familiar with [2], where a method was presented, how the error of a reduced problem can be controlled with a reduced dual solution. In the cited paper, the problem

$$a(u, \phi) + d(u, \phi) = (f, \phi) \quad \forall \phi \in V$$

was considered, where the reduced solution $u_m$ solves only the equation

$$a(u_m, \phi) = (f, \phi) \quad \forall \phi \in V.$$

The equations describe for example a reaction–diffusion equation, where the term $a$ describes the diffusion phenomena by Fick's law and $d$ the difference between this relatively simple diffusion model and the more detailed multicomponent model. It is shown that the error identity

$$j(e) = -d(u_m, z) = -d(u_m, z_m) + O(\|d\|^2)$$

holds, where $e = u - u_m$ is the difference between the detailed and reduced solutions and $z$ denotes the detailed dual solution. The reduced dual solution is denoted by $z_m$. This

identity indicates clearly, that the error can be estimated by the reduced dual solution. But this identity can only be achieved, because

$$\|z - z_m\| \leq \alpha \|d\| \|z\|$$

holds, which shows that the difference between the detailed and reduced dual solution is relatively bounded. The observation of the reduced dual solutions for the problems of this thesis shows clearly, that such an estimate cannot be achieved for the reduction methods QSSA and ILDM.

### 7.2.4 Numerical example

In order to treat a more realistic example, a reaction mechanism with 20 species is considered. The mechanism itself is given by the reactions

$$\begin{gathered}
sp_1 \underset{1}{\overset{2}{\rightleftharpoons}} sp_2 \\
sp_1 + sp_2 \xrightarrow{1000} sp_1 + sp_3 \\
sp_2 \xleftarrow{90} sp_3 \\
sp_3 \underset{90}{\overset{1000}{\rightleftharpoons}} sp_4 \underset{90}{\overset{1000}{\rightleftharpoons}} sp_5 \quad ... \quad sp_{18} \underset{90}{\overset{1000}{\rightleftharpoons}} sp_{19} \xrightarrow{500} sp_{20} \\
3sp_4 + 2sp_5 \underset{90}{\overset{1000}{\rightleftharpoons}} 5sp_{18}.
\end{gathered} \tag{7.8}$$

The second elementary reaction contains $sp_1$ as third body, which can formally be interpreted as a reaction from $sp_2$ to $sp_3$ with the reaction rate $k = 1000 \cdot y_1$ depending on the mass fraction of $sp_1$. A second source for nonlinearities is the last reaction. It ensures that each time step takes on average 3 Newton steps to converge. Note that more nonlinearities in the reaction mechanism does not necessarily increase the number of Newton steps required in each time step.

In order to simplify the investigations of this mechanism, the source term $f$ is reduced by the mass fraction of the 19–th species. This value can be obtained by $y_{19} = 1 - \sum_{i \neq 19} y_i$. The solution to the ODE $\dot{u} = f(u)$ with $f$ being the source term of the above reaction mechanism and initial conditions $u(0) = (1, 0, ..., 0)^T$ can be seen in figure 7.10. The figure on the right hand side shows clearly the effect of the second elementary reaction, which depends strongly on the current state of species $sp_1$. For an even further understanding of this mechanism, the evolution of the eigenvalues along the solution trajectory is visualized in figure 7.11. The left eigenvector to the slowest eigenvalue is largest in its first entry, which suggests to take the mass fraction of species $sp_1$ as process variable. The variable related to the second slowest eigenvalue is the mass fraction of species $sp_2$. $sp_2$ can therefore be treated as process variable or as fast variable. Three variables are now favorable as control variables for the error estimation:

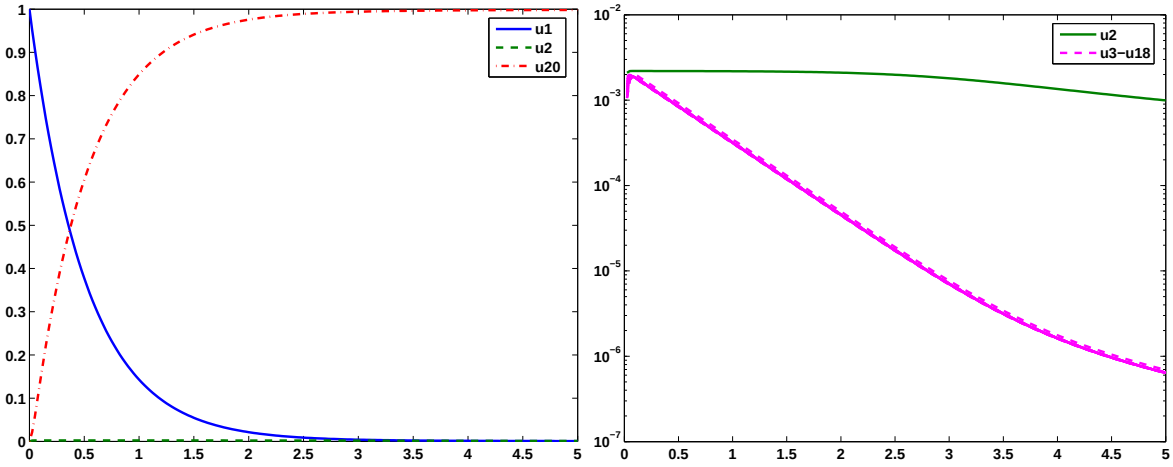


**Figure 7.10:** The solution to the ODE $\dot{u} = f(u)$ with $f$ describing the source term of (7.8), the right figure shows a different scale.

- species $sp_1$, because it is the process variable
- species $sp_2$, which is a physically fast variable in the beginning of the solution process, but becomes slow close to the stationary point
- species $sp_{20}$, because it is the product.

This small analysis is substantiated by figure 7.12, which shows that these three variable have a different error behavior compared to the rest of the species.

### Adaptive solution process

The adaptive solution strategy leads to the errors shown in figure 7.13, if the process variable $sp_1$ is controlled. This improvement compared to the reduced error is obtained by doing 40% (52%) of the time steps with a detailed solver, if the error at the final time (averaged error) is controlled. Considering other variables as control variables leads basically to the same solution. The only difference occurs by controlling the averaged error of $sp_{20}$, because the error at the beginning of the solution process is very high. For that reason, detailed solution steps are performed also at the start of the iteration.

The computation time for the solution processes for the ODEs $\dot{u} = f(u)$ with $f$ describing the above reaction mechanism (7.8) or similar mechanisms with 40 and 80 variables can be seen in table 7.6. The values describe the computation times for the detailed solver, the reduced solver and the adaptive solver. The adaptive solver takes 2 adaptive cycles in order to reduce the controlled error by $1/10$, therefore the computation time for only one cycle is also given. But even for performing only one adaptive cycle, the detailed solver turns out to be cheaper.

Due to the exact solution process for the dual solution, this result is probably not surprising. In order to obtain reasonable results, the dual problem has to be solved with the reduction

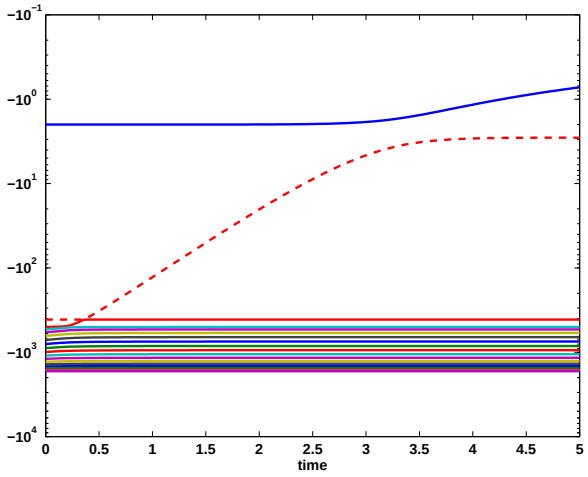


**Figure 7.11:** The evolution of the eigenvalues of reaction mechanism (7.8) along the solution trajectory (compare figure 7.10).

| | 20 | 40 | 80 |
|---|---|---|---|
| detailed | 0.6 | 4.1 | 32 |
| reduced | 0.3 | 0.7 | 2.1 |
| adaptive | 2.0 | 12.7 | 94 |
| adaptive (1 cycle) | 1.2 | 6.6 | 47 |

| | estimated | measured |
|---|---|---|
| reduced error | $2.7e-4$ | $2.4e-4$ |
| adaptive error | $1.4e-5$ | $1.4e-5$ |

**Table 7.6:** The computation times for reaction mechanisms with 20, 40 and 80 variables. The number of process variables is in all cases 1. The right table shows the estimated and exactly calculated errors.

methods. In the above constellation, this is clearly impossible, because the product $sp_{20}$ is considered as a fast variable, the reduced dual problem can therefore not be solved numerically. So $sp_{20}$ has either to be the process variable or the variable, which is obtained via post–processing by the evaluation of the conservation laws. For the latter, the computation times can be seen in 7.7. The computational costs for the detailed and reduced solver are clearly similar to the case, where the species $sp_{19}$ was neglected in the computation, compare table 7.6.

## 7.3 A posteriori control of modeling and discretization errors

The application of the reduction methods to the solution process for ODEs leads to two different errors, namely to the error produced by the reduction itself and the discretization error. The last two sections concentrated on the reduction errors, in this section, a strategy will be presented, how the reduction and the discretization error can be estimated at the same time. This will be done by considering the dG(0) algorithm, which is equivalent to the implicit Euler. For general considerations, see for example [2].

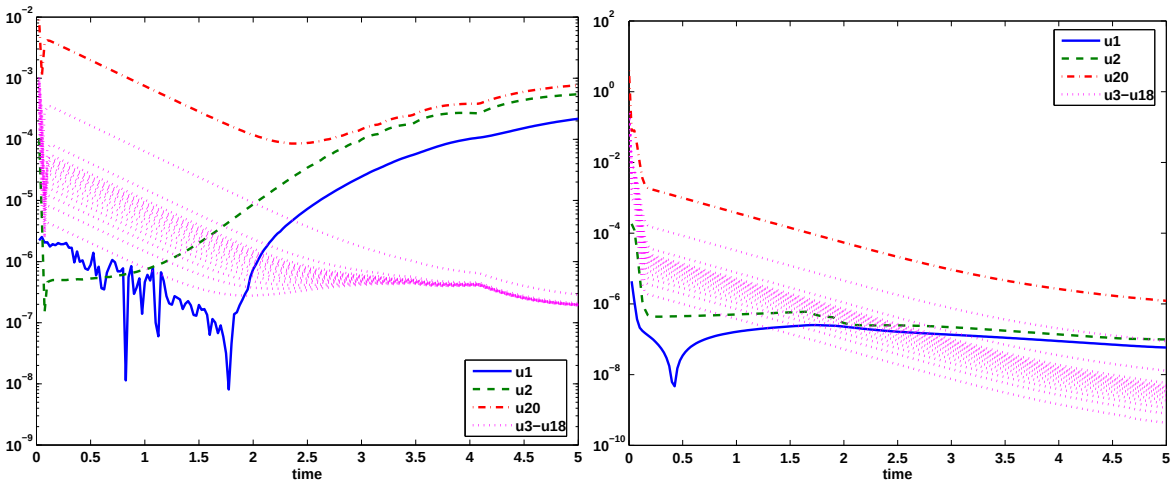


**Figure 7.12:** The difference between the exact and the reduced solution to $\dot{u} = f(u)$ with reaction (7.8). The reduced solution for the left figure was calculated with one process variable, the right figure with two.

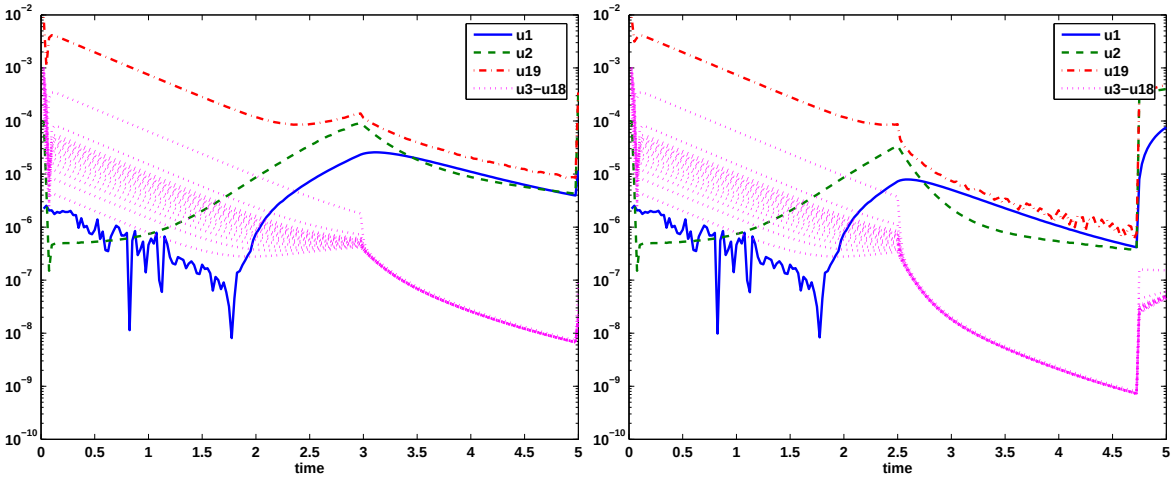


**Figure 7.13:** The errors produced by the adaptive strategy, if the error of species $sp_1$ is controlled. In the left figure, the error at the final time is controlled, in the right figure the averaged error over the whole time interval.

| | 20 | 40 | 80 |
|---|---|---|---|
| detailed | 0.6 | 4.6 | 36 |
| reduced | 0.4 | 0.8 | 2.4 |
| adaptive | 1.1 | 3.4 | 17.2 |

| | estimated | measured |
|---|---|---|
| reduced error | $1.9e-4$ | $8.4e-4$ |
| adaptive error | $5.9e-5$ | $3.2e-5$ |

**Table 7.7:** The computation times for reaction mechanisms with 20, 40 and 80 variables. The number of process variables is in all cases 1. For the adaptive calculation, the dual solution is obtained by a reduced solver. The second table shows the errors at time $t = 5$.

For the reduced implicit Euler, four problems are considered: The detailed problem, the reduced problem and the detailed and reduced discretized problems, compare figure 7.14. The reduced implicit Euler leads to the solution $u_{h,red}$ of the reduced discretized problem,

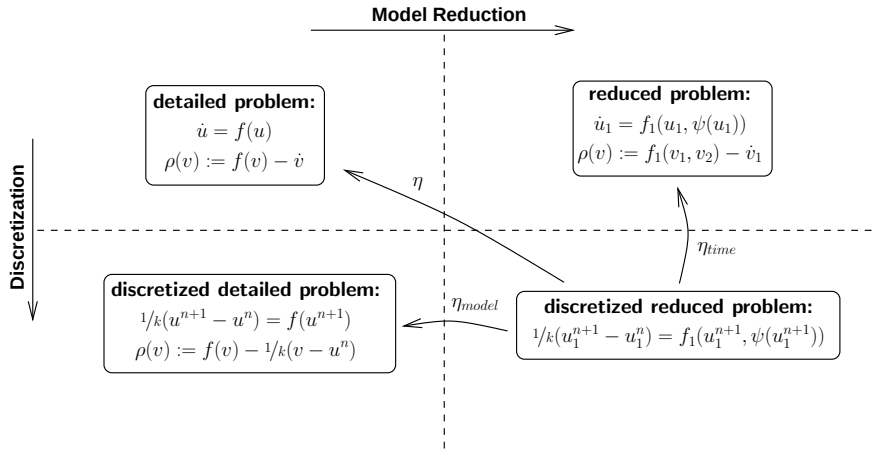


**Figure 7.14:** The error estimators for the reduced discretized problem.

where its difference to the solution of the detailed problem is to be estimated. On top of that, the estimate is to be split into a part resulting from the model reduction and a second part, which represents the error resulting from the time discretization. Instead of calculating

$$\eta = \rho(u_{h,red})(z),$$

the estimates

$$\eta_{time} = \rho_h(u_{h,red})(z) \quad \text{and} \quad \eta_{model} = \rho_m(u_{h,red})(z)$$

are calculated for a suitable $z$ and $\eta$ is estimated by $\eta \approx \eta_{time} + \eta_{model}$.

The estimate $\eta_{model}$ can be calculated by considering the dual solution, which solves the dual problem of the discretized detailed problem, compare section 7.2. The estimator $\eta_{model}$ is then obtained by testing the residual with the obtained dual solution. In order to obtain a reasonable estimator $\eta_{time}$, it is clearly necessary to calculate the dual solution of the reduced problem created with the discretized reduced solution.

This strategy leads clearly to exact error estimators $\eta_{model}$ and $\eta_{time}$, but has the big disadvantage that two dual problems have to be solved, where at least the dual problem for the estimator of the discretization error can only be approximated. Less accurate estimators can be obtained by considering an approximation of the dual solution of the detailed problem for both estimators. Then only one dual problem is to be solved, which reduces the computational costs.

This approximation of the dual solution cannot be obtained by accepting the dual solution $z_h$ of the discretized detailed problem, because $z_h \in S^{(r)}(I)$ and therefore $\rho(u_h)(z_h) = 0$, which leads to a bad approximation of the estimator $\eta_{time}$. It has therefore to be guaranteed that the accepted approximation $\tilde{z}_h$ is not in $S^{(r)}(I)$. This can be achieved by two possibilities:

- Calculate $\tilde{z}_h$ with a higher order solver.

- Accept a higher order interpolation of $z_h$ as $\tilde{z}_h$.

Here, the second possibility is taken. The by the implicit Euler obtained dual solution $z_h$ is discontinuous and piece–wise constant and can easily be interpolated to a continuous and piece–wise linear function by

$$\tilde{z}_h(t) = z_h(t_n) + \frac{t - t_n}{t_{n+1} - t_n}(z_h(t_{n+1}) - z_h(t_n))$$

for $t_n \leq t \leq t_{n+1}$. Clearly, $\tilde{z}_h \notin S^{(r)}(I)$, which implies $\eta_{time} = \rho_{red}(u_{h,red})(\tilde{z}_h) \neq 0$.

This leads to the following solution process:

1. Calculate the reduced solution for a given time discretization.
2. Solve the discretized dual problem and obtain $z_h$.
3. Interpolate $z_h$ and obtain $\tilde{z}_h$.
4. The residuals tested with the dual solutions $z_h$ and $\tilde{z}_h$ lead to the estimates $\eta_{model}$ and $\eta_{time}$ and therefore to a total estimate $\eta \approx \eta_{model} + \eta_{time}$.
5. Find the intervals, which have the biggest influence on the estimators $\eta_{time}$ and $\eta_{model}$. If the effect of the model error in the considered interval is higher than the effect of the discretization error, mark the interval, such that a detailed solver is to be applied in future, and set $\nu_{model} = 0$ in this interval. Otherwise split the corresponding time interval in two parts and set $\nu_{time} = 0$ herein.
6. Repeat step 5, until a tolerance is reached.
7. Calculate a new solution and the corresponding error estimators. If the error is still too large, repeat steps 5–7.

The performance of this strategy shall exemplarily be shown with problem (7.6). The initial time discretization shall be set to 50 equidistant time steps. The error of the fast variable of the reduced solution is $1.1 \cdot 10^{-1}$ and shall be reduced by the adaptive strategy by 90%. After four adaptive cycles, the error is $1.5 \cdot 10^{-2}$ (estimator: $1.1 \cdot 10^{-2}$), where the error by the reduction mechanisms is $7.6 \cdot 10^{-4}$ (estimator: $7.5 \cdot 10^{-4}$) and the discretization error is estimated to be $10^{-2}$. An exact measure of the discretization error is not possible. In the last adaptive cycle, 85 time steps were performed.

In the final solution cycle, the error is dominated by the discretization error for an obvious reason: Doing a detailed instead of a reduced solution step, leads to an almost zero model error in the corresponding interval, whereas an error remains, if the the time interval is splitted into two parts.

The distribution of the time steps is visualized in figure 7.15, as well as the difference between the obtained solution and the solution obtained by a detailed solver with the same time discretization. The figure shows clearly, that the switching in the type of the solver has a bigger influence than the mesh refinement. It also shows that the model adaption is independent of the refinement of the time discretization: The problem is solved with the reduction methods at time $t = 0.6$, but a mesh refinement was performed. And even though

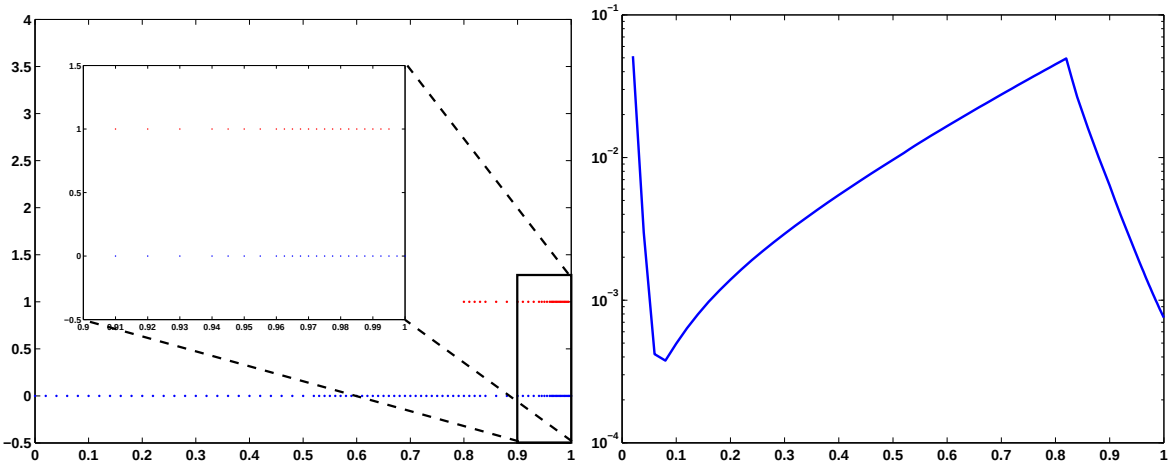


**Figure 7.15:** The time distribution for problem 7.6 and the resulting error, if the adaptive algorithm is applied.

the mesh is twice refined towards $t = 1$, the last two time steps are solved with the reduced solver.

# 8 Conclusion and Outlook

## 8.1 Conclusion

This thesis showed first various possibilities to apply the reduction methods QSSA and ILDM to stationary problems with linear and nonlinear reaction mechanisms. A reduction method using the tables was presented as well as several possibilities, how the tables may be applied as preconditioners. Secondly, the application of the QSSA– and ILDM–method to instationary homogeneous reactors was presented.

Even though it could be shown that at least ILDM reduces some problems exactly, the results for the stationary problems are rather frustrating:

- Preconditioning with the pure reduction methods cannot lead to convergence. An analytic proof for this statement was presented.
- Applying modified tables as preconditioners can have nice convergence properties, but the iteration is almost as expensive as preconditioning with standard methods like ILU.
- The quasi–Newton method, where the Newton update is calculated with a reduced solver, does not converge.
- As in the case of the preconditioners, the modified tables can lead to convergence of the quasi–Newton method, but the computational price is very high.

This leads to the conclusion that QSSA and ILDM are not suitable for the calculation of stationary problems, if systematic errors are to be avoided. This is especially true for iterations, where the pure reduction methods are used as preconditioners.

The application of the reduction methods to instationary problems leads to more optimistic results. By calculating the dual solutions, a relation between the current residual and the produced error is obtained, therefore an optimal distribution of detailedly calculated time steps can be found, such that the error is reduced to a certain fraction of the error produced by the totally reduced solution strategy. But even here, the computational costs are very high due to the effort, which is to be spent for the dual solution. Still, the presented iteration can successfully be applied, if the source term is highly nonlinear, such that the Newton iteration for each time step needs several iterations to converge. The more Newton iterations per time step are needed for the detailed model, the more effective is the presented mixed solution strategy.

## 8.2 Outlook

### 8.2.1 Application of incomplete tables

In this thesis, the reducing function $\psi$ obtained from QSSA or ILDM was always well defined for the whole possible domain of the process variables. As a consequence, the representing table was complete, the value $\psi(x_1)$ can therefore be interpolated by the tabulated surrounding values for all $x_1$. But for practical problems, the equations defining the function $\psi$ implicitly are not necessarily solvable. This effect leads to incomplete tables, where some of the tabulated points do not contain information about the fast variables. Therefore methods have to be invented, how the introduced applications of the tables can be transformed, such that incomplete tables can also be used.

Another possibility to circumvent this problem is the creation of trajectories, where all values of the process variables have a corresponding value for the fast variables. A quite promising method minimizes the entropy production and is introduced in [27].

### 8.2.2 Adaptivity in ODEs

In section 7.1, it was shown, how the existence of a table can be used to solve ODEs adaptively, such that the computational effort is reduced, but also the introduced systematic error is controlled. The control switched from the reduced solver to the detailed solution process, if the error is too big. As already seen, the switching from considering only few process variables to considering all variables as process variables, makes the strategy very expensive.

An interesting continuative question is therefore, if there is a possibility to increase the accuracy of the obtained solution with less effort than the detailed solver needs. One possibility is to have more than only one table in a hierarchy with different numbers of process variables. Tables with more process variables perform better, but are also more expensive in the application, because more equations are to be solved in the solution process. Then the controlling function has the possibility to redo a reduced time step with slightly higher effort, but still with much less computational costs than the detailed solver produces. The applicability of this idea and its performance is part of forthcoming research.

### 8.2.3 Adaptivity in PDEs

A PDE involving a given reaction system with a corresponding table $\psi$ offers up to now only two possibilities: Reduce the problem with ILDM or do not use the reduction methods at all. But in many practical applications, the reduction methods lead to very accurate results in certain regions, but produce large systematic errors in other parts of the domain, for example the cold regions in a flame. It is therefore desirable to have a solution strategy, which allows the reduction on some nodes of the grid, but uses the detailed solving strategy in other nodes.

Forthcoming projects will therefore have to create and investigate algorithms, which allow to solve problems of the form

$$\begin{pmatrix} A_1 + B_1 & B_2 \\ B_3 & A_2 + B_4 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \end{pmatrix},$$

which represents a discretized PDE on a grid with two nodes. The algorithm for the solution process shall allow to solve the problem with reduction methods for the first node and with a detailed solving strategy for the second node.

One possibility to tackle this problem is the application of splitting strategies, which split the above matrix into two operators, where one contains only the chemical parts $A_i$ and the other contains the blocks $B_i$. Then the chemical part can be solved node–wise, which allows the application of the reduction methods on certain nodes.

The Strang splitting (see also section 5.4) strategy might for example be applied, if the above equation results from a time–dependent problem. Then it is to be investigated, to which extend the theory of dual problems as in section 7.2 can be applied. The problem is that within the Strang splitting only one chemical time step will be applied, whereas the theory about the dual problems is based on several time steps.

Other splitting schemes, which might also be applied, are the Godunov splitting scheme [20]. For several other splitting schemes, see [19].

## Acknowledgments

This thesis was supported by the International Graduiertenkolleg IGK 710 (“Complex processes: Modeling, Simulation and Optimization”) and by the SFB 359 (“Reaktive Strömungen, Diffusion und Transport”).

I would like to express my gratitude to my supervisor Rolf Rannacher for the endless and fruitful discussions, especially towards the end of this thesis. His help and encouragement was very much appreciated.

A special thank goes also to Malte Braack, Dominik Meidner and Michael Schmich, who did not hesitate to discuss more detailed and technical aspects. I also want to thank Dirk Lebiedz and Volkmar Reinhardt, who provided me with insight into the computation of ILDM–tables in their work group.

# Bibliography


[1] Arrhenius, S.: "Über die Reaktionsgeschwindigkeit bei der Inversion von Rohrzucker durch Säuren", *Zeitschrift für Physikalische Chemie*, Vol. 4, 226–248 (1889)

[2] Braack, M., Ern, A.: "A posteriori control of modeling errors and discretization errors", *Multiscale Model. Simul.*, Vol. 1, No. 2, 221–238 (2003)

[3] Becker, R., Braack, M.: Gascoigne — High Performance Adaptive Finite Element Toolkit, http://www.gascoigne.de/

[4] Becker, R., Rannacher, R.: "An optimal control approach to error estimation and mesh adaptation in finite element methods", *Acta Numerica 2000*, 1–101, Cambridge University Press (2001)

[5] Bilger, R.W., Kee, R.J.: *Simplified Kinetics for Diffusion Flames of Methane in Air.* Western States Fall Meeting, The Combustion Institute, Paper no. 87-85, 1987.

[6] Birgin, E.G., Krejić, N., Martinez, J.M.: "Globally convergent inexact quasi–Newton methods for solving nonlinear systems", *Numerical Algorithms*, Vol. 32, 249–260 (2003)

[7] Bodenstein, M.: "Chemische Kinetik", *Ergebn. exakt. Naturwiss.*, Vol. 1, 197–209 (1922)

[8] Bongers, H., van Oijen, J.A., de Goey, L.P.H., "Intrinsic Low-Dimensional Manifold Method Extended with Diffusion", *Proc. Combust. Inst.*, Vol. 29, 1371-1378 (2002)

[9] Böttcher, K., Rannacher, R.: "Adaptive Error Control in Solving Ordinary Differential Equations by the Discontinuous Galerkin Method", Preprint 96–53, SFB 359, Universität Heidelberg, Germany (1996)

[10] Buleev, N.I.: "A numerical method for the solution of two–dimensional and three–dimensional equations of diffusion", *Math. Sb.*, Vol. 51, 227–238 (1960)

[11] Correa, C., Niemann, H., Schramm, B., Warnatz, J.: "Reaction Mechanism Reduction for Higher Hydrocarbons by the ILDM Method", *Proc. Comb. Inst.*, Vol. 28, 1607-1614 (2001)

[12] Deuflhard, P., Heroth, J.: "Dynamic Dimension Reduction in ODE Models", *Scientific Computing in Chemical Engineering*, John Wiley & Sons Ltd., 205–212

[13] Deuflhard, P., Hohmann, A.: *Numerische Mathematik I*, Walter de Gruyter, 1993

[14] Dibble, R.W., Maas, U., Warnatz, J.: *Verbrennung*, Springer–Verlag, 2001.

[15] Eisenstat, S.C., Walker, H.F.: "Globally convergent inexact Newton methods", *Siam J. Optimization*, Vol. 4, 393–422 (1994)

[16] Dembo, R.S., Eisenstat, S.C., Steihaug, T.: "Inexact Newton methods", *Siam J. on Numerical Analysis*, Vol. 19, 400–408 (1982)

[17] Fick, A.: "Über Diffusion", *Annu. Phys. Leipzig*, Vol. 94, 59–86 (1855)

[18] Gerschgorin, S.: "Über die Abgrenzung der Eigenwerte einer Matrix.", *Izv. Akad. Nauk. UdSSR Otd. Fiz.-Mat. Nauk*, Vol. 7, 749–754 (1931)

[19] Glowinski, R., Le Tallec, P.: *Augmented Lagrangian and Operator Splitting Methods in Nonlinear Mechanics*, Siam Studies in Applied Mathematics, Philadelphia, 1989

[20] Godunov, S.K.: "Finite difference method for numerical computation of discontinuous solution of the equations of fluid dynamics", *Matematicheskii Sbornik*, Vol. 47, 271–295 (1959)

[21] Golub, G.H., Van Loan, C.F.: *Matrix Computations*, John Hopkins University Press, 1990

[22] Gonnet, G.H.: *Handbook of Algorithms and Data Structures*, Addison–Wesley, 1984

[23] Kelley, C.T.: "Iterative methods for linear and nonlinear equations", *Siam Publications*, Philadelphia (1995)

[24] Kozlov, R., Kværnø, A., Owren, B.: "The behaviour of the local error in splitting methods applied to stiff problems", *Journal of Computational Physics*, Vol. 195, 576–593 (2004)

[25] Lam, S.H., Goussis, D.A.: "The CSP Method for Simplifying Kinetics", *International Journal of Chemical Kinetics*, Vol. 26, 461–486 (1994)

[26] Lam, S.H.: "Using CSP to Understand Complex Chemical Kinetics", *Combustion Science and Technology*, Vol. 89, 375–404 (1993)

[27] Lebiedz, D.: "Computing minimal entropy production trajectories – an approach to model reduction in chemical kinetics", *Journal of Chemical Physics*, Vol. 15, 6890–6897 (2004)

[28] Maas, U., Pope, S.B.: "Simplifying Chemical Kinetics: Intrinsic Low–Dimensional Manifolds in Composition Space", *Combustion and Flame*, Vol. 88, 239–264 (1992)

[29] Maas, U.: "Efficient calculation of intrinsic low–dimensional manifolds for the simplification of chemical kinetics", *Computing and Visualization in Science*, Vol. 1, 69–81 (1998)

[30] Mott, D.R.: "New Quasi-Steady-State and Partial-Equilibrium Methods for Integrating Chemically Reacting Systems", Dissertation, The University of Michigan, 1999

[31] Mohr, P. J., Taylor, N.: "CODATA recommended values of the fundamental physical constants: 2002", *Rev. Mod. Phys.*, Vol. 77, 1–107 (2005)

[32] Nelder, J.A., Mead, R.: "A Simplex method for function minimization", *Computer Journal*, Vol. 7, 308–313 (1965)

[33] Newton, I.: "Methodus fluxionum et serierum infinitarum.", (1664-1671)

[34] Niemann, H.: "Niedrigdimensionale Modellierung Dynamischer Systeme am Beispiel reduzierter Reaktionsmechanismen", Dissertation, Universität Heidelberg, 2002.

[35] Niemann, H., Schmidt, D., Maas, U.: "An efficient storage scheme for reduced chemical kinetics based on orthogonal polynomials", *Journal of Engineering Mathematics*, Vol. 31, Issue 2–3, 131–142 (1997)

[36] http://kinetics.nist.gov/

[37] Van Oijen, J.A., de Goey, L.P.H.: "Modelling of premixed laminar flames using flamelet–generated manifolds", *Combust. Sci. and Tech.*, Vol. 161, 113–137 (2000)

[38] Oliphant, T.A.: "An extrapolation process for solving linear systems", *Quart. Appl. Math.*, Vol. 20, 257–267 (1962)

[39] Peters, N. und Rogg, B.: "Reduced Kinetic Mechanisms for Applications in Combustion Systems", Springer, Berlin, 1993.

[40] Petzold, L., Zhu, W.: "Model reduction for chemical kinetics: An optimization approach", *AIChE Journal*, Vol. 20, 869–886 (1999)

[41] Polak, E.: *Optimization – Algorithms and Consistent Approximations*, Springer, 1997

[42] Ramshaw, J.D.: "Partial chemical equilibrium in fluid dynamics", *Phys. Fluids*, Vol. 23, 675–680 (1980).

[43] Ramshaw, J.D., Cloutman, L.D.: "Numerical method for partial equilibrium flow", *J. Comput. Phys.*, Vol. 39, 405–417 (1981)

[44] Schmidt, D.: "Modellierung reaktiver Strömungen unter Verwendung automatisch reduzierter Reaktionsmechanismen", Dissertation, Universität Heidelberg, 1996.

[45] Semenov, N.N., *Z. Phys. Chem.*, 48:571 (1928).

[46] Smooke, M. D. (Hrsg): *Reduced Kinetic Mechanisms and Asymptotic Approximations for Methane-Air Flames*, Lecture Notes in Physics 384, Springer, 1991.

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

[48] Walter, W.: *Gewöhnliche Differentialgleichungen*, Springer, 1997

[49] Watkins, D.S.: "Understanding the QR algorithm", *SIAM Review*, Vol. 24, No. 4, 427–440 (1982)

[50] Wilkinson, J.H.: *The Algebraic Eigenvalue Problem*, Oxford Science Publications, 1988.