Stiff Well–Posedness and Asymptotic Expansions for Hyperbolic Systems with Relaxation

Jens Lorenz * and H. Joachim Schroll †

October 27, 1997

Institut Mittag–Leffler, Report No. 3, 1997/98

Abstract

The Cauchy problem for linear constant–coefficient hyperbolic systems $u_t + \sum_j A^{(j)} u_{x_j} = (1/\delta)Bu + Cu$ in d space dimensions is analyzed. Here $(1/\delta)Bu$ is a large relaxation term, and we are mostly interested in the critical case where B has a non–trivial null–space. A concept of stiff well–posedness is introduced that ensures solution estimates independent of $0 < \delta \ll 1$. Under suitable assumptions, an asymptotic expansion for small δ is presented. Furthermore, we give sharp conditions for the existence of a limit of the L_2–solution as δ tends to zero. The evolution of this limit is governed by a reduced hyperbolic system, the so–called equilibrium system. The original system and the equilibrium system are related by a phase speed condition.

The theory is applied to a linearized version of a generic model of two–phase flow in a porous medium. Stiff well–posedness of this system is analyzed in detail, and the leading terms of the expansion are computed numerically.

Key words: Hyperbolic system, relaxation, well–posedness, asymptotic expansion, equilibrium limit, phase speed condition.

AMS subject classification (1991): 35L40, 35L45, 35C20

*Department of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131, USA (`lorenz@math.unm.edu`).

†Numerical Mathematics, RWTH–Aachen, D-52056 Aachen, Germany (`schroll@igpm.rwth-aachen.de`).

1 Introduction and Main Results

We consider multidimensional first–order systems with a large relaxation term

$$u_t + \sum_{j=1}^{d} A^{(j)} u_{x_j} = \frac{1}{\delta} Bu + Cu, \quad 0 < \delta \ll 1, \quad x \in \mathcal{R}^d, \quad t \geq 0. \tag{1}$$

Here $u(x,t,\delta) \in \mathcal{C}^n$ and $A^{(j)}$, B, C are constant complex $n \times n$ matrices. The number d of space dimensions is not restricted. It is assumed that the system is strongly hyperbolic. (This concept is reviewed in Section 2.) We are mostly interested in the critical case, where the matrix B has a non–trivial null–space. In applications, the large term $\frac{1}{\delta}Bu$ typically has a damping effect, leading to an initial layer except in the null–space of B. The term Cu, on the other hand, may lead to exponential growth, at rates independent of δ. Therefore, we refer to Cu as a source term.

The present paper extends the concept of stiff well–posedness, originally introduced for 1D problems [4], to systems in any number of space dimensions. In addition, an asymptotic expansion of the solution is derived and an application to two–phase flows is discussed.

The system (1) is called *stiff* if the relaxation parameter δ is small compared to the time scale determined by the maximal speed of propagation and some appropriate length scale.

Our aim is to study the behavior of the solutions of (1) when Cauchy data are given at time $t = 0$,

$$u(x, 0, \delta) = f(x), \quad x \in \mathcal{R}^d \ . \tag{2}$$

We will give necessary and sufficient conditions which ensure that for all initial data $f \in L_2$ and all $t > 0$ the solution $u(\cdot, t, \delta) = S(t,\delta)f$ of (1), (2) converges in L_2 as $\delta \to 0$. Here $S(t,\delta)$ denotes the solution operator of (1). Under our assumptions, we can also describe the evolution of the limit function

$$\breve{u}(\cdot, t) = \lim_{\delta \to 0+} u(\cdot, t, \delta), \quad t > 0 \ ,$$

by a reduced strongly hyperbolic system, which will be independent of δ. This system can be determined easily from a canonical form of the original stiffly well–posed system. The limit $\breve{u}$ evolves in the null–space of B.

An asymptotic expansion

$$u = u_0 + \delta u_1 + \ldots + \delta^m u_m + \mathcal{O}(\delta^{m+1})$$

of the solution will be derived in Section 3. The leading term u_0 consists of the equilibrium limit and an initial layer,

$$u_0(x,t,\delta) = \breve{u}(x,t) + u_{initial}(x, \frac{t}{\delta}).$$

Here $u_{initial}$ is determined by a family of stiff ODEs. The higher order terms u_1, u_2, etc. have similar decompositions.

It will become more clear in Section 3, how this expansion separates the two different time scales present in the original stiff system (1). As this decomposition is not limited to first order accuracy, like other straight forward splitting methods, it might be useful in deriving efficient numerical methods for stiff hyperbolic systems. Nevertheless, the present paper is devoted to the analysis of solutions of hyperbolic systems with stiff relaxation terms, rather than computational aspects. For the design of numerical methods, we refer to the work of Engquist et al. which is summarized in [5].

Before stating our main results we introduce some notation. With

$$\langle u, v\rangle = \sum_j \bar{u}_j v_j \quad \text{and} \quad |u| = \langle u, u\rangle^{1/2}$$

we denote the Euclidean inner product and norm in $\mathcal{C}^n$. The corresponding matrix norm is also denoted by $|\cdot|$. The L_2–norm of a vector function $u : \mathcal{R}^d \to \mathcal{C}^n$, for which every component is square integrable, is defined by

$$\|u\|^2 = \int_{\mathcal{R}^d} |u(x)|^2 dx \ .$$

Similarly, if $L : L_2 \to L_2$ is a bounded linear operator, its norm is $\|L\|$. With $\|\cdot\|_{H_m}$ we denote the usual Sobolev norm based on L_2,

$$\|u\|^2_{H_m} = \sum_{|\beta|\le m} \|D^\beta u\|^2, \quad D^\beta = \frac{\partial^{|\beta|}}{\partial x_1^{\beta_1} \ldots \partial x_d^{\beta_d}} \ , \quad |\beta| = \sum_{j=1}^d \beta_j, \quad \beta \in \mathcal{N}_0^d \ .$$

The result stated next characterizes all strongly hyperbolic systems (1) for which $S(t,\delta)f$ converges for every $f \in L_2$ as $\delta \to 0$.

Theorem 1.1 *Assume that the system (1) is strongly hyperbolic.*
a. Suppose the following two conditions are satisfied:

- *COND1 There is a constant c_1 so that*

$$\left| \exp\left(\left[B - i \sum_{j=1}^d \omega_j A^{(j)} \right] \tau \right) \right| \le c_1 \quad \forall \omega \in \mathcal{R}^d, \quad \forall \tau \ge 0 \ ;$$

- *COND2 The matrix B has no purely imaginary eigenvalue different from zero.*

Then, for all $f \in L_2$ and all $t > 0$, the limit of $S(t,\delta)f$ exists in L_2 as $\delta \to 0$.

b. Conversely, if for some $t > 0$ the limit of $S(t,\delta)f$ as $\delta \to 0$ exists in L_2 for every $f \in L_2$, then the above two conditions hold.

Let us note that necessary and sufficient conditions for *COND1* are given by the Kreiss–matrix–theorem. (See, for example, [3], Theorem 2.3.2.)

It will be convenient to formalize our concept of stiff well–posedness as follows.

Definition 1.1 *The Cauchy problem for the family of systems (1) is called stiffly well–posed if there are constants K, α so that*

$$\left| \exp\left(\left[\frac{1}{\delta} B + C - i \sum_{j=1}^{d} k_j A^{(j)} \right] t \right) \right| \le K e^{\alpha t} \quad \forall k \in \mathcal{R}^d,\ t \ge 0,\ \delta \in (0,1]\ . \qquad (3)$$

As we will show in Section 2, this concept is a natural generalization of the usual concept of well–posedness (in L_2) for a single system independent of a parameter δ. Also, we will show that stiff well–posedness is *equivalent* to *COND1*, which implies, in particular, that the concept is independent of the zero–order term Cu. This is important for extensions to variable coefficient problems, which we plan to present in future work. See, for example, [3] for general principles of such extensions.

Our second main result shows how one can obtain the evolution of the limit $\breve{u}(\cdot,t) = \lim_{\delta\to 0} u(\cdot,t,\delta)$ if *COND1* and *COND2* are satisfied. In fact, we will derive a so–called *equilibrium system*, which can be used to determine $\breve{u}(x,t)$. First note that *COND1* and *COND2* imply the following property of the matrix B. (Just consider *COND1* with $\omega = 0$.)

- *COND3 All eigenvalues λ of B satisfy $\mathrm{Re}\,\lambda < 0$ or $\lambda = 0$. If $\lambda = 0$ is an eigenvalue of B, then $\lambda = 0$ is semisimple, i.e., the algebraic and geometric multiplicities of $\lambda = 0$ are the same.*

Whenever *COND3* holds, there is a transformation $T \in \mathcal{C}^{n\times n}$ so that

$$T^{-1} B T = \begin{pmatrix} 0 & 0 \\ 0 & \tilde{B}_{22} \end{pmatrix} =: \tilde{B} \qquad (4)$$

where all eigenvalues of $\widetilde{B}_{22}$ have negative real part. Setting $u(x,t,\delta) = Tv(x,t,\delta)$, $f(x) = Tg(x)$, and

$$\widetilde{A}^{(j)} = T^{-1}A^{(j)}T = \begin{pmatrix} \widetilde{A}^{(j)}_{11} & \widetilde{A}^{(j)}_{12} \\ \widetilde{A}^{(j)}_{21} & \widetilde{A}^{(j)}_{22} \end{pmatrix}, \quad \widetilde{C} = T^{-1}CT = \begin{pmatrix} \widetilde{C}_{11} & \widetilde{C}_{12} \\ \widetilde{C}_{21} & \widetilde{C}_{22} \end{pmatrix}$$

one obtains the transformed system

$$v_t + \sum_{j=1}^{d} \widetilde{A}^{(j)} v_{x_j} = \frac{1}{\delta}\widetilde{B}v + \widetilde{C}v, \quad v(x,0,\delta) = g(x) \ .$$

We partition

$$v = \begin{pmatrix} v^I \\ v^{II} \end{pmatrix}$$

where the dimension of v^I is the dimension of the null–space of B. Then it is plausible that the negativity of the spectrum of $\widetilde{B}_{22}$ leads to rapid decay in time of v^{II} for small $\delta > 0$, and this suggests the limit system

$$\check{v}^I_t + \sum_{j=1}^{d} \widetilde{A}^{(j)}_{11} \check{v}^I_{x_j} = \widetilde{C}_{11}\check{v}^I \ . \tag{5}$$

We will prove:

Theorem 1.2 *Suppose the system (1) is strongly hyperbolic and COND1+2 are satisfied. Let u=Tv, f=Tg denote the transformation described above with $f \in L_2$. Then, for any $t > 0$,*

$$v(\cdot,t,\delta) \to \begin{pmatrix} \check{v}^I(\cdot,t) \\ 0 \end{pmatrix} \quad as \quad \delta \to 0 \ .$$

Here $\check{v}^I(\cdot,t)$ is the L_2–solution of (5) with initial data $\check{v}^I(x,0) = g^I(x)$, and convergence holds w.r.t. the L_2–norm. Furthermore, the system (5) is strongly hyperbolic, and for all $k \in \mathcal{R}^d$ all eigenvalues of $\sum_j k_j \widetilde{A}^{(j)}_{11}$ lie between the minimal and maximal eigenvalue of $\sum_j k_j A^{(j)}$.

The eigenvalues of $\sum_j k_j \widetilde{A}^{(j)}_{11}$ for $|k| = 1$ are the phase speeds of the equilibrium system (5) in the following sense: If

$$\left(\sum_{j=1}^{d} k_j \widetilde{A}^{(j)}_{11} \right) w^I = \lambda w^I, \quad \lambda \in \mathcal{R}, \quad w^I \neq 0 \ ,$$

then $v^I(x,t) = e^{i\langle k,x\rangle - i\lambda t} w^I$ solves (5) and describes a wave moving with speed λ in direction k. According to the above theorem, the phase speeds of the equilibrium system (5) lie between the minimal and maximal phase speeds of the system (1), corresponding to the same wave vector k, if *COND1* and *COND2* are satisfied.

Whenever *COND3* holds for the matrix B — but *COND1* is not necessarily fulfilled — a transformation T which blocks B as in (4) can be obtained, and a system (5) can be derived. Then we say that the *phase speed condition* is fulfilled if (5) is strongly hyperbolic and for all $k \in \mathcal{R}^d$ all eigenvalues of $\sum_j k_j \tilde{A}^{(j)}_{11}$ lie between the minimal and maximal eigenvalue of $\sum_j k_j A^{(j)}$. Theorem 1.2 says that the phase speed condition is a consequence of *COND1* and *COND2*. Conversely, *COND3* together with the phase speed condition generally do not imply *COND1*, as was shown by Example 4.4 in [4].

The above results all apply to general initial data f in L_2. The conditions for convergence $u(\cdot,t,\delta) \to \breve{u}(\cdot,t)$ as $\delta \to 0$ are *different*, however, if one puts more stringent assumptions on the initial function f. Let M_0 denote the space of all $f : \mathcal{R}^d \to \mathcal{C}^n$ whose Fourier transform is in C_0^∞. See Section 2 for details. The space M_0 is dense in L_2 and consists of analytic functions. Then, for arbitrary $A^{(j)}$, $B \in \mathcal{C}^{n\times n}$, i.e., without requesting strong hyperbolicity, the initial value problem (1), (2) has an analytic solution $u(x,t,\delta)$ satisfying $u(\cdot,t,\delta) \in M_0$. We write $u(\cdot,t,\delta) = S_0(t,\delta)f$ and call $S_0(t,\delta)$ the M_0–solution operator. The following convergence result will be shown.

Theorem 1.3 *Suppose COND3 holds. Then, for all $f \in M_0$ and all $t > 0$, the M_0–solution $u(\cdot,t,\delta) = S_0(t,\delta)f$ converges (w.r.t. the H_m–norm, for any m) to a limit $\breve{u}(\cdot,t)$ as $\delta \to 0$. It holds that $\breve{u}(\cdot,t) \in M_0$, and if $u = Tv$, $f = Tg$ denotes the transformation described above, then*

$$v(\cdot,t,\delta) \to \begin{pmatrix} \breve{v}^I(\cdot,t) \\ 0 \end{pmatrix} =: \breve{v}(\cdot,t) \quad \textit{as} \quad \delta \to 0 \quad \textit{for} \quad t > 0 \ .$$

Here $\breve{u} = T\breve{v}$ and the function $\breve{v}^I$ is the M_0–solution of (5) with initial data $\breve{v}^I(x,0) = g^I(x)$.

Under the assumption of Theorem 1.3 neither system (1) nor (5) is necessarily hyperbolic, and even if both systems are hyperbolic, the phase speed condition might be violated.

The paper is organized as follows: In Section 2 we review well–posedness for non–stiff systems motivating our concept of stiff well–posedness. In Section 3 an asymptotic expansion is derived and in Section 4 a model for two–phase porous medium flow is analyzed. Furthermore, we compute the first three terms in the asymptotic expansion of this example numerically. Finally, in the last section Theorems 1.1–1.3 shall be proved.

2 Well–Posedness, Source Terms and Entropy

For every fixed positive δ the system (1) has the form

$$u_t + \sum_{j=1}^{d} A^{(j)} u_{x_j} = Cu\ . \tag{6}$$

The Cauchy problem for such systems with zero–order term Cu is well understood, and the results are included here to motivate our concept of stiff well–posedness. For a general account on well–posedness (in the L_2–sense) of Cauchy problems, see [3], for example. Let us first assume that $A^{(j)}$, $C \in \mathcal{C}^{n\times n}$ are arbitrary matrices. Given initial data in

$$M_0 = \left\{ f : \mathcal{R}^d \to \mathcal{C}^n \,\middle|\, f(x) = \frac{1}{(2\pi)^{d/2}} \int\limits_{\mathcal{R}^d} e^{i\langle x,k\rangle} \hat{f}(k)\, dk,\ \hat{f} \in C_0^\infty \right\}$$

there is a unique M_0–solution of (6), namely

$$u(x,t) = \frac{1}{(2\pi)^{d/2}} \int\limits_{\mathcal{R}^d} e^{i\langle k,x\rangle} e^{P(ik)t} \hat{f}(k)\, dk\ .$$

Here

$$P(ik) = C - i \sum_{j=1}^{d} k_j A^{(j)}$$

is the symbol of the operator $C - \sum_j A^{(j)} \frac{\partial}{\partial x_j}$. The M_0–solution $u(\cdot,t) = S_0(t)f$ has the following properties:

- $u(x,t)$ is an analytic function of (x,t) and a classical solution;
- $u(\cdot,t) \in M_0 \quad \forall\, t$;
- $\hat{u}(k,t) = e^{P(ik)t}\hat{f}(k)$ is C^∞ in (k,t);
- $\hat{u}(k,t) = 0$ for $|k| > k_0$, k_0 independent of t.

2.1 Well–Posedness

The Cauchy problem for (6) is called *well–posed* (in the L_2 sense) if the M_0–solutions can be estimated by their initial data f for $t \geq 0$. More precisely, well–posedness requires the existence of constants K and α such that

$$\|u(\cdot,t)\| \leq Ke^{\alpha t}\|f\| \quad \forall\, f \in M_0,\, t \geq 0 \tag{7}$$

or equivalently if ([3], Theorem 2.2.2)

$$|e^{P(ik)t}| \leq Ke^{\alpha t} \quad \forall\, k \in \mathcal{R}^d,\ t \geq 0\,. \tag{8}$$

(Here K and α are independent of $t \geq 0$ and $k \in \mathcal{R}^d$.)

If the Cauchy problem for (6) is well–posed, then the M_0–solution operator

$$S_0(t): M_0 \to M_0, \quad f \to u(\cdot,t)$$

is a bounded linear operator (with respect to the L_2–norm) for each $t \geq 0$. Since M_0 is dense in L_2, we obtain a unique bounded linear extension $S(t): L_2 \to L_2$ of $S_0(t)$, and call $S(t), t \geq 0$, the L_2–solution operator of (6). We have

$$\|S(t)\| \leq Ke^{\alpha t}, \quad t \geq 0\ ,$$

as follows directly from (7).

It is not difficult to show that well–posedness of the Cauchy problem for (6) is independent of the zero–order term Cu, cf. [3] Lemma 2.3.5. The next theorem gives an algebraic characterization of well–posedness. The following notations are used. The principle part of the symbol $P(ik)$ is

$$P_0(ik) := -i\sum_{j=1}^{d} k_j A^{(j)}.$$

By B^* we denote the complex conjugate transpose of a matrix B, and if Q_1, Q_2 are Hermitian matrices, then we write $Q_1 \leq Q_2$ if $y^*Q_1y \leq y^*Q_2y$ for all $y \in \mathcal{C}^n$.

Theorem 2.1 *Let $A^{(j)}$, $C \in \mathcal{C}^{n\times n}$, $j = 1,2,\ldots,d$, be arbitrary. The following conditions are equivalent.*

1. *The Cauchy Problem for $u_t + \sum_j A^{(j)}u_{x_j} = Cu$ is well–posed.*
2. *For all $k \in \mathcal{R}^d$ the eigenvalues of $P_0(ik)$ are purely imaginary and semi-simple. Furthermore, there exists a transformation $S(k)$ to diagonal form, $S(k)P_0(ik)S^{-1}(k) = diag$, which is uniformly bounded, $|S(k)| + |S^{-1}(k)| \leq K$.*

3. *There exists a uniformly positive, bounded Hermitian matrix function $H(k) \in C^{n\times n}$, $K^{-1}I \le H(k) \le KI$, such that $H(k)P_0(ik)$ is anti–Hermitian, i.e., $HP_0 + P_0^* H = 0$.*

Proof: Equivalence of 1. and 2. follows from the Kreiss–matrix–theorem, see [3] Theorem 2.4.1. If 2. is assumed, then the matrix $H(k) = S(k)^* S(k)$ satisfies the conditions in 3. Conversely, if 3. is assumed let $H = S^2$, $S = S^* > 0$; then SP_0S^{-1} is anti–Hermitian, and condition 2 follows.

If the conditions of the above theorem are met, then $u_t + \sum_j A^{(j)} u_{x_j} = Cu$ is called *strongly hyperbolic*. A matrix function $H(k)$ satisfying condition 3 is called a *symmetrizer* of the system.

A necessary condition for strong hyperbolicity follows from Theorem 2.1, namely:

All eigenvalues of the matrices $A^{(j)}$, $j = 1, 2, \ldots, d$ are real and semisimple.

In the one–dimensional case strong hyperbolicity is equivalent to the well known spectral condition

All eigenvalues of the matrix A are real and semisimple.

If $H \in \mathcal{C}^{n\times n}$ is a positive definite Hermitian matrix, then

$$\langle u, v\rangle_H = u^* H v, \quad |u|_H^2 = u^* H u$$

defines an inner product and a norm on $\mathcal{C}^n$. The proof of the following elementary result is omitted.

Lemma 2.1 *Assume $c^{-1}I \le H = H^* \le cI$ with $c > 0$, i.e.,*

$$\frac{1}{c}|y|^2 \le |y|_H^2 \le c|y|^2 \quad \textit{for all} \quad y \in \mathcal{C}^n \ .$$

Then we have $|H| \le c$ and $|H^{-1}| \le c$. Furthermore, if $C \in \mathcal{C}^{n\times n}$ is arbitrary, then

$$HC + C^* H \le 2c^2 |C| H \quad \textit{and} \quad |C| \le c|C|_H \ .$$

2.2 Source Terms and Stiff Well–Posedness

Let us turn to the parameter dependent system

$$u_t + \sum_{j=1}^{d} A^{(j)} u_{x_j} = \frac{1}{\delta} B u, \quad 0 < \delta \ll 1. \tag{9}$$

We assume strong hyperbolicity, i.e., $HP_0 + P_0^* H = 0$ for a suitable symmetrizer $H = H(k)$, and consider the ODE

$$\dot{y} = P(ik,\delta) y, \quad P(ik,\delta) = P_0(ik) + \frac{1}{\delta} B = \frac{1}{\delta} B - i \sum_{j=1}^{d} k_j A^{(j)}$$

for fixed $k \in \mathcal{R}^d$. Clearly,

$$\frac{d}{dt}\langle y, Hy\rangle = \langle \dot{y}, Hy\rangle + \langle y, H\dot{y}\rangle = \frac{1}{\delta}\langle By, Hy\rangle + \frac{1}{\delta}\langle y, HBy\rangle.$$

Assuming that $c^{-1} I \le H \le cI$, we obtain from the previous lemma

$$\frac{d}{dt}\langle y, Hy\rangle \le 2\,\frac{\beta}{\delta}\,\langle y, Hy\rangle,$$

with $\beta = |B| c^2$. This implies the estimate $|e^{P(ik,\delta)t}|_H \le e^{\beta t/\delta}$, thus

$$|e^{P(ik,\delta)t}| \le c e^{\beta t/\delta}\ . \tag{10}$$

Clearly, if $0 < \delta \ll 1$, this bound becomes practically useless if $\beta > 0$. But of course, the estimate is not always sharp, and it will be our task to improve it under more specific assumptions on $A^{(j)}$ and B.

Instead of (10) a δ–independent bound like (8)

$$\left|\exp\left(\left[\frac{1}{\delta} B - i \sum_{j=1}^{d} k_j A^{(j)}\right] t\right)\right| \le K e^{\alpha t} \quad \forall\, k \in \mathcal{R}^d,\ t \ge 0,\ \delta \in (0,1] \tag{11}$$

where K and α are independent of k, t and δ will turn out to be useful. Introducing the variables $\omega = \delta k$ and $\tau = t/\delta$, inequality (11) reads

$$\left|\exp\left(\left[B - i \sum_{j=1}^{d} \omega_j A^{(j)}\right] \tau\right)\right| \le K e^{\alpha \tau \delta} \quad \forall\, \omega \in \mathcal{R}^d,\ \tau \ge 0,\ \delta \in (0,1]\,. \tag{12}$$

The requirement that (12) holds uniformly in δ is equivalent to *COND1*,

$$\left|\exp\left(\left[B - i \sum_{j=1}^{d} \omega_j A^{(j)}\right] \tau\right)\right| \le K \quad \forall\, \omega \in \mathcal{R}^d,\ \tau \ge 0\,.$$

Thus, if (11) holds for some α, then it already holds for $\alpha = 0$, and there is no exponential growth. In other words, postulating the usual well–posedness estimate

$$\|u(\cdot, t, \delta)\| \leq K e^{\alpha t} \|f\|, \quad t \geq 0, \quad \delta \in (0, 1] \tag{13}$$

uniformly in δ for the parameter dependent system (9) leads to *COND1*.

Let us now include a bounded source term Cu and consider

$$u_t + \sum_{j=1}^{d} A^{(j)} u_{x_j} = \frac{1}{\delta} Bu + Cu, \quad 0 < \delta \leq 1. \tag{14}$$

We will show that stiff well–posedness is equivalent to *COND1*. As a notation, we let

$$\mathrm{Re}\, A = \frac{1}{2}(A + A^*)$$

denote the symmetric part of $A \in \mathcal{C}^{n \times n}$.

Lemma 2.2 *Consider the family of systems (14). The existence of constants K and α with (3) is equivalent to COND1.*

Proof: As above, let $P_0 = P_0(ik) = -i \sum_{j=1}^{d} k_j A^{(j)}$.

First assume that *COND1* holds. By the Kreiss–matrix–theorem there is a bounded symmetrizer $c^{-1} I \leq H(\omega) = H^*(\omega) \leq cI$, such that

$$\mathrm{Re}(H(\omega)\,(B + P_0(i\omega))) \leq 0.$$

Now consider (with $H = H(\delta k)$)

$$\mathrm{Re}(H(\frac{1}{\delta} B + C + P_0(ik))) \leq \mathrm{Re}(HC) \leq c^2 |C| H =: \alpha H\ .$$

This implies

$$\mathrm{Re}(H(P(ik, \delta) - \alpha I)) \leq 0\ ,$$

and therefore,

$$|e^{(P(ik,\delta) - \alpha I)t}| \leq K\ .$$

Now assume conversely that there exist K and α with

$$|\exp\left((B/\delta + P_0(ik) + C)\, t\right)| \leq K e^{\alpha t}, \quad t \geq 0, \quad k \in \mathcal{R}^d, \quad \delta \in (0, 1].$$

By the Kreiss–matrix–theorem there is a bounded symmetrizer $c^{-1}I \le H \le cI$ with

$$\mathrm{Re}(H(k)\,(B/\delta + P_0(ik) + C)) \le \alpha H(k).$$

It follows that

$$\mathrm{Re}(H(k)\,(B/\delta + P_0(ik))) \le \beta H(k)$$

with a new constant β. As we have shown above, this implies *COND1*, and the lemma is proved.

A simple sufficient criterion for stiff well–posedness is given next.

Theorem 2.2 *Consider* $u_t + \sum_j A^{(j)} u_{x_j} = (1/\delta)Bu + Cu$, $0 < \delta \le 1$, *where* $A^{(j)}$, B, $C \in \mathcal{C}^{n\times n}$. *Assume the system is strongly hyperbolic,* H *is a symmetrizer and*

$$H(k)B + B^* H(k) \le 0, \quad \forall k \in \mathcal{R}^d \ .$$

Then COND1 holds, thus the system is stiffly well–posed.

Proof: Fix some $\omega \in \mathcal{R}^d$, let $P = P(i\omega) = B + P_0(i\omega) = B - i\sum_j \omega_j A^{(j)}$ and define $y(t)$ by

$$\dot{y} = Py, \quad y(0) = y^0.$$

As $HP_0 + P_0^* H = 0$, it follows

$$\begin{aligned} \frac{d}{dt}|y|_H^2 &= \langle \dot{y}, Hy\rangle \ + \ \langle y, H\dot{y}\rangle \\ &= \langle By, Hy\rangle \ + \ \langle y, HBy\rangle \\ &\le \ 0 \ . \end{aligned}$$

Therefore, $|y(t)|_H \le |y_0|_H$ and consequently $|e^{Pt}|_H \le 1$ for $t \ge 0$. As ω was arbitrary the theorem is proved.

2.3 Entropies and Symmetrizers

Entropies and symmetrizers are both well–established tools for the study of hyperbolic PDEs and also for systems of other types. In the case of a constant coefficient system, a symmetrizer is often constructed in Fourier space. Using the tools of pseudo–differential operators, the construction can then be extended to variable coefficient problems with the aim to obtain a norm in which the solution can be controlled.

In contrast, entropies are typically constructed directly in terms of physical variables. The aim is, again, to obtain a functional of the solution which can be controlled. In this section we try to relate symmetrizers and entropies in the simplest setting.

For a system in one space dimension

$$u_t + Au_x = 0 \tag{15}$$

a symmetrizer is, by definition, a matrix function $H = H(k), k \in \mathcal{R}$, with the following properties:

1) There exists $c > 0$ with

$$\frac{1}{c}I \leq H(k) = H^*(k) \leq cI, \quad k \in \mathcal{R} .$$

2)

$$\mathrm{Re}(H(k)P(ik)) = 0, \quad k \in \mathcal{R} ,$$

where $P(ik) = -ikA$. For $k \neq 0$, the condition $\mathrm{Re}(H(k)P(ik)) = 0$ is equivalent to

$$(H(k)A)^* = H(k)A . \tag{16}$$

Therefore, if no additional properties of $H(k)$ are needed, it is natural to assume that $H(k) = H_0$ does not depend on k. Let us assume this and let us assume for simplicity that A and H_0 are real. Then (16) requires *symmetry* of the product H_0A.

Recall that a smooth scalar function $\varphi : \mathcal{R}^n \to \mathcal{R}$ is called an *entropy* for the system of conservation laws

$$u_t + f(u)_x = 0, \quad x \in \mathcal{R}, \tag{17}$$

if the matrix $\varphi''(u)f'(u)$ is symmetric for all $u \in \mathcal{R}^n$. (Here $f : \mathcal{R}^n \to \mathcal{R}^n$ is a flux function with Jacobian $f'(u)$ and $\varphi''(u)$ is the Hessian of φ.) To explain the significance of the symmetry of $\varphi''(u)f'(u)$, let $\varphi : \mathcal{R}^n \to \mathcal{R}$ denote an arbitray smooth function and consider the vector field

$$V(u) = \varphi'(u)f'(u)$$

with Jacobian

$$V'(u) = \varphi''(u)f'(u) + \varphi'(u)f''(u).$$

The term $\varphi'(u)f''(u)$ is always symmetric and, consequently, the symmetry of $\varphi''(u)f'(u)$ is *equivalent* to the symmetry of $V'(u)$. Therefore, if $\varphi : \mathcal{R}^n \to \mathcal{R}$ is an entropy, then $V'(u)$ is symmetric, and the vector field $V(u)$ has a potential $\psi(u)$,

$$\varphi'(u)f'(u) = V(u) = \psi'(u).$$

It is easy to show, that the function $\psi(u)$, called an entropy flux, has the following property: If $u(x,t)$ is a smooth solution of (17), then

$$\varphi(u)_t + \psi(u)_x = 0.$$

Under appropriate assumptions, this fundamental relation is then used to obtain estimates for $\varphi(u(x,t))$ and then for u.

Now consider the simple case where

$$f(u) = Au, \quad u \in \mathcal{R}^n,$$

is linear and

$$\varphi(u) = \frac{1}{2}u^T H_0 u, \quad u \in \mathcal{R}^n, \tag{18}$$

is quadratic. (We assume here that A and $H_0 = H_0^T$ are real.) Then

$$\varphi''(u)f'(u) = H_0 A,$$

i.e., $\varphi(u)$ is an entropy if and only if $H_0 A$ is symmetric. Clearly, if $H_0 = H_0^T$ is a positive definite matrix (as is required for a symmetrizer), then $\varphi(u)$ is a positive definite quadratic form, and in particular $\varphi(u)$ is strictly convex. If (18) is an entropy, then

$$\psi(u) = \frac{1}{2}u^T H_0 A u$$

is the corresponding entropy flux.

Now consider

$$u_t + Au_x = \frac{1}{\delta}Bu, \quad 0 < \delta \le 1\ . \tag{19}$$

It is assumed that the term $\frac{1}{\delta}Bu$ "models viscosity" in some sense, which is not made precise here. (For a discussion see [1] and the references therein.) If we assume, as above,

$$H_0 = H_0^T, \quad H_0 A = (H_0 A)^T, \quad \varphi(u) = \frac{1}{2}u^T H_0 u, \quad \psi(u) = \frac{1}{2}u^T H_0 A u,$$

then we obtain

$$\varphi(u)_t + \psi(u)_x = u^T H_0 u_t + u^T H_0 A u_x = \frac{1}{\delta}u^T H_0 B u$$

for any smooth solution $u(x,t)$ of (19). If we make the requirement

$$H_0 B + B^* H_0 \le 0\ , \tag{20}$$

which yields stiff well–posedness by Theorem 2.2, we obtain the entropy inequality

$$\varphi(u)_t + \psi(u)_x \le 0$$

for all smooth solutions of (19). Thus (20) is consistent with an entropy inequality.

In future work we want to investigate the relations between the following three statements, where A, $B \in \mathcal{R}^{n\times n}$ and the eigenvalues of A are real and semisimple. (In fact, one may assume that A is real and diagonal.)

Statement 1 *There exists a symmetric, positive definite H_0 so that $H_0 A$ is symmetric and $H_0 B + B^T H_0 \le 0$.*

Statement 2 *The Cauchy problem for the family of strongly hyperbolic systems (19) is stiffly well–posed.*

Statement 3 *There exists a strictly convex function $\varphi : \mathcal{R}^n \to \mathcal{R}$ (an entropy) and a smooth function $\psi : \mathcal{R}^n \to \mathcal{R}$ (an entropy flux) with $\varphi''(u)A = (\varphi''(u)A)^T$, $\psi'(u) = \varphi'(u)A$, and $\varphi(u)_t + \psi(u)_x \le 0$ for all smooth solutions of (19).*

Our discussion has shown that Statement 1 implies Statements 2 and 3. Furthermore, by the Kreiss–matrix–theorem, Statement 2 is equivalent to

Statement 2* *There is $c > 0$ and for each $\omega \in \mathcal{R}$ there is a Hermitean matrix $H(\omega)$ with $c^{-1} I \le H(\omega) \le cI$ and $\mathrm{Re}\,[H(\omega)(B - i\omega A)] \le 0$.*

If Statement 1 is true, we may choose $H(\omega) = H_0$ in Statement 2*. We want to investigate if there are systems (19) satisfying Statement 2, but not Statement 1.

3 Asymptotic Expansion

In this section we derive an asymptotic expansion of the solution of (1), (2) assuming that the system is stiffly well–posed and *COND2* holds. For simplicity of presentation we omit the bounded source term Cu and restrict ourselves to the case of one space variable, but will comment below on the multidimensional case. Without loss of generality we assume that the matrix B is already in block form. Thus we consider a system

$$\begin{pmatrix} u \\ v \end{pmatrix}_t + \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix}_x = \frac{1}{\delta} \begin{pmatrix} 0 & 0 \\ 0 & B_{22} \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} + \begin{pmatrix} F \\ G \end{pmatrix} \tag{21}$$

with initial condition

$$\begin{pmatrix} u \\ v \end{pmatrix}(x, 0, \delta) = \begin{pmatrix} f \\ g \end{pmatrix}(x), \quad x \in \mathcal{R}, \quad \delta \in (0, 1]. \tag{22}$$

Here

$$A = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 \\ 0 & B_{22} \end{pmatrix}$$

are constant complex $n \times n$ matrices. We have introduced forcing terms into the system, because even if we start out without forcing, forced equations like (21) will appear in the process of the asymptotic expansion. As this section is essentially self contained, the obvious change of notation should not cause any problem.

The assumption of stiff well–posedness requires:

(a) There is a constant $c_1 > 0$ such that $|\exp(B - i\omega A)\tau)| \le c_1$, for all $\omega \in \mathcal{R}$ and $\tau \ge 0$.

Furthermore, *COND2* reads:

(b) $\operatorname{Re}\lambda \le -\gamma < 0$ for all $\lambda \in \sigma(B_{22})$.

Let $T > 0$ be some fixed time. For simplicity of presentation we will make strong smoothness assumptions for the data $h(x) = (f,g)^T(x)$, $H(x,t) = (F,G)^T(x,t)$ and require

$$h \in C_0^\infty(\mathcal{R}); \quad H \in C^\infty(\mathcal{R} \times [0,T])$$

and there is a $\kappa \ge 0$ such that

$$H(x,t) = 0 \quad \text{if} \quad |x| \ge \kappa, \quad 0 \le t \le T.$$

Thus, the data are C^∞–smooth and have a compact support in x, which is uniform for $0 \le t \le T$. Then, because of the finite speed of propagation in hyperbolic systems, which is independent of δ, the solution and all approximating functions appearing below are also C^∞–smooth and compactly supported in some common x–region. (One can relax the smoothness requirement for the data and the assumption of a compact support by counting derivatives and requiring sufficiently strong decay as $|x| \to \infty$.)

For sufficiently regular vector functions

$$\varphi : \mathcal{R} \times [0,T] \to \mathcal{R}^k$$

we use the notations

$$M_j(\varphi) = \max_{0 \le t \le T} \|\partial_x^j \varphi(\cdot,t)\|, \quad I_j(\varphi) = \int_0^T \|\partial_x^j \varphi(\cdot,t)\|\, dt, \quad j = 0,1,\ldots$$

Here $\partial_x^j = \partial^j/\partial x^j$. (Thus, M_j takes a maximum over $0 \le t \le T$, whereas I_j takes an integral. We always use L_2–norms over space.) Setting $w = (u, v)^T$ the initial value problem (21), (22) reads

$$w_t + Aw_x = \frac{1}{\delta} Bw + H(x,t), \quad w(x,0,\delta) = h(x). \tag{23}$$

As above, $S(t,\delta)$ denotes the solution operator of $w_t + Aw_x = 1/\delta Bw$. Then by Duhamel's principle it follows that

$$w(\cdot,t,\delta) = S(t,\delta)h + \int_0^t S(t-\xi,\delta)H(\cdot,\xi)\, d\xi. \tag{24}$$

Assumption (a) yields

$$\|S(t,\delta)\| \le c_1, \quad \forall t \ge 0,$$

and therefore

$$\|w(\cdot,t,\delta)\| \le c_1\|h\| + c_1 \int_0^t \|H(\cdot,\xi)\|\, d\xi,$$

thus

$$\max_{0\le t\le T} \|w(\cdot,t,\delta)\| \le c_1 \left\{\|h\| + I_0(H)\right\}.$$

If we differentiate (23) j–times with respect to x, we obtain

$$\max_{0\le t\le T} \|\partial_x^j w(\cdot,t,\delta)\| \le c_1 \left\{\|\partial_x^j h\| + I_j(H)\right\}.$$

Lemma 3.1 *Under the smoothness assumption stated above, the solution of the forced stiff system (23) is bounded in terms of the data as follows:*

$$M_j(w(\cdot,\cdot,\delta)) \le c_1 \left\{\|\partial_x^j h\| + I_j(H)\right\}, \quad \delta \in (0,1], \quad j \in \mathcal{N}_0.$$

After these preliminary considerations, we derive an asymptotic expansion.

3.1 Recursive Definition of the Expansion Terms

Because of assumption (b) we expect v to decay like $e^{-\gamma\tau}$, where $\tau = t/\delta$ is the fast time variable. Therefore the following ansatz is reasonable

$$w(x,t,\delta) = \sum_{l=0}^m \delta^l w_l(x,t) + \sum_{l=0}^m \delta^l W_l(x,t/\delta) + \Delta w^{(m)}(x,t,\delta). \tag{25}$$

Here $\sum_l \delta^l w_l$ is a so–called outer approximation, whereas the inner approximation $\sum_l \delta^l W_l$ describes the initial layer. We shall see below that the error $\Delta w^{(m)}$ is of order δ^{m+1} in maximum norm.

Substituting the ansatz (25) into the Cauchy problem (21), (22) and collecting the coefficients of equal powers of δ one obtains defining equations for the expansion terms $w_l = (u_l, v_l)^T$ and $W_l = (U_l, V_l)^T$.

The leading order terms are determined as follows.

a-0) $$U_0 = 0.$$

b-0) $$v_0 = 0.$$

c-0) $$\partial_t u_0 + A_{11}\partial_x u_0 = F, \quad u_0(x,0) = f(x).$$

d-0) $$\partial_\tau V_0 = B_{22}V_0, \quad V_0(x,0) = g(x).$$

Thus u_0 is given by the equilibrium system which is strongly hyperbolic by Theorem 1.2. From Lemma 3.1 it follows that

$$M_j(u_0) \leq K\left\{\|\partial_x^j f\| + I_j(F)\right\} \leq C.$$

Here and throughout this section C denotes a generic constant, which may depend on the data f, g, F, G and its derivatives, but not on δ. Using the PDE, time derivatives are expressed as space derivatives, hence

$$M_j(\partial_t^k u_0) \leq M_{j+k}(u_0) + M_j(\partial_t^k F) \leq C.$$

Because of (b), the initial layer $V_0(x,\tau) = e^{B_{22}\tau}g(x)$ is bounded by

$$\|\partial_x^j V_0(\cdot,\tau)\| \leq C\|\partial_x^j g\|e^{-\gamma\tau} \leq Ce^{-\gamma\tau}.$$

Using the ODE $\partial_\tau V_0 = B_{22}V_0$ it follows that

$$\|\partial_x^j \partial_\tau^k V_0(\cdot,\tau)\| \leq Ce^{-\gamma\tau}.$$

The first–order correction is determined as follows.

a-1) $$\partial_\tau U_1 = -A_{12}\partial_x V_0, \quad U_1(x,\infty) = 0.$$

b-1) $$B_{22}v_1 = A_{21}\partial_x u_0 - G.$$

c-1) $$\partial_t u_1 + A_{11}\partial_x u_1 = -A_{12}\partial_x v_1, \quad u_1(x,0) = -U_1(x,0).$$

d-1) $$\partial_\tau V_1 = B_{22}V_1 - A_{22}\partial_x V_0, \quad V_1(x,0) = -v_1(x,0).$$

Integrating backwards in time we find

$$U_1(x,\tau) = \int_\tau^\infty A_{12}\partial_x V_0 \; d\xi = A_{12}B_{22}e^{B_{22}\tau}g_x(x).$$

Note that this function 'generates' initial data $U_1(x,0) \neq 0$, but it decays exponentially

$$\|\partial_x^j U_1(\cdot,\tau)\| \leq Ce^{-\gamma\tau}.$$

Again, τ–derivatives of U_1 are given in terms of space derivatives of V_0, which are bounded

$$\|\partial_x^j\partial_\tau^k U_1(\cdot,\tau)\| \leq \|\partial_x^{j+1}\partial_\tau^{k-1}V_0(\cdot,\tau)\| \leq Ce^{-\gamma\tau}.$$

The next outer expansion term v_1 is explicitly given in terms of u_0 and G, hence

$$M_j(\partial_t^k v_1) \leq C\left\{M_{j+1}(\partial_t^{k-1}u_0) + M_j(\partial_t^k G)\right\} \leq C.$$

Again, this term generates initial data $v_1(x,0) \neq 0$. The term u_1 of the outer expansion eliminates the data created by U_1 and is bounded by

$$M_j(u_1) \leq K\left\{\|\partial_x^j U_1(x,0)\| + I_{j+1}(v_1)\right\}.$$

Using the bounds for U_1 (at $t = \tau = 0$) and v_1 it follows that

$$M_j(u_1) \leq C.$$

Again time derivatives are expressed as space derivatives of u_1 and v_1, hence

$$M_j(\partial_t^k u_1) \leq C\left\{M_{j+k}(u_1) + M_{j+k}(v_1)\right\} \leq C.$$

Finally, V_1 compensates the data $v_1(x,0)$. It is given by Duhamel's principle

$$V_1(x,\tau) = -e^{B_{22}\tau}v_1(x,0) - \int_0^\tau e^{B_{22}(\tau-\xi)}A_{22}e^{B_{22}\xi}g_x(x)\; d\xi.$$

By assumption (b) it follows that

$$\|\partial_x^j V_1(\cdot,\tau)\| \leq C\left\{e^{-\gamma\tau}\|\partial_x^j v_1(\cdot,0)\| + \int_0^\tau e^{-\gamma\tau}\; d\xi\|\partial_x^{j+1}g\|\right\} \leq C(1+\tau)e^{-\gamma\tau}.$$

Time derivatives and mixed derivatives are bounded in the same way

$$\|\partial_\tau^k V_1(\cdot,\tau)\| \leq C\left\{\|V_1(\cdot,\tau)\| + \|\partial_\tau^{k-1}\partial_x V_0(\cdot,\tau)\|\right\} \leq C(1+\tau)e^{-\gamma\tau},$$

$$\|\partial_x^j\partial_\tau^k V_1(\cdot,\tau)\| \leq C(1+\tau)e^{-\gamma\tau}.$$

Having determined w_{m-1} and W_{m-1}, the next terms are given by

a-m) $$\partial_\tau U_m = -A_{11}\partial_x U_{m-1} - A_{12}\partial_x V_{m-1}, \quad U_m(x,\infty) = 0.$$

b-m) $$B_{22} v_m = A_{21}\partial_x u_{m-1} + A_{22}\partial_x v_{m-1} + \partial_t v_{m-1}.$$

c-m) $$\partial_t u_m + A_{11}\partial_x u_m = -A_{12}\partial_x v_m, \quad u_m(x,0) = -U_m(x,0).$$

d-m) $$\partial_\tau V_m = B_{22} V_m - A_{21}\partial_x U_{m-1} - A_{22}\partial_x V_{m-1}, \quad V_m(x,0) = -v_m(x,0).$$

Corresponding bounds are formulated in the next lemma.

Lemma 3.2 *Assume the data is smooth enough (see above), the Cauchy problem for the blocked system (21) is stiffly well–posed and* $\operatorname{Re}\sigma(B_{22}) \le -\gamma < 0$. *Let* $(u_m, v_m)^T$ *and* $(U_m, V_m)^T$ *be recursively be defined by a-m)–d-m). Then the following holds*

i) $\|\partial_x^j \partial_\tau^k U_m(\cdot,\tau)\| \le C \sum_{l=0}^{m-1} \tau^l e^{-\gamma\tau}$.

ii) $M_j(\partial_t^k v_m) \le C$.

iii) $M_j(\partial_t^k u_m) \le C$.

iv) $\|\partial_x^j \partial_\tau^k V_m(\cdot,\tau)\| \le C \sum_{l=0}^{m} \tau^l e^{-\gamma\tau}$.

Here the constant C *depends on the data and its derivatives up to some finite order, determined by* m.

Proof: Assume the statements hold for some $m \ge 1$. The task is to prove it for $m+1$.

Consider U_{m+1} first. From the ODE for U_{m+1} it follows that

$$U_{m+1}(x,\tau) = \int_\tau^\infty A_{11}\partial_x U_m(x,\xi) + A_{12}\partial V_m(x,\xi)\, d\xi$$

and hence

$$\partial_x^j U_{m+1}(x,\tau) = \int_\tau^\infty A_{11}\partial_x^j\partial_x U_m(x,\xi) + A_{12}\partial_x^j\partial_x V_m(x,\xi)\, d\xi.$$

Using the assumption it follows that

$$\|\partial_x^j U_{m+1}(\cdot,\tau)\| \le C \int_\tau^\infty \sum_{l=0}^{m} \xi^l e^{-\gamma\xi}\, d\xi.$$

Integration by parts shows that

$$\|\partial_x^j U_{m+1}(\cdot,\tau)\| \le C\sum_{l=0}^{m} \tau^l e^{-\gamma\tau}.$$

This is statement *i)*, with $k = 0$. If $k \ge 1$, then the same bound follows directly from the ODE and the assumption.

Next we consider v_{m+1} which is explicitly given by u_m and v_m. Therefore, the bound *ii)* follows directly from the assumption.

The term u_{m+1} solves the equilibrium system with a modified forcing, hence Lemma 3.1 applies and yields that

$$M_j(u_{m+1}) \le C.$$

Again, by the PDE time derivatives are expressed as space derivatives and *iii)* holds for u_{m+1}.

It remains to bound V_{m+1}. Using Duhamel's principle it holds that

$$V_{m+1}(x,\tau) = -e^{B_{22}\tau}v_m(x,0) - \int_0^T e^{B_{22}(\tau-\xi)}\left(A_{21}\partial_x U_m(x,\xi) + A_{22}\partial_x V_m(x,\xi)\right)\,d\xi.$$

By (b) and the induction assumption it follows that

$$\|V_{m+1}(\cdot,\tau)\| \le Ce^{-\gamma\tau} + C\int_0^\tau \sum_{l=0}^{m} \xi^l\,d\xi e^{-\gamma\tau} \le C\sum_{l=0}^{m+1} \tau^l e^{-\gamma\tau}.$$

The corresponding bound for mixed space and time derivatives follows in the same way by applying the derivatives to V_{m+1}.

This concludes the proof of the lemma.

Remark: Note that the construction of the asymptotic expansion works in any number of space dimensions $x \in \mathcal{R}^d$:

- $U_m(x,\tau)$ is determined by integration from $\tau = \infty$ to $\tau = 0$.
- $v_m(x,t)$ is determined by inverting B_{22}.
- $u_m(x,t)$ is determined by an inhomogeneous equilibrium system.
- $V_m(x,\tau)$ is determined by an ODE initial value problem.

3.2 The Error–Estimate

Since we are dealing with linear problems here, governing equations for the approximation error $\Delta w^{(m)} = w - w^{(m)}$ are obtained by subtracting the equation for $w^{(m)}$ from the original system (21). Equations for the m–th order approximation $w^{(m)} = \sum_{l=0}^{m} \delta^l (w_l + W_l)$ are obtained as follows.

Consider the inner parts U_l first. By definition (see a-m))

$$\sum_{l=0}^{m} \delta^l \partial_\tau U_{l+1} + A_{11} \sum_{l=0}^{m} \delta^l \partial_x U_l + A_{12} \sum_{l=0}^{m} \delta^l \partial_x V_l = 0.$$

Note that $\partial_\tau = \delta \partial_t$ and $U_0 = 0$, therefore

$$\sum_{l=0}^{m+1} \delta^l \partial_t U_l + A_{11} \sum_{l=0}^{m} \delta^l \partial_x U_l + A_{12} \sum_{l=0}^{m} \delta^l \partial_x V_l = 0.$$

Also by construction (see c-m))

$$\sum_{l=0}^{m} \delta^l \partial_t u_l + A_{11} \sum_{l=0}^{m} \delta^l \partial_x u_l + A_{12} \sum_{l=0}^{m} \delta^l \partial_x v_l = F(x,t),$$

$$\sum_{l=0}^{m} \delta^l u_l(x,0) = f(x) - \sum_{l=0}^{m} \delta^l U_l(x,0).$$

Therefore, for the sum $u^{(m)}$ it holds that

$$u_t^{(m)} + A_{11} u_x^{(m)} + A_{12} v_x^{(m)} = F(x,t) - \delta^{m+1} \partial_t U_{m+1}(x, t/\delta),$$

$$u^{(m)}(x,0,\delta) = f(x).$$

For the outer part v_l it holds by construction (c.f. b-m))

$$\sum_{l=0}^{m} \delta^l \partial_t v_l + A_{21} \sum_{l=0}^{m} \delta^l \partial_x u_l + A_{22} \sum_{l=0}^{m} \delta^l \partial_x v_l = \frac{1}{\delta} B_{22} \sum_{l=0}^{m+1} \delta^l v_l + G(x,t).$$

Summing up the equations d-m) for the inner parts V_l, we find

$$\sum_{l=0}^{m} \delta^l \partial_\tau V_{l+1} + A_{21} \sum_{l=0}^{m} \delta^l \partial_x U_l + A_{22} \sum_{l=0}^{m} \delta^l \partial_x V_l = \frac{1}{\delta} B_{22} \sum_{l=0}^{m} \delta^{l+1} V_{l+1}.$$

With $\partial_\tau = \delta \partial_t$ and d-0) it follows that

$$\sum_{l=0}^{m+1} \delta^l \partial_t V_l + A_{21} \sum_{l=0}^{m} \delta^l \partial_x U_l + A_{22} \sum_{l=0}^{m} \delta^l \partial_x V_l = \frac{1}{\delta} B_{22} \sum_{l=0}^{m+1} \delta^l V_l,$$

$$\sum_{l=0}^{m} \delta^l V_l(x,0) = g(x) - \sum_{l=0}^{m} \delta^l v_l(x,0).$$

Therefore,

$$\begin{aligned} v_t^{(m)} + A_{21}u_x^{(m)} + A_{22}v_x^{(m)} &= \frac{1}{\delta}B_{22}v^{(m)} + G(x,t) \\ &\quad + \delta^m B_{22}(v_{m+1} + V_{m+1}) - \delta^{m+1}\partial_t V_{m+1}(x, t/\delta), \end{aligned}$$

$$v^{(m)}(x,0,\delta) = g(x).$$

As the systems are linear, the equations for the error $\Delta w^{(m)} = w - w^{(m)}$ are:

$$\begin{aligned} \Delta u_t^{(m)} + A_{11}\Delta u_x^{(m)} + A_{12}\Delta v_x^{(m)} &= \delta^{m+1}\partial_t U_{m+1}, \\ \Delta v_t^{(m)} + A_{21}\Delta u_x^{(m)} + A_{22}\Delta v_x^{(m)} &= \frac{1}{\delta}B_{22}\Delta v^{(m)} \\ &+ \delta^{m+1}\partial_t V_{m+1} - \delta^m B_{22}(v_{m+1} + V_{m+1}), \end{aligned}$$

$$\Delta u^{(m)}(x,0,\delta) = \Delta v^{(m)}(x,0,\delta) = 0.$$

At this stage an application of Lemma 3.1 would yield an error estimate of order δ^m, because of the forcing $\delta^m B_{22}(v_{m+1} + V_{m+1})$. Nevertheless, the following bound holds:

Theorem 3.1 *Under the assumptions stated in Lemma 3.2 the error $\Delta w^{(m)}$ in the asymptotic expansion (25) is bounded by*

$$\max_{0 \le t \le T} \sup_{x \in \mathcal{R}} |\Delta w^{(m)}(\cdot,\cdot,\delta)| \le C\delta^{m+1}.$$

Again, the constant C depends on m, but not on δ.

Proof: Let $\Delta \bar{v} = \delta^{m+1}(v_{m+1} + V_{m+1})$. Clearly, by Lemma 3.2

$$M_j(\Delta \bar{v}(\cdot,\cdot,\delta)) \le C\delta^{m+1}.$$

Now we study $\Delta \hat{v}$ given by $\Delta v^{(m)} = \Delta \hat{v} + \Delta \bar{v}$. The system vor $\Delta u^{(m)}$ and $\Delta \hat{v}$ reads:

$$\begin{aligned} \Delta u_t^{(m)} + A_{11}\Delta u_x^{(m)} + A_{12}\Delta \hat{v}_x &= \delta^{m+1}\partial_t U_{m+1} - A_{12}\Delta \bar{v}_x, \\ \Delta \hat{v}_t + A_{21}\Delta u_x^{(m)} + A_{22}\Delta \hat{v}_x &= \frac{1}{\delta}B_{22}\Delta \hat{v} - \delta^{m+1}\partial_t v_{m+1} - A_{22}\Delta \bar{v}_x, \end{aligned}$$

$$\Delta u^{(m)}(x,0,\delta) = \Delta\hat{v}(x,0,\delta) = 0.$$

Lemma 3.1 applied to this system yields

$$M_j\left(\left(\begin{array}{c}\Delta u^{(m)}\\ \Delta\hat{v}\end{array}\right)\right) \le C\left\{\delta^{m+1} I_j\left(\partial_t\left(\begin{array}{c}U_{m+1}\\ v_{m+1}\end{array}\right)\right) + I_j(\Delta\bar{v}_x)\right\}.$$

Again, by Lemma 3.2, $I_j(\Delta\bar{v}_x)$ is of the desired order

$$I_j(\Delta\bar{v}_x) \le I_{j+1}(\Delta\bar{v}) \le C\delta^{m+1}$$

and

$$I_j(\partial_t v_{m+1}) \le T M_j(\partial_t v_{m+1}) \le C.$$

Finally,

$$\begin{aligned} I_j(\partial_t U_{m+1}) &= \int_0^T \|\partial_x^j \partial_t U_{m+1}(\cdot, t/\delta)\|\, dt \\ &\le C/\delta \int_0^T \sum_{l=0}^m \left(\frac{t}{\delta}\right)^l e^{-\gamma t/\delta}\, dt \\ &\le C\int_0^{T/\delta} \sum_{l=0}^m \tau^l e^{-\gamma\tau}\, d\tau \le C. \end{aligned}$$

Now it is obvious that

$$M_j(\Delta w^{(m)}(\cdot,\cdot,\delta)) \le C\delta^{m+1}.$$

Since this bound holds for any $j = 0, 1, 2, \ldots$, Sobolev's inequality applies, the pointwise bound follows and the theorem is proved.

Remark: The theorem and the proof directly generalize to any number of space dimensions and the introduction of a δ–independent source term causes no problems.

4 An Example

In this section we study a linearized version of the following nonlinear system

$$\begin{aligned} s_t + f(s,c)_x &= 0, \\ (cs + a)_t + (cf(s,c))_x &= 0, \\ \delta a_t &= E(c) - a, \quad 0 < \delta \ll 1. \end{aligned} \tag{26}$$

This system is considered as a model for two–phase flow in a porous medium. (For example, when oil is displaced in the sea ground by a mixture of water and dissolved polymer.) In this context s is the saturation of the aqueous phase. It is assumed that the void volume is filled with fluid, hence $1-s$ is the saturation of the oleic phase. The concentration of the dissolved polymer in the aqueous phase is approximated by c, and a models the adsorption of the polymer in the sea ground. The adsorption reaction is fast compared to the characteristic speeds and is therefore represented by a stiff rate equation in the model. The equilibrium state of the adsorption process is given by $a = E(c)$, where E is a given smooth and increasing function. Finally, the fractional flow function f is a smooth, given function. Typically, when c is fixed, $f(\cdot, c)$ is an S–shaped function of s. This S–shape affects the hyperbolic character of the system and causes problems for the well–posedness of the initial value problem (cf. [2] and [6]).

To apply the theory of stiff well–posedness, we linearize the system at a constant state $(\bar{s}, \bar{c}, \bar{a})$ with $\bar{a} = E(\bar{c})$. Then the triple $(\bar{s}, \bar{c}, \bar{a})$ is a solution of system (26). We substitute the ansatz

$$s = \bar{s} + \sigma, \quad c = \bar{c} + \gamma, \quad a = \bar{a} + \alpha$$

into the system and neglect second order terms. The result is a linear constant coefficient system for the deviation $u = (\sigma, \gamma, \alpha)^T$. To simplify the notation, superscript bars shall be omitted. We write s for $\bar{s}$ and let $f = f(\bar{s}, \bar{c})$, $f_s = f_s(\bar{s}, \bar{c})$, $f_c = f_c(\bar{s}, \bar{c})$ and $e = E'(\bar{c})$. Then the linearized system reads

$$u_t + \begin{pmatrix} f_s & f_c & 0 \\ 0 & f/s & 0 \\ 0 & 0 & 0 \end{pmatrix} u_x = \frac{1}{\delta} \begin{pmatrix} 0 & 0 & 0 \\ 0 & -e/s & s^{-1} \\ 0 & e & -1 \end{pmatrix} u. \tag{27}$$

If $f_s \neq f/s$, this system is strongly hyperbolic. The spectrum of B is $\sigma(B) = \{0, 0, \lambda_3 = -1 - e/s\}$, and *COND3* is fulfilled. The system in canonical block form is

$$v_t + C v_x = \frac{1}{\delta} D v, \tag{28}$$

where

$$C := \begin{pmatrix} f_s & f_c & -f_c \\ 0 & a & -a \\ 0 & -ae/s & ae/s \end{pmatrix}, \quad a := \frac{f}{e+s}$$

and $D :=$diag$(0, 0, \lambda_3)$. The equilibrium system is

$$\breve{v}_t^I + \begin{pmatrix} f_s & f_c \\ 0 & a \end{pmatrix} \breve{v}_x^I = 0. \tag{29}$$

Obviously, critical situations can occur when either $f = s f_s$ or $f = (e+s) f_s$, because then the system (27) or (29) is not strongly hyperbolic, respectively.

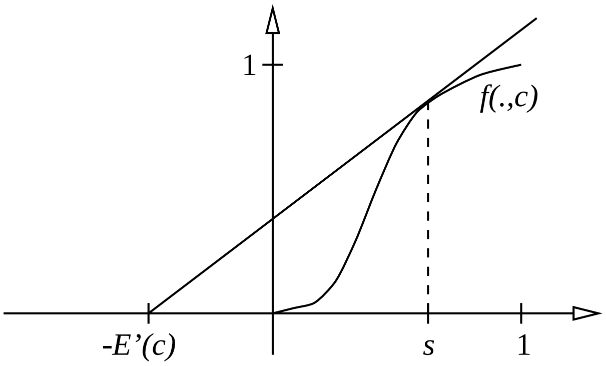

Figure 1: A typical fractional flow function f

Due to the S–shape of $f(\cdot, c)$, for any given c a critical value for s exists, see Figure 1. The next lemma gives precise conditions for the stiff well–posedness of system (27).

Lemma 4.1 *Consider the stiff hyperbolic system (27) where all coefficients are real and*

$$s > 0, \quad f > 0, \quad f_s > 0, \quad e > 0, \quad 0 < \delta \ll 1.$$

The well–posedness and stiff well–posedness of this system is determined by the parameters as follows.

1. *If $f_c = 0$, the Cauchy problem is stiffly well–posed.*
2. *If $f/s \neq f_s \neq a := f/(e+s)$, the Cauchy problem is stiffly well–posed.*
3. *If $f_c \neq 0$, $f_s \neq f/s$ but $f_s = a$, the Cauchy problem is well–posed for each δ, but not stiffly well–posed.*
4. *If $f_c \neq 0$ and $f_s = f/s$, the system is not strongly hyperbolic. Therefore, the Cauchy problem is not well–posed and, in particular, not stiffly well–posed.*

Before we prove the lemma, let us report on a numerical experiment illustrating the asymptotic expansion in case of the present example. For the experiment we have chosen the parameters $f = 1/3$, $s = 1/3$, $f_s = 1/3$, $f_c = 1$, $e = 1/3$ and $\delta = 1/10$. Then, by the lemma, the system is stiffly well–posed. We have computed the solution and the first–, second– and third–order approximations of the Cauchy problem for the blocked system (28). The initial data is $v(x, 0, \delta) = (\sin(\pi x), \cos(\pi x), 0)^T$. Since the data is in the null–space of the relaxation term D, we expect only a small initial layer (with initial amplitude of order δ). The computations are performed in the space interval $[0, 2]$. At $x = 0$ and

$x = 2$ periodic boundary conditions where imposed. The PDEs where discretized by first–order upwind differences for the convective terms and a trapezoidal rule for the time integration of the forcing. The space derivative in the forcing terms was approximated by central differences. The meshsizes in space and time where uniform, $\Delta x = \Delta t = 1/400$.

Figure 4 shows the first component of the solution, while Figures 5, 6, and 7 show the leading–, first–, and second–order errors respectively. We clearly observe that the m–th order errors are of order δ^{m+1}, as predicted by the theory.

Proof of Lemma 4.1 Consider the system in canonical block form (28).

Case 1. By the assumption $f_c = 0$, the matrix $H = \mathrm{diag}(1, e/s, 1)$ is a symmetrizer for

$$C = \begin{pmatrix} f_s & 0 & 0 \\ 0 & a & -a \\ 0 & -ae/s & ae/s \end{pmatrix}.$$

Since $D = HD$ and $D + D^* = 2D \leq 0$, *COND1* follows from Theorem 2.2.

In **Case 3** the full system (27) is strongly hyperbolic, but the equilibrium system (29) is not strongly hyperbolic. By Theorem 1.2, *COND1* does not hold.

Under the conditions of **Case 4** system (27) is not strongly hyperbolic. In this situation $\exp[(B/\delta - ikA)t]$ is not uniformly bounded for $k \in \mathcal{R}$, $t \geq 0$, and again *COND1* does not hold.

The assumptions $f/s \neq f_s \neq a$ in **Case 2** ensure that both, the full system as well as the equilibrium system, are strongly hyperbolic. To bound the matrix exponential of

$$P(\omega) = D - i\omega C = \begin{pmatrix} -i\omega f_s & -i\omega f_c & i\omega f_c \\ 0 & -i\omega a & i\omega a \\ 0 & i\omega ae/s & \lambda_3 - i\omega ae/s \end{pmatrix}$$

a transformation T will be constructed, such that $|T| + |T^{-1}| \leq const.$ and

$$T^{-1}PT = \begin{pmatrix} -i\omega f_s & -i\omega f_c(1-\varrho) & 0 \\ 0 & -i\omega a(1-\varrho) & 0 \\ 0 & 0 & \lambda_3 - i\omega a(e/s + \varrho) \end{pmatrix}. \tag{30}$$

Here ϱ, a root of the quadratic equation (35), is defined by

$$\varrho(\omega) := \begin{cases} \dfrac{\lambda_* + \sqrt{\lambda_*^2 - 4\omega^2 a^2 e/s}}{2i\omega a}, & \omega \neq 0 \\ 0, & \omega = 0 \end{cases}, \quad \lambda_* := \lambda_3 - i\omega a\,(e/s - 1)\,.$$

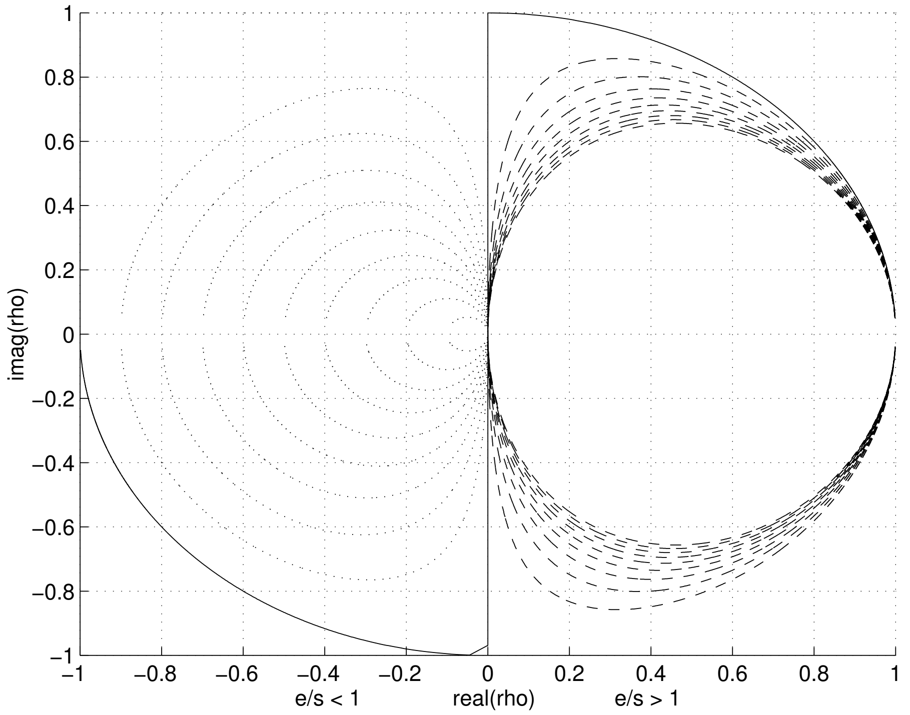

Figure 2: ϱ as function of ω for various values of e/s — dotted lines $e/s < 1$, solid $e/s = 1$, dashed $e/s > 1$.

The following properties of ϱ are obvious from Figure 2.

$$|\varrho(\omega)| \leq const, \quad \omega \in \mathcal{R} \tag{31}$$

and

$$\lim_{\omega \to \pm\infty} \varrho = \begin{cases} -e/s & e/s \in (0,1) \\ \pm 1 & e/s = 1 \\ 1 & e/s > 1 \end{cases} . \tag{32}$$

Furthermore, Figure 3 shows

$$\lambda_3 \leq 2\,\mathrm{Re}(i\omega a\varrho) = \lambda_3 + \mathrm{Re}\left(\sqrt{\lambda_*^2 - 4\omega^2 a^2 e/s}\right) \leq 0 \tag{33}$$

and

$$\mathrm{Re}(i\omega a\varrho) < 0, \quad \omega \neq 0. \tag{34}$$

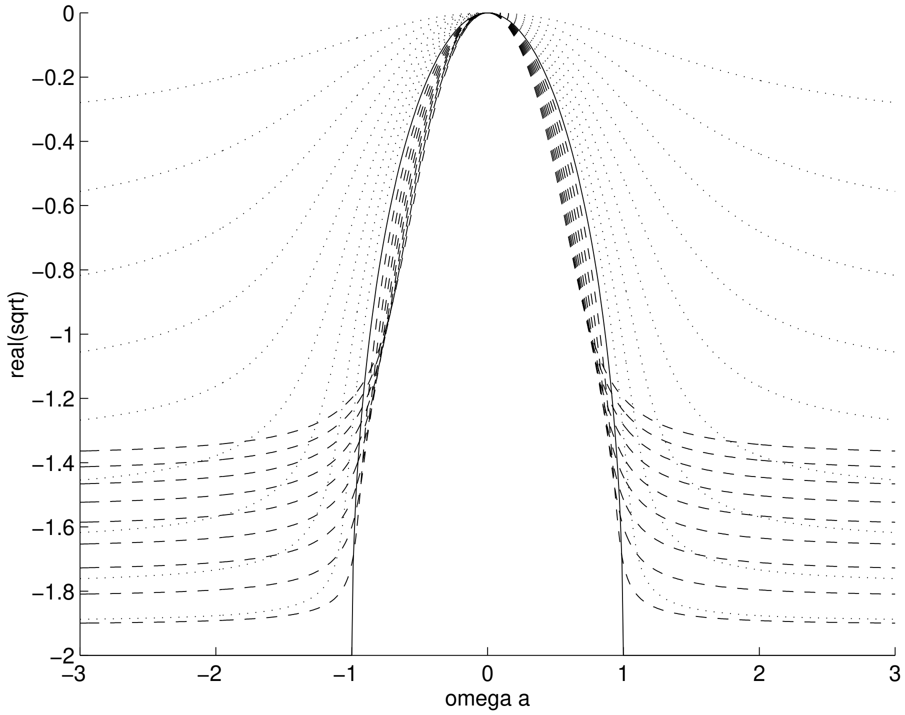

Figure 3: $2\,\mathrm{Re}(i\omega a\varrho)$ as function of ωa for various values of e/s — dotted lines $e/s < 1$, solid $e/s = 1$, dashed $e/s > 1$. $\lambda_3 = -2$ is fixed.

As the transformation T will be uniformly bounded, it remains to bound the exponentials of each block in (30) separately. Concerning the scalar entry $\mu_3 := \lambda_3 - i\omega a(e/s + \varrho)$ we find from (33)

$$\mathrm{Re}(\mu_3) = \lambda_3 - \mathrm{Re}(i\omega a\varrho) \leq \frac{\lambda_3}{2} \leq -\frac{1}{2}.$$

Therefore, the exponential is uniformly bounded

$$|\exp(\mu_3\tau)| \leq \exp(-\frac{\tau}{2}) \leq 1, \quad \forall \omega \in \mathcal{R}, \quad \tau \geq 0.$$

Next consider the remaining block in (30)

$$Q = -i\omega \left(\begin{array}{cc} f_s & f_c(1-\varrho) \\ 0 & a(1-\varrho) \end{array} \right).$$

Again, by (33) the eigenvalues $\mu_1 = -i\omega f_s$, $\mu_2 = -i\omega a(1-\varrho)$ have non–positive real part

$$\mathrm{Re}(\mu_2) = \mathrm{Re}(i\omega a\varrho) \leq \mathrm{Re}(\mu_1) = 0.$$

By the Kreiss–matrix–theorem ([3] Theorem 2.3.2) the exponential $\exp(Q\tau)$ is bounded if and only if there is a transformation R such that $|R|+|R^{-1}|$ is bounded and

$$R^{-1}QR = \mathrm{diag}(\mu_1, \mu_2).$$

Observe that μ_1 and μ_2 are distinct [when $\omega = 0$ then by assumption $f_s \neq a$ and otherwise by (34) $\mathrm{Re}(\mu_2) = \mathrm{Re}(i\omega a\varrho) < 0$]. Therefore, the matrix of right eigenvectors of

$$\begin{pmatrix} f_s & f_c(1-\varrho) \\ 0 & a(1-\varrho) \end{pmatrix}$$

is regular and transforms Q to diagonal form. Clearly, an eigenvector corresponding to the eigenvalue f_s is $(1,0)^T$. The eigenvector associated with $a(1-\varrho)$ is $(r,1)^T$, where

$$r := \frac{f_c(1-\varrho)}{a(1-\varrho)-f_s}.$$

Consequently,

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

The boundedness of R and $R^{-1} = \begin{pmatrix} 1 & -r \\ 0 & 1 \end{pmatrix}$ depends only on r. Since, by (31), ϱ is bounded, we need to bound the denominator $a(1-\varrho) - f_s$ away from zero. Observe that ϱ is real only for $\omega = 0$ or in the limit $|\omega| \to \infty$. When $\omega = 0$ then $\varrho = 0$ and by assumption $f_s \neq a$. The limits for $|\omega| \to \infty$ depend on the parameter e/s. When $e/s \in (0,1)$ then by (32) $\varrho(\omega) \to -e/s$ and again by assumption

$$\lim_{|\omega|\to\infty} a(1-\varrho) = a(1+\frac{e}{s}) = \frac{f}{s} \neq f_s.$$

When $e/s \geq 1$ then the limit of $a(1-\varrho)$ can become zero, but a general assumption is $f_s > 0$. To summarize, it follows that r, R, R^{-1} and $\exp(Q\tau)$ are bounded.

It remains to construct the transformation T. This is done in two steps. First the lower off–diagonal element in P is transformed to zero by

$$T_1 := \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & \varrho & 1 \end{pmatrix}.$$

By construction $\varrho = \varrho(\omega)$ is a root of the equation

$$i\omega a\varrho^2 - \lambda_*\varrho - i\omega a\frac{e}{s} = 0 \tag{35}$$

and therefore

$$T_1^{-1}PT_1 = \begin{pmatrix} -i\omega f_s & -i\omega f_c(1-\varrho) & i\omega f_c \\ 0 & -i\omega a(1-\varrho) & i\omega a \\ 0 & 0 & \lambda_3 - i\omega a(e/s+\varrho) \end{pmatrix}.$$

Again as ϱ is bounded T_1 and T_1^{-1} are bounded. In the second step the third column is transformed using

$$T_2 := \begin{pmatrix} 1 & 0 & \sigma_1 \\ 0 & 1 & \sigma_2 \\ 0 & 0 & 1 \end{pmatrix},$$

where $S := (\sigma_1, \sigma_2)^T$ is defined by

$$\left[(\lambda_3 - i\omega a(e/s+\varrho))\, I + i\omega \begin{pmatrix} f_s & f_c(1-\varrho) \\ 0 & a(1-\varrho) \end{pmatrix}\right] S = i\omega \begin{pmatrix} f_c \\ a \end{pmatrix}.$$

Clearly, when $\omega = 0$, $S = 0$. In the limit $|\omega| \to \infty$ it holds that

$$\begin{pmatrix} f_s - a(e/s+\varrho_\infty) & f_c(1-\varrho_\infty) \\ 0 & a(1-2\varrho_\infty - e/s) \end{pmatrix} S_\infty = \begin{pmatrix} f_c \\ a \end{pmatrix}.$$

Obviously, S_∞ is well defined and bounded when this system is regular.

If $e/s \in (0,1)$ then $\varrho_\infty = -e/s$ and the system is regular since $f_s > 0$ and $a(1+e/s) = f/s > 0$. The only other possible limit of ϱ is $\varrho_\infty = 1$. Then the system is again regular, since $f_s \neq f/s = a(e/s+1)$ and $-a(1+e/s) = -f/s < 0$.

After all, $T := T_1T_2$ is the desired transformation and Lemma 4.1 is proved.

5 Proofs of Theorems 1.1–1.3

Consider the family of equations

$$u_t + \sum_{j=1}^{d} A^{(j)} u_{x_j} = \frac{1}{\delta} Bu + Cu, \quad 0 < \delta \le 1, \quad t \ge 0 \tag{36}$$

with initial condition

$$u(x,0,\delta) = f(x), \quad x \in \mathcal{R}^d\ . \tag{37}$$

If $f \in M_0$ then (for arbitrary coefficient matrices $A^{(j)}$, B, C) the unique M_0–solution is

$$u(x,t,\delta) = \frac{1}{(2\pi)^{d/2}} \int\limits_{\mathcal{R}^d} e^{i\langle k,x\rangle} e^{P(ik,\delta)t} \hat{f}(k)\, dk\ ,$$

where $P(ik,\delta) = \frac{1}{\delta}B + C - i\sum_j k_j A^{(j)}$, and we write $u(\cdot,t,\delta) = S_0(t,\delta)f$. If the system is strongly hyperbolic, then $S_0(t,\delta) : M_0 \to M_0$ can be extended uniquely to a bounded linear operator $S(t,\delta) : L_2 \to L_2$. For initial data $f \in L_2$ we call $u(\cdot,t,\delta) = S(t,\delta)f$ the L_2–solution of (36).

Let us first consider the question of convergence of the M_0–solution $u(\cdot,t,\delta) = S_0(t,\delta)f$ as $\delta \to 0$. We assume that *COND3* is satisfied, but otherwise $A^{(j)}$, B, $C \in \mathcal{C}^{n\times n}$ are arbitrary. There is a transformation $T \in \mathcal{C}^{n\times n}$ such that $T^{-1}BT$ has block form (4) and all eigenvalues of $\widetilde{B}_{22}$ have negative real parts. Using the transformations

$$u = Tv, \quad f = Tg, \quad \widetilde{A}^{(j)} = T^{-1}A^{(j)}T, \quad \widetilde{C} = T^{-1}CT$$

we obtain the equivalent system

$$v_t + \sum_{j=1}^{d} \widetilde{A}^{(j)} v_{x_j} = \frac{1}{\delta}\widetilde{B}v + \widetilde{C}v, \quad v(x,0,\delta) = g(x) \ .$$

Fourier transformation yields

$$\hat{v}(k,t,\delta) = e^{\widetilde{P}(ik,\delta)t}\hat{g}(k), \quad k \in \mathcal{R}^d$$

with $\widetilde{P}(ik,\delta) = \frac{1}{\delta}\widetilde{B} + \widetilde{C} - i\sum_j k_j \widetilde{A}^{(j)}$. Our convergence results are based on the following lemma.

Lemma 5.1 *Assume COND3 and let $\widetilde{P}(ik,\delta)$ be determined as described above. Then, for all $t > 0$ and all $k \in \mathcal{R}^d$,*

$$\lim_{\delta\to 0} e^{\widetilde{P}(ik,\delta)t} = \begin{pmatrix} \exp(t\widetilde{C}_{11} - it\sum_j k_j \widetilde{A}^{(j)}_{11}) & 0 \\ 0 & 0 \end{pmatrix} . \tag{38}$$

For any fixed $t > 0$ and $k_0 > 0$ the convergence is uniform in $-k_0 \le k_j \le k_0$.

Proof: We write

$$\widetilde{P}(ik,\delta)t = \frac{t}{\delta}\left(\widetilde{B} + \delta\widetilde{C} - i\delta\sum_{j=1}^{d} k_j \widetilde{A}^{(j)}\right) .$$

For small δ, $\epsilon = \delta \max_j |k_j|$ is small and $\widetilde{B} + \delta\widetilde{C} - i\delta\sum_j k_j\widetilde{A}^{(j)} = \widetilde{B} + \mathcal{O}(\epsilon)$ is a perturbation of $\widetilde{B}$.

As the spectra of $\widetilde{B}_{11} = 0$ and $\widetilde{B}_{22}$ are disjoint, we can apply Lemma 6.1 in [4]. Hence, there is a transformation $T_\epsilon = I + \mathcal{O}(\epsilon)$ such that

$$\delta T_\epsilon^{-1} \widetilde{P}(ik, \delta) T_\epsilon = \begin{pmatrix} \delta \widetilde{C}_{11} - i\delta \sum_j k_j \widetilde{A}_{11}^{(j)} + \mathcal{O}(\epsilon^2) & 0 \\ 0 & \tilde{B}_{22} + \mathcal{O}(\epsilon) \end{pmatrix} .$$

Therefore,

$$e^{\widetilde{P}(ik,\delta)t} = T_\epsilon \begin{pmatrix} \exp(t\widetilde{C}_{11} - it\sum_j k_j \widetilde{A}_{11}^{(j)} + \mathcal{O}(\delta)) & 0 \\ 0 & \exp(\frac{t}{\delta}\tilde{B}_{22} + \mathcal{O}(1)) \end{pmatrix} T_\epsilon^{-1} .$$

As $\delta \to 0$ we have $\epsilon \to 0$, uniformly for $-k_0 \le k_j \le k_0$, and $T_\epsilon \to I$ as well as

$$\exp\left(\frac{t}{\delta}\tilde{B}_{22} + \mathcal{O}(1)\right) \to 0 .$$

This proves the lemma.

Proof of Theorem 1.3: Let us denote

$$\lim_{\delta\to 0} e^{\widetilde{P}(ik,\delta)t} = \begin{pmatrix} \exp(t\widetilde{C}_{11} - it\sum_j k_j \widetilde{A}_{11}^{(j)}) & 0 \\ 0 & 0 \end{pmatrix} =: Q(k,t), \quad t > 0 . \qquad (39)$$

Since $\hat{g}(k)$ has compact support, we obtain from the lemma that

$$v(x,t,\delta) = \frac{1}{(2\pi)^{d/2}} \int_{\mathcal{R}^d} e^{ikx} e^{\widetilde{P}(ik,\delta)t} \hat{g}(k) dk$$

converges to

$$\breve{v}(x,t) = \frac{1}{(2\pi)^{d/2}} \int_{\mathcal{R}^d} e^{ikx} Q(k,t) \hat{g}(k) dk \qquad (40)$$

as $\delta \to 0$. Here $x \in \mathcal{R}^d$ and $t > 0$ are arbitrary but fixed. The form of $Q(k,t)$ given above implies that $\breve{v}(\cdot,t) \in M_0$. Furthermore, the Fourier transforms of $v(\cdot,t,\delta)$ and $\breve{v}(\cdot,t)$ are $e^{\widetilde{P}(ik,\delta)t}\hat{g}(k)$ and $Q(k,t)\hat{g}(k)$, respectively. Therefore, the previous lemma and Parseval's relation imply

$$\|v(\cdot,t,\delta) - \breve{v}(\cdot,t)\| \to 0 \quad \text{as} \quad \delta \to 0$$

for any $t > 0$. Also, we can differentiate the Fourier representations of v and $\breve{v}$ with respect to x under the integral sign, obtaining convergence in H_m.

The form of $\breve{v}(x,t)$ and $Q(k,t)$ further imply that

$$\breve{v}_t^I + \sum_{j=1}^d \widetilde{A}_{11}^{(j)} \breve{v}_{x_j}^I = \widetilde{C}_{11}\breve{v}^I, \quad \breve{v}^{II} = 0, \quad t > 0 .$$

Thus far, we have defined $\breve{v}(x,t)$ only for $t > 0$, but we can use (39) and (40) to define the analytic function $\breve{v}(x,t)$ for all real x and t. Then we have $\breve{v}^I(x,0) = g^I(x)$. This finishes the proof of Theorem 1.3.

Proof of Theorem 1.1, Part a: By assumption *COND1* holds, which (cf. Lemma 2.2) is equivalent to the existence of constants K and α such that

$$|\exp\left(P(ik,\delta)t\right)| \le Ke^{\alpha t}, \quad \forall k \in \mathcal{R}^d,\ t \ge 0,\ 0 < \delta \le 1.$$

Therefore,

$$\|S(t,\delta)\| \le Ke^{\alpha t}, \quad t \ge 0, \quad 0 < \delta \le 1 \ . \tag{41}$$

Also note that we can apply Theorem 1.3 since *COND3* follows from *COND1* and *COND2*.

Let $f \in L_2$ be given and choose a sequence $f_j \in M_0$ with $f_j \to f$. (Convergence is w.r.t. the L_2–norm.) Then Theorem 1.3 yields that for $t > 0$

$$\lim_{\delta\to 0} S_0(t,\delta)f_j =: u_j$$

exists in M_0. By (41) we have

$$\|S_0(t,\delta)(f_j - f_l)\| \le Ke^{\alpha t}\|f_j - f_l\|, \quad 0 < \delta \le 1 \ ,$$

and we obtain for $\delta \to 0$,

$$\|u_j - u_l\| \le Ke^{\alpha t}\|f_j - f_l\| \ .$$

This implies convergence of u_j to a limit $\breve{u}$ in L_2. Finally, the inequalities

$$\begin{aligned}\|\breve{u} - S(t,\delta)f\| &\le \|\breve{u} - u_j\| + \|u_j - S(t,\delta)f_j\| + \|S(t,\delta)f_j - S(t,\delta)f\| \\ &\le \|\breve{u} - u_j\| + \|u_j - S(t,\delta)f_j\| + Ke^{\alpha t}\|f_j - f\|\end{aligned}$$

imply that

$$\lim_{\delta\to 0} S(t,\delta)f = \breve{u}, \quad t > 0 \ .$$

This proves part a) of Theorem 1.1.

Proof of Theorem 1.1, Part b: The system (36) is strongly hyperbolic, which implies that $S(t,\delta)f$ is defined for all $f \in L_2$. By an energy estimate and Lemma 2.1, we always have an estimate

$$\|S(t,\delta)\| \le Ke^{(\alpha+\beta/\delta)t}, \quad t \ge 0, \quad 0 < \delta \le 1 \ . \tag{42}$$

By assumption, for some $t^* > 0$ the limit $\lim_{\delta\to 0} S(t^*,\delta)f$ exists for all $f \in L_2$. Then, because of (42), we have boundedness,

$$\|S(t^*,\delta)f\| \le C(f), \quad 0 < \delta \le 1$$

for every $f \in L_2$. The principle of uniform boundedness implies existence of a constant c_1 such that

$$\|S(t^*, \delta)\| \leq c_1, \quad 0 < \delta \leq 1 \ ,$$

and therefore,

$$\left|\exp\left(\left[\frac{1}{\delta}B + C - i\sum_{j=1}^{d} k_j A^{(j)}\right] t^*\right)\right| \leq c_1, \quad 0 < \delta \leq 1, \quad k \in \mathcal{R}^d \ .$$

Setting $\tau = t^*/\delta$, $\omega = k\delta$, we have

$$\left|\exp\left(\left[B + \delta C - i\sum_{j=1}^{d} \omega_j A^{(j)}\right] \tau\right)\right| \leq c_1, \quad \tau \geq t^*, \quad 0 < \delta \leq 1, \quad \omega \in \mathcal{R}^d \ .$$

Due to continuity of the matrix exponential, we can let δ to zero and find

$$\left|\exp\left(\left[B - i\sum_{j=1}^{d} \omega_j A^{(j)}\right] \tau\right)\right| \leq c_1, \quad \tau \geq t^*, \quad \omega \in \mathcal{R}^d \ .$$

It remains to bound $\exp\left((B - i\sum_j \omega_j A^{(j)})\tau\right)$ for $0 \leq \tau \leq t^*$. According to (10), the estimate

$$\left|\exp\left(\left[\frac{1}{\delta}B - i\sum_{j=1}^{d} k_j A^{(j)}\right] t\right)\right| \leq K e^{\beta t/\delta}$$

holds for all $t \geq 0$, all $k \in \mathcal{R}^d$, and all $0 < \delta \leq 1$. Setting again $\tau = t/\delta$, $\omega = k\delta$, we obtain

$$\left|\exp\left(\left[B - i\sum_{j=1}^{d} \omega_j A^{(j)}\right] \tau\right)\right| \leq K e^{\beta\tau}, \quad \tau \geq 0, \quad \omega \in \mathcal{R}^d \ .$$

Therefore, if we define $c_2 := \max\{c_1, K e^{\beta t^*}\}$, then

$$\left|\exp\left(\left[B - i\sum_{j=1}^{d} \omega_j A^{(j)}\right] \tau\right)\right| \leq c_2, \quad \forall\tau \geq 0, \quad \forall\omega \in \mathcal{R}^d \ . \tag{43}$$

This shows that *COND1* holds.

It remains to show *COND2*. Let $\omega = 0$ in (43), then by [3] Lemma 2.3.1 all eigenvalues of B have a non–positive real part and purely imaginary eigenvalues are semisimple. Suppose that B has a purely imaginary eigenvalue $i\lambda \neq 0$, $\lambda \in \mathcal{R}$ and B is already blocked in the form

$$B = \begin{pmatrix} i\lambda I & 0 \\ 0 & B_{II} \end{pmatrix}, \quad i\lambda \notin \sigma(B_{II}).$$

Denote by

$$C = \begin{pmatrix} C_I & * \\ * & C_{II} \end{pmatrix}$$

the corresponding blocks of the matrix C. Then for small δ the spectra of $i\lambda I + \delta C_I$ and $B_{II} + \delta C_{II}$ are disjoint. Again by [4] Lemma 6.1 there is a transformation $T_\delta = I + \mathcal{O}(\delta)$ such that

$$T_\delta^{-1}\,(B + \delta C)\,T_\delta = \begin{pmatrix} i\lambda I + \delta C_I + \mathcal{O}(\delta^2) & 0 \\ 0 & B_{II} + \delta C_{II} + \mathcal{O}(\delta^2) \end{pmatrix}. \tag{44}$$

Next we construct special initial data for the stiff system (36) as follows. For sufficiently large $L > 0$ let $\phi : \mathcal{R}^d \to \mathcal{R}$ denote a function in C_0^∞ with $\phi(x) = 1$ for $x \in [-2L, 2L]^d$. Furthermore, let $y = (y_I, 0)^T \in \mathcal{C}^n$ such that y_I is an eigenvector of C_I with eigenvalue λ_I. We consider the solution of (36) with initial condition

$$u(x, 0, \delta) = \phi(x) y\ .$$

The finite speed of propagation (which only depends on the $A^{(j)}$, not on $(1/\delta)B$ and C) implies that in the smaller cube $[-L, L]^d$ the solution is still constant in space

$$u(x, t, \delta) = \exp\left(\left[\frac{1}{\delta} B + C\right] t\right) y, \quad x \in [-L, L]^d, \quad 0 \leq t \leq t_L.$$

With (44) and $T_\delta y = (y_I + \mathcal{O}(\delta), 0)^T$ it follows

$$\begin{aligned} (T_\delta u(x,t,\delta))_I &= \exp\left(\left[\frac{i\lambda}{\delta} I + C_I + \mathcal{O}(\delta)\right] t\right) y + \mathcal{O}(\delta) \\ &= e^{(i\lambda/\delta + \lambda_I)t} y_I + \mathcal{O}(\delta), \quad x \in [-L, L]^d, \quad 0 \leq t \leq t_L. \end{aligned}$$

Here $t_L \to \infty$ as $L \to \infty$. This demonstrates that $u(\cdot, t, \delta)$ does not converge as $\delta \to 0$, contrary to our assumption. Therefore, B does not have a purely imaginary eigenvalue, and Theorem 1.1 is proved.

Proof of Theorem 1.2: Let us first note that all solution operators appearing below are bounded in norm by $Ke^{\alpha t}$, where K and α are the constants appearing in (3). We will derive Theorem 1.2 from Theorem 1.1 and Theorem 1.3 and — as an essential step — have to show two limit processes to commute. In the following let $t > 0$ be fixed. Recall that $S_0(t,\delta) : M_0 \to M_0$ is the M_0–solution operator of (36) and $S(t,\delta) : L_2 \to L_2$ is its continuous extension. As $\delta \to 0$, we have by Theorem 1.1,

$$\lim_{\delta\to 0} S(t,\delta)f =: S(t)f, \quad f \in L_2 \ .$$

This defines a bounded linear operator $S(t) : L_2 \to L_2$. On the other hand, applying Theorem 1.3 to an arbitrary $f \in M_0$, we find

$$\lim_{\delta\to 0} S_0(t,\delta)f =: S_0(t)f, \quad f \in M_0 \ .$$

This determines a bounded linear operator $S_0(t) : M_0 \to M_0$. We claim that $S(t)$ defined above equals the continuous extension of $S_0(t)$. In other words, if we first extend $S_0(t,\delta)$ to L_2 and then send $\delta \to 0$, or if we first send $\delta \to 0$ —to obtain $S_0(t)$— and then extend to L_2, the result is the same. To see this, let $f \in L_2$ be arbitrary but fixed and choose a sequence $f_j \in M_0$ with $f_j \to f$. Then we have

$$\begin{aligned} \|S(t)f - S_0(t)f_j\| &\leq \|S(t)f - S(t,\delta)f\| + \|S(t,\delta)f - S_0(t,\delta)f_j\| \\ &\quad + \|S_0(t,\delta)f_j - S_0(t)f_j\| \end{aligned}$$

Here

$$\|S(t,\delta)f - S_0(t,\delta)f_j\| \leq Ke^{\alpha t}\|f - f_j\| \ ,$$

and we obtain $S_0(t)f_j \to S(t)f$ as $j \to \infty$. This proves that $S(t)$ is the continuous extension of $S_0(t)$ to L_2.

We can phrase the result also as follows: If $f \in L_2$ and $f_j \in M_0$ with $f_j \to f$, then

$$\lim_{j\to\infty}\lim_{\delta\to 0} S_0(t,\delta)f_j = \lim_{\delta\to 0}\lim_{j\to\infty} S_0(t,\delta)f_j \ . \tag{45}$$

To see this, note that the left–hand side equals

$$\lim_{j\to\infty} S_0(t)f_j = S(t)f$$

(as established above), whereas the right–hand side equals

$$\lim_{\delta\to 0} S(t,\delta)f = S(t)f \ .$$

(This is our definition of $S(t)f$ by Theorem 1.1.)

Next we show that the equilibrium system (5) is strongly hyperbolic. To this end, let $g^I \in M_0$ and set

$$g(x) = \begin{pmatrix} g^I(x) \\ 0 \end{pmatrix} .$$

If $u(\cdot, t, \delta) = S_0(t, \delta)Tg$ and $v = T^{-1}u$, then

$$\|v(\cdot, t, \delta)\| \leq Ke^{\alpha t}|T||T^{-1}|\|g\| = \widetilde{K}e^{\alpha t}\|g^I\| .$$

Furthermore, by Theorem 1.3 we have

$$v(\cdot, t, \delta) \to \breve{v}(\cdot, t) = \begin{pmatrix} \breve{v}^I(\cdot, t) \\ 0 \end{pmatrix} \quad \text{as} \quad \delta \to 0 ,$$

where $\breve{v}^I$ is the M_0–solution of (5) with $\breve{v}^I(x, 0) = g^I(x)$. This implies

$$\|\breve{v}^I(\cdot, t)\| \leq \widetilde{K}e^{\alpha t}\|g^I\| .$$

Since $g^I \in M_0$ and $t > 0$ were arbitrary, it follows that the Cauchy problem for (5) is well–posed and consequently (by definition) (5) is strongly hyperbolic. As a consequence, the L_2–solution operator for (5) is well–defined.

Let $S_0^E(t)$ denote the M_0–solution operator of the equilibrium system (5) and let $S^E(t)$ denote its continuous extension to L_2. To show the convergence statement of Theorem 1.2, we let $f \in L_2$ be arbitrary and choose a sequence $f_j \in M_0$ with $f_j \to f$. By Theorem 1.3,

$$\lim_{\delta \to 0} T^{-1}S_0(t, \delta)f_j = \begin{pmatrix} S_0^E(t)(T^{-1}f_j)^I \\ 0 \end{pmatrix}$$

for every f_j. Sending $j \to \infty$ we obtain

$$\lim_{j \to \infty} \lim_{\delta \to 0} T^{-1}S_0(t, \delta)f_j = \begin{pmatrix} S^E(t)(T^{-1}f)^I \\ 0 \end{pmatrix}$$

By (45) the left–hand side equals

$$\lim_{\delta \to 0} \lim_{j \to \infty} T^{-1}S_0(t, \delta)f_j = \lim_{\delta \to 0} T^{-1}S(t, \delta)f .$$

Thus we have proved

$$\lim_{\delta \to 0} T^{-1}S(t, \delta)f = \begin{pmatrix} S^E(t)(T^{-1}f)^I \\ 0 \end{pmatrix} ,$$

which is exactly the convergence claimed in Theorem 1.2.

It remains to show that the equilibrium system (5) satisfies the phase speed condition. This can be done easily by using the corresponding result in one space dimension shown in [4]. The following result is a consequence of Theorem 1.2 of [4]:

Lemma 5.2 *Let $A, B \in \mathcal{C}^{n\times n}$. Assume that the eigenvalues of A are real and semisimple, and assume there is c_1 such that*

$$|e^{(B-i\omega A)\tau}| \le c_1, \quad \forall \omega \in \mathcal{R}, \quad \forall \tau \ge 0 \ .$$

Further assume that B has no purely imaginary eigenvalue different from zero, and let T denote a transformation such that

$$T^{-1}BT = \begin{pmatrix} 0 & 0 \\ 0 & \widetilde{B}_{22} \end{pmatrix} \tag{46}$$

where all eigenvalues of $\widetilde{B}_{22}$ have negative real part. If

$$T^{-1}AT = \begin{pmatrix} \widetilde{A}_{11} & \widetilde{A}_{12} \\ \widetilde{A}_{21} & \widetilde{A}_{22} \end{pmatrix}$$

then the eigenvalues of $\tilde{A}_{11}$ are real and semisimple and lie between the minimal and maximal eigenvalue of A.

For any fixed $k \in \mathcal{R}^d$ we can apply this result to $A = \sum_j k_j A^{(j)}$ and the matrix B under consideration here to obtain the inclusion of the eigenvalues of $\sum_j k_j \widetilde{A}^{(j)}_{11}$ between the minimal and maximal eigenvalues of $\sum_j k_j A^{(j)}$.

We remark that the proof of Theorem 1.2 in [4], which yields the above Lemma, used the established convergence result for $\delta \to 0$ of solutions of corresponding hyperbolic systems. It seems worthwhile to find a more direct proof of the above Lemma.

Acknowledgement. We would like to thank Professor Kjell–Ove Widman for the kind invitation to the Mittag–Leffler Institute. We enjoyed the special atmosphere at this unique institut.

Our thanks also go to Professor Björn Engquist, the director of a great and inspiring scientific program during the fall 1997.

Furthermore we wish to acknowledge the support of Jens Lorenz's research by the US–National Science Foundation (contract number DMS–9404124) and by the US–Department of Energy (contract number DE–FG03–95ER25235). Finally, H. Joachim Schroll is supported by the German Research Foundation (DFG) through SFB 401.

References

[1] G. Q. Chen, C. D. Levermore, and T. P. Liu, *Hyperbolic Conservation Laws with Stiff Relaxation Terms and Entropy*, Comm. on Pure and Appl. Math., XLVII, 1994, pp. 787–830.

[2] E. Isaacson and B. Temple, *The structure of Asymptotic States in a Singular System of Conservation Laws*, Adv. in Appl. Math., 11, 1990, pp. 205–219.

[3] H.-O. Kreiss and J. Lorenz, *Initial–Boundary Value Problems and the Navier–Stokes Equations*, Academic Press, London, 1989.

[4] J. Lorenz and H. J. Schroll, *Stiff Well–Posedness for Hyperbolic Systems with Large Relaxation Terms*, Adv. in Differential Equations 2(4), 1997, pp 643–666.

[5] B. Sjögreen and B. Engquist, *Numerical Approximation of Hyperbolic Conservation Laws with Stiff Terms*, in proceedings of the Third International Conference on Hyperbolic Problems, B. Engquist and B. Gustafsson, eds., Studentlitteratur, 1991, pp 848–860.

[6] A. Tveito and R. Winther, *A Well–Posed System of Hyperbolic Conservation Laws*, in proceedings of the Third International Conference on Hyperbolic Problems, B. Engquist and B. Gustafsson, eds., Studentlitteratur, 1991, pp 888–898.

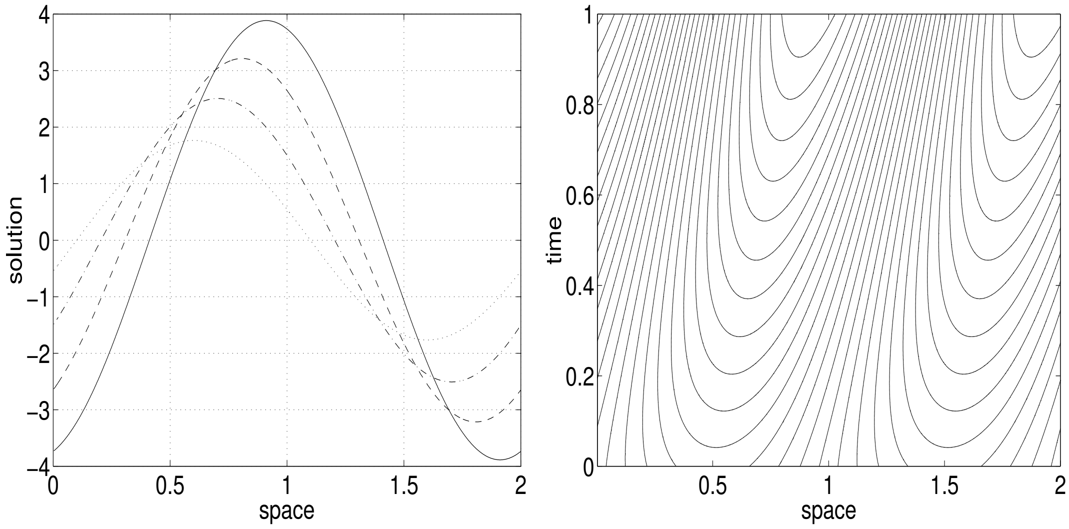

Figure 4: The first component of the solution of (28).
Left side: Dotted line $t = 1/4$, dash–dotted $t = 1/2$, dashed $t = 3/4$, solid $t = 1$.

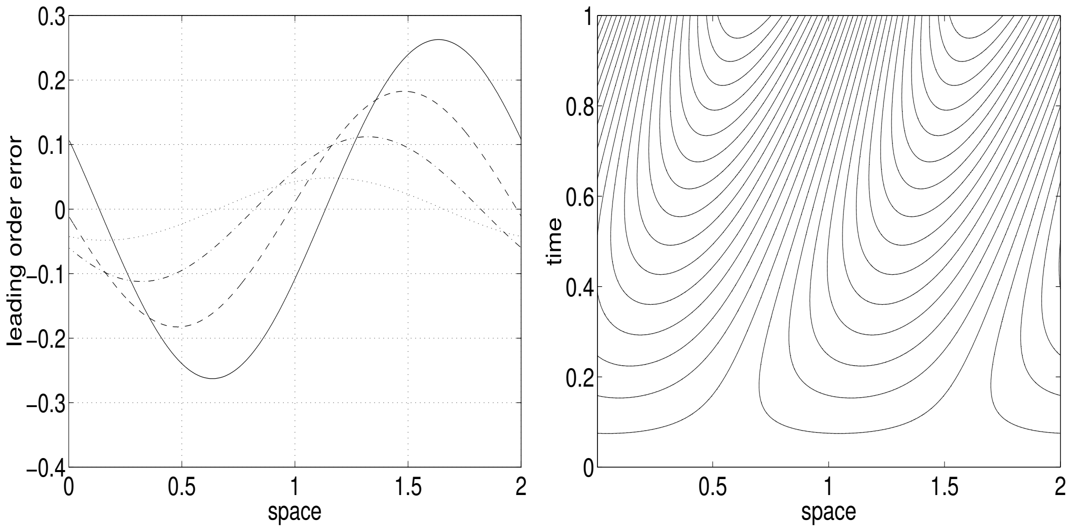

Figure 5: The leading order error.
Left side: Dotted line $t = 1/4$, dash–dotted $t = 1/2$, dashed $t = 3/4$, solid $t = 1$.

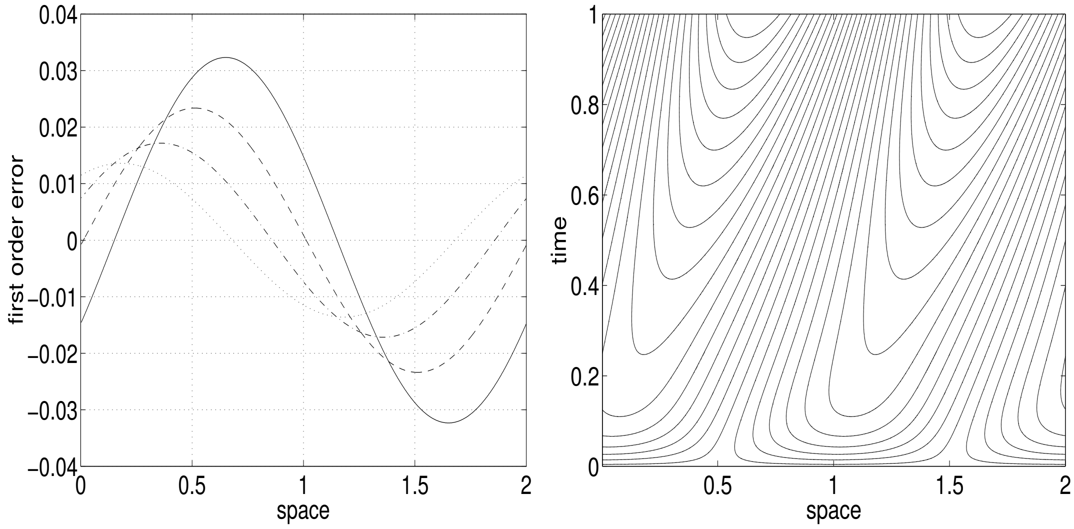

Figure 6: The first order error.
Left side: Dotted line $t = 1/4$, dash–dotted $t = 1/2$, dashed $t = 3/4$, solid $t = 1$.

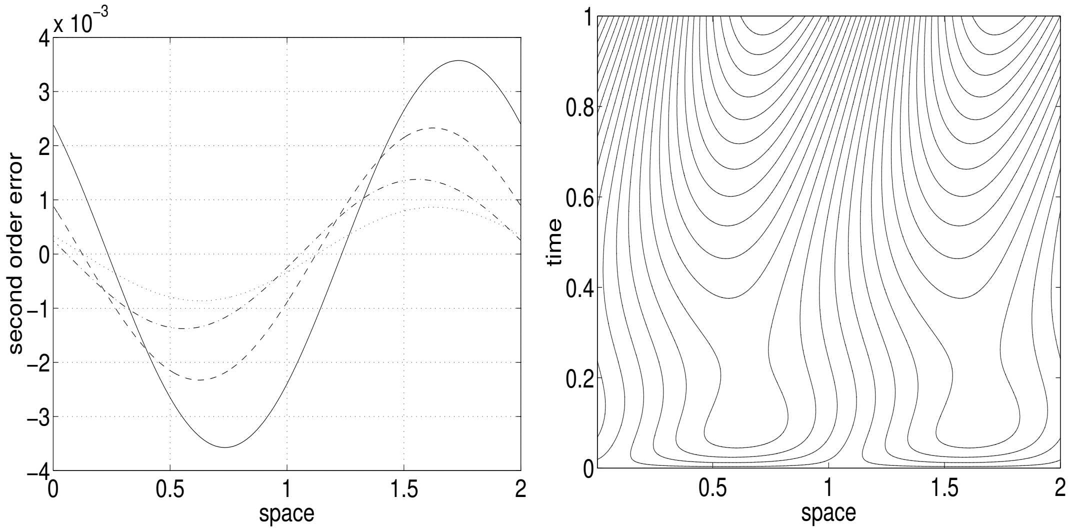

Figure 7: The second order error.
Left side: Dotted line $t = 1/4$, dash–dotted $t = 1/2$, dashed $t = 3/4$, solid $t = 1$.