Stiff Well–Posedness for Hyperbolic Systems with Large Relaxation Terms

(Linear Constant–Coefficient problems)

Jens Lorenz * and H. Joachim Schroll †

IGPM-Report 125, January 1996

Abstract

The Cauchy problem for linear constant–coefficient hyperbolic systems $u_t + Au_x = (1/\delta)Bu$ 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 nullspace. A concept of stiff well–posedness is introduced that ensures solution estimates independent of $0 < \delta \leq 1$. Under suitable assumptions, we prove convergence of the L_2–solution to a limit as δ tends to zero. The limit solves a reduced strongly hyperbolic system without zero–order term, the so–called equilibrium system, and we present a method to determine this limit system. For $2{\times}2$ systems the requirement of stiff well–posedness is shown to be equivalent to the well–known subcharacteristic condition, but in general the subcharacteristic condition is not suffient for stiff well–posedness. The theory is illustrated by examples.

Key words: Hyperbolic system, relaxation, well–posedness, subcharacteristic condition, equilibrium limit

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

*Numerical Mathematics, RWTH–Aachen, D-52056 Aachen, Germany and Dept. of Mathematics and Statistics, University of New Mexico, Albuquerque, NM 87131, USA.
Research supported by NSF grant DMS-9404124 and DOE grant DE-FG03-95ER2535.

†Numerical Mathematics, RWTH–Aachen, D-52056 Aachen, Germany
e-mail: `lorenz@igpm.rwth-aachen.de, schroll@igpm.rwth-aachen.de`

1 Introduction

We consider first–order systems with a large relaxation term

$$u_t + Au_x = \frac{1}{\delta} Bu, \quad 0 < \delta \ll 1, \quad x \in \mathcal{R}, \quad t \geq 0. \tag{1}$$

Here $u(x,t) \in \mathcal{C}^n$ and A, B are constant complex $n \times n$ matrices. It is assumed that the system is strongly hyperbolic, i.e., the eigenvalues of A are real, and A has a complete set of eigenvectors. We are mostly interested in the critical case where the matrix B has a non–trivial null space.

The system is called *stiff* if the relaxation parameter δ is small compared to the time scale determined by the characteristic speeds and some appropriate length scale.

Our aim is to study the behavior of the solutions of (1) when Cauchy data $f \in L_2$ are given at time $t = 0$,

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

We will give necessary and sufficient conditions which ensure that for all initial data $f \in L_2$ and all $t \geq 0$ 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. The derivation of this so–called equilibrium system will be given.

Before describing our general results, we consider an example by Chen, Levermore and Liu [1]. It shows that naive considerations do not always tell whether the solutions of (1), (2) have limits as δ tends to zero. The example is the following 3×3 system

$$u_t + \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix} u_x = \frac{1}{\delta} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ -2 & 0 & 1 \end{pmatrix} u, \quad 0 < \delta \ll 1\ . \tag{3}$$

Setting formally $\delta = 0$, one obtains the equilibrium condition $\breve{u}_3 = 2\breve{u}_1$. Assuming this relation in the first and second equation, one obtains the Cauchy–Riemann differential system for $\breve{u}_1$ and $\breve{u}_2$:

$$\begin{array}{rcrcl} \breve{u}_{1t} & + & \breve{u}_{2x} & = & 0 \\ \breve{u}_{2t} & - & \breve{u}_{1x} & = & 0. \end{array} \tag{4}$$

As this system is elliptic, the corresponding Cauchy problem is ill–posed. Therefore, it is doubtful if the system will describe the solution behavior as $\delta \to 0$.

If we instead substitute the equilibrium assumption $\breve{u}_3 = 2\breve{u}_1$ into the second and third equation of (3), we obtain the strictly hyperbolic system

$$\begin{aligned} \breve{u}_{1t} &- \frac{1}{2}\breve{u}_{2x} &= 0 \\ \breve{u}_{2t} &- \breve{u}_{1x} &= 0. \end{aligned} \tag{5}$$

The corresponding Cauchy problem is well–posed, but has one obtained the equilibrium dynamics?

Clearly, a more rigorous approach is called for. As will follow from our results, the solutions of (3) generally do not converge as $\delta \to 0$, and a well–posed equilibrium dynamics does not exist. The problem is not stiffly well–posed.

This is in contrast to the modified system

$$u_t + \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix} u_x = \frac{1}{\delta}\begin{pmatrix} -2 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} u, \quad 0 < \delta \ll 1 \tag{6}$$

Our results will imply that the solutions of (6) converge as $\delta \to 0$, and (5) describes the evolution of the limit.

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 $u : \mathcal{R} \to \mathcal{C}^n$, for which every component is square integrable, is defined by

$$\|u\|^2 = \int_{-\infty}^{\infty} |u(x)|^2 dx \ .$$

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,

$$\|u\|_{H_m}^2 = \sum_{j=0}^{m} \|D^j u\|^2, \quad D = \frac{d}{dx} \ .$$

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*

$$|e^{(B-i\omega A)\tau}| \leq c_1 \quad \forall \omega \in \mathcal{R}, \quad \forall \tau \geq 0 \ ;$$

- *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, [2], Theorem 2.3.2). Furthermore, let us introduce the term *stiff well–posedness* here. We say the Cauchy problem for (1) is *stiffly well–posed* if *COND1* holds and will show in Section 2 that stiff well–posedness is a natural extension of the usual well–posedness concept to the parameter dependent system (1).

Considering the matrix B of example (3), it is clear that 1 is an eigenvalue of B and therefore $|e^{B\tau}|$ is unbounded as $\tau \to \infty$. Consequently, *COND1* is violated for $\omega = 0$. For example (6), on the other hand, *COND1* will be shown below. Since *COND2* is obviously satisfied, Theorem 1.1 is applicable.

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

$$T^{-1}BT = \begin{pmatrix} 0 & 0 \\ 0 & \tilde{B}_{22} \end{pmatrix} =: \tilde{B} \tag{7}$$

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

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

one obtains the transformed system

$$v_t + \tilde{A}v_x = \frac{1}{\delta}\tilde{B}v, \quad v(x,0,\delta) = g(x) \ .$$

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 $\tilde{B}_{22}$ leads to rapid decay of v^{II} as $\delta \to 0$, and this suggests the limit system

$$\breve{v}^I_t + \tilde{A}_{11}\breve{v}^I_x = 0 \ . \tag{8}$$

We will prove:

Theorem 1.2 *Suppose the system (1) is strongly hyperbolic and COND1, COND2 are satisfied. Let u=Tv, f=Tg denote the transformation described above with $f \in L_2$. Then, for any $t > 0$,*

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

Here $\breve{v}^I(\cdot,t)$ is the L_2–solution of (8) with initial data $\breve{v}^I(x,0) = g^I(x)$, and convergence holds w.r.t. the L_2–norm. Furthermore, the system (8) is strongly hyperbolic and all eigenvalues of $\tilde{A}_{11}$ lie between the minimal and maximal eigenvalue of A.

The eigenvalues of $\tilde{A}_{11}$ are the characteristic speeds of the equilibrium system (8), and according to the above theorem these lie between the minimal and maximal characteristic speeds of (1) 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 (7) can be obtained, and a system (8) can be derived. Then we say that the *subcharacteristic* *condition* is fulfilled if (8) is strongly hyperbolic and all eigenvalues of $\tilde{A}_{11}$ lie between the minimal and maximal eigenvalue of A. Theorem 1.2 says that the subcharacteristic condition is a consequence of *COND1* and *COND2*.

Conversely, *COND3* together with the subcharacteristic condition generally do not imply *COND1*, as we will show by Example 4.4 below.

In the 2×2 case, however, *COND3* together with the subcharacteristic condition implies *COND1*.

Theorem 1.3 *Consider a 2×2 system (1). Assume*

- *the system is strongly hyperbolic;*
- *COND3 holds and the null space of B has dimension one;*
- *the subcharacteristic condition is valid.*

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

In the above theorem the assumption of strong hyperbolicity is essential as we will show by Example 4.5.

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. We let M_0 denote the space of all $f : \mathcal{R} \to \mathcal{C}^n$ whose Fourier transform is in C_0^∞. See Sect. 2 for details. The space M_0 is dense in L_2 and consists of analytic functions. Then, for arbitrary $A, B \in \mathcal{C}^{n\times n}$, 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 1.4 *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*

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

Here $\breve{u} = T\breve{v}$ and the function $\breve{v}^I$ is the M_0–solution of (8) with initial data $\breve{v}^I(x,0) = g^I(x)$.

Under the assumption of Theorem 1.4 neither system (1) nor (8) is necessarily hyperbolic, and even if both systems are hyperbolic, the subcharacteristic condition might be violated.

The rest of the paper is organized as follows: First we review some basic properties of non-stiff systems. In Section 3 the main Theorems 1.1–1.4 shall be proved. In Section 4 several examples from the literature are discussed. Finally, in Section 5 numerical experiments illustrate the theoretical results.

2 Well–Posedness and Stiff Well–Posedness

For every fixed positive δ the system (1) is of the form

$$u_t + Au_x = Cu \,. \tag{9}$$

The Cauchy problem for such systems with zero–order term Cu is well understood and the results are included here for completeness. (For a more general account see [2], for example.) Let us first assume that $A, C \in \mathcal{C}^{n\times n}$ are arbitrary matrices. Given initial data in

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

there is a unique M_0–solution of (9), namely

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

Here $P(ik) = C - ikA$ is the symbol of the operator $C - A\frac{\partial}{\partial x}$. 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.

The Cauchy problem for (9) is called *well–posed* (in the L_2 sense) if the M_0–solutions can be estimated by there initial data f for $t \geq 0$. More precisely, well–posedness requires the existence of constants K and α such that

$$\|u(\cdot,t)\| \leq Ke^{\alpha t}\|f\| \quad \forall\, f \in M_0,\, t \geq 0 \tag{10}$$

or equivalently if ([2], Theorem 2.2.2)

$$|e^{P(ik)t}| \leq Ke^{\alpha t} \quad \forall\, k \in \mathcal{R}, t \geq 0\,. \tag{11}$$

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

If the Cauchy problem for (9) 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 (9). We have

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

as follows directly from (10).

It is not difficult to show that well–posedness of the Cauchy problem for (9) is independent of the zero–order term Cu, and one has the following result.

Theorem 2.1 *Let $A, C \in \mathcal{C}^{n\times n}$ be arbitrary. The following conditions are equivalent.*

1. *The Cauchy Problem for $u_t + Au_x = Cu$ is well–posed.*
2. *All eigenvalues of A are real and semisimple.*
3. *There exists a Hermitian positive definite matrix H such that HA is Hermitian, i.e., $HA = A^*H$.*

Proof: Equivalence of 1. and 2. is elementary. See [2] Theorem 2.1.2, for example. If 2. is assumed, then there exists $U \in \mathcal{C}^{n\times n}$ such that $U^{-1}AU$ is real and diagonal. The matrix $H = (U^{-1})^*U^{-1}$ satisfies the condition in 3. Conversely, if 3. is assumed and $H = S^2, S = S^* > 0$, then SAS^{-1} is Hermitian. Condition 2 follows.

If the conditions of the above theorem are met, then $u_t + Au_x = Cu$ is called *strongly hyperbolic*. The matrix H of condition 3 is called a *symmetrizer* of the system.

Assuming the conditions of Theorem 2.1 to be fulfilled, consider the ODE

$$\dot{y} = P(ik)y, \quad P(ik) = C - ikA, \quad (k \in \mathcal{R} \text{ fixed})$$

and obtain

$$\begin{aligned} \frac{d}{dt}\langle y, Hy\rangle &= \langle \dot{y}, Hy\rangle + \langle y, H\dot{y}\rangle \\ &= \langle Cy, Hy\rangle + \langle y, HCy\rangle \end{aligned}$$

since $HA = A^*H$. Therefore,

$$\frac{d}{dt}\langle y, Hy\rangle \leq 2|C||H||H^{-1}|\langle y, Hy\rangle \ .$$

Denoting the norm corresponding to the inner product $\langle u, Hv\rangle$ by $|\cdot|_H$, we obtain the estimate

$$|e^{P(ik)t}|_H \leq e^{|C||H||H^{-1}|t}$$

and therefore

$$|e^{P(ik)t}| \leq Ke^{|C|K^2 t}, \quad K^2 = |H||H^{-1}| \ .$$

This shows the following: a) If $C = 0$, then one can take $\alpha = 0$ in the well–posedness estimate, i.e., there is no exponential growth. b) If $C = \frac{1}{\delta}B, \delta > 0$, then one obtains

$$|e^{P(ik,\delta)t}| \leq Ke^{ct/\delta}, \quad P(ik, \delta) = \frac{1}{\delta}B - ikA \tag{12}$$

with $c = |B||H||H^{-1}|$. Clearly, if $0 < \delta \ll c$, the last estimate becomes practically useless. But of course, the estimate is not always sharp, and it will be our task to improve it under more specific assumptions on A and B.

Instead of (12) a δ–independent bound like (11)

$$|e^{P(ik,\delta)t}| \leq Ke^{\alpha t} \quad \forall\ k \in \mathcal{R}, t \geq 0, 0 < \delta \leq 1 \ . \tag{13}$$

where K and α are independent of k, t and δ would be useful. Introducing the variables $\omega = \delta k$ and $\tau = t/\delta$ this inequality reads

$$|e^{(B - i\omega A)\tau}| \leq Ke^{\alpha\tau\delta} \quad \forall\ \omega \in \mathcal{R}, \tau \geq 0 \ . \tag{14}$$

The requirement that (14) holds uniformly for $0 < \delta \leq 1$ implies $|e^{(B-i\omega A)\tau}| \leq K$; thus one can always assume $\alpha = 0$ in (13) without loss of generality, and again there is no exponential growth. In other words, requiring *COND1* is nothing else than postulating the usual well–posedness estimate to hold uniformly for the parameter dependent system (1). In this sense, stiff well–posedness is a natural extension of the non–stiff concept.

We end this section by giving a simple sufficient criterion for stiff well–posedness. Here we use the following notations: B^* denotes the complex conjugate transpose of a matrix B, and if Q is a Hermitian matrix, then we write $Q \leq 0$ if $y^* Q y \leq 0$ for all $y \in \mathcal{C}^n$.

Theorem 2.2 *Consider $u_t + Au_x = (1/\delta)Bu, 0 < \delta \leq 1$, where $A, B \in \mathcal{C}^{n\times n}$. Assume that there is a positive definite Hermitian matrix H such that HA is Hermitian and*

$$HB + B^*H \leq 0 \ .$$

Then $|e^{(B-i\omega A)\tau}| \leq const$ for all $\omega \in \mathcal{R}$ and all $\tau \geq 0$, i.e., the system is stiffly well–posed.

Proof: Let $P = P(\omega) = B - i\omega A$ and let $y(t)$ solve $\dot{y} = Py, y(0) = y_0$. Then we have

$$\begin{aligned} \frac{d}{dt}|y|_H^2 &= \langle \dot{y}, Hy \rangle \ + \ \langle y, H\dot{y} \rangle \\ &= (1/\delta)\Big(\langle By, Hy \rangle \ + \ \langle y, HBy \rangle\Big) \\ &\leq 0 \ . \end{aligned}$$

Therefore, $|y(t)|_H \leq |y_0|_H$ for $t \geq 0$ and consequently $|e^{Pt}|_H \leq 1$ for $t \geq 0$. This proves the theorem.

3 Proofs of Theorems 1.1–1.4

Consider the family of equations

$$u_t + Au_x = \frac{1}{\delta}Bu, \quad 0 < \delta \leq 1, \quad t \geq 0 \qquad (15)$$

with initial condition

$$u(x, 0, \delta) = f(x), \quad x \in \mathcal{R} \ . \tag{16}$$

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

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

where $P(ik, \delta) = \frac{1}{\delta}B - ikA$, and we write $u(\cdot, t, \delta) = S_0(t, \delta)f$. If the system is strongly hyperbolic, i.e., all eigenvalues of A are real and semisimple, 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 (15).

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 $A, B \in \mathcal{C}^{n\times n}$ are arbitrary. There is a transformation $T \in \mathcal{C}^{n\times n}$ such that $T^{-1}BT$ has block form (7) and all eigenvalues of $\tilde{B}_{22}$ have negative real parts. Using the transformations

$$u = Tv, \quad f = Tg, \quad \tilde{A} = T^{-1}AT$$

we obtain the equivalent system

$$v_t + \tilde{A}v_x = \frac{1}{\delta}\tilde{B}v, \quad v(x, 0, \delta) = g(x) \ .$$

Fourier transformation yields

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

with $\tilde{P}(ik, \delta) = \frac{1}{\delta}\tilde{B} - ik\tilde{A}$. Our convergence results are based on the following lemma.

Lemma 3.1 *Assume COND3 and let $\tilde{P}(ik, \delta)$ be determined as described above. Then, for all $t > 0$ and all real k,*

$$\lim_{\delta\to 0} e^{\tilde{P}(ik,\delta)t} = \begin{pmatrix} \exp(-ikt\tilde{A}_{11}) & 0 \\ 0 & 0 \end{pmatrix} \ . \tag{17}$$

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

Proof: We write

$$\tilde{P}(ik,\delta)t = \frac{t}{\delta}(\tilde{B}+\epsilon\tilde{A}), \quad \epsilon = -ik\delta \ .$$

For small δ, the term $\epsilon\tilde{A}$ is a perturbation of $\tilde{B}$, and we can apply Lemma 6.1 given in the Appendix since the spectra of $\tilde{B}_{11}=0$ and $\tilde{B}_{22}$ are disjoint. If $|\epsilon|$ is small enough, then there is a transformation $T_\epsilon = I+\mathcal{O}(\epsilon)$ such that

$$T_\epsilon^{-1}(\tilde{B}+\epsilon\tilde{A})T_\epsilon = \begin{pmatrix} \epsilon\widetilde{A}_{11}+\mathcal{O}(\epsilon^2) & 0 \\ 0 & \tilde{B}_{22}+\mathcal{O}(\epsilon) \end{pmatrix} \ .$$

Therefore,

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

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

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

This proves the lemma.

Proof of Theorem 1.4: Let us denote

$$\lim_{\delta\to 0} e^{\tilde{P}(ik,\delta)t} = \begin{pmatrix} \exp(-ikt\widetilde{A}_{11}) & 0 \\ 0 & 0 \end{pmatrix} =: Q(k,t), \quad t>0 \ . \tag{18}$$

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

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

converges to

$$\breve{v}(x,t) = \frac{1}{\sqrt{2\pi}}\int_{\mathcal{R}} e^{ikx}Q(k,t)\hat{g}(k)dk \tag{19}$$

as $\delta \to 0$. Here $x \in \mathcal{R}$ 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 $e^{\tilde{P}(ik,\delta)t}\hat{g}(k)$ 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 $\|\cdot\|_{H_m}$.

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

$$\breve{v}^I_t + \tilde{A}_{11}\breve{v}^I_x = 0, \quad \breve{v}^{II} = 0, \quad t>0 \; .$$

Thus far, we have defined $\breve{v}(x,t)$ only for $t>0$, but we can use (18) and (19) 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 1.4.

Proof of Theorem 1.1, Part a: Assuming *COND1*, we have

$$|e^{((1/\delta)B-ikA)t}| \le c_1$$

for all $k\in\mathcal{R}, t\ge 0, 0<\delta\le 1$, and therefore,

$$\|S(t,\delta)\| \le c_1, \quad t\ge 0, \quad 0<\delta\le 1 \; . \tag{20}$$

Also note that we can apply Theorem 1.4 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 1.4 yields that

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

exists in M_0. By (20) we have

$$\|S_0(t,\delta)(f_j-f_l)\| \le c_1\|f_j-f_l\|, \quad 0<\delta\le 1 \; ,$$

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

$$\|u_j-u_l\| \le c_1\|f_j-f_l\| \; .$$

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

$$\begin{aligned}
\|\breve{u}-S(t,\delta)f\| &\le \|\breve{u}-u_j\| + \|u_j - S(t,\delta)f_j\| \\
&\quad + \|S(t,\delta)f_j - S(t,\delta)f\| \\
&\le \|\breve{u}-u_j\| + \|u_j - S(t,\delta)f_j\| \\
&\quad + c_1\|f_j - f\|
\end{aligned}$$

imply that

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

This proves part a) of Theorem 1.1.

Proof of Theorem 1.1, Part b: The system (15) is strongly hyperbolic, which implies that $S(t,\delta)f$ is defined for all $f \in L_2$. Further, according to (12), we always have an estimate

$$\|S(t,\delta)\| \le Ke^{ct/\delta}, \quad t \ge 0, \quad 0 < \delta \le 1 \ . \tag{21}$$

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 (21), 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

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

and therefore,

$$|e^{((1/\delta)B - ikA)t^*}| \le c_1, \quad 0 < \delta \le 1, \quad k \in \mathcal{R} \ .$$

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

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

and it remains to bound $|e^{(B-i\omega A)\tau}|$ for $0 \le \tau \le t^*$. According to (12), the estimate

$$|e^{((1/\delta)B - ikA)t}| \le Ke^{ct/\delta}$$

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

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

Therefore, if we define

$$c_2 = \max\{c_1, Ke^{ct^*}\} \ ,$$

then

$$|e^{(B - i\omega A)\tau}| \leq c_2, \quad \forall \tau \geq 0, \quad \forall \omega \in \mathcal{R} \ .$$

This shows that *COND1* holds.

Now suppose that B has a purely imaginary eigenvalue $\lambda = i\beta, \beta \in \mathcal{R}, \beta \neq 0$, and let $Bw = i\beta w, w \neq 0$. For sufficiently large $L > 0$ let $\phi : \mathcal{R} \to \mathcal{R}$ denote a function in C_0^∞ with $\phi(x) = 1$ for $-2L \leq x \leq 2L$. We consider the solution $u(x, t, \delta)$ of (15) with initial condition

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

The finite speed of propagation (which only depends on the eigenvalues of A) implies that u is given by

$$u(x, t, \delta) = e^{(1/\delta)Bt} w = e^{i\beta t/\delta} w$$

for $-L \leq x \leq L, 0 \leq t \leq t_L$. 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 c_1, the constant appearing in *COND1*. We will derive Theorem 1.2 from Theorem 1.1 and Theorem 1.4 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 (15) 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 1.4 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

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

Here

$$\|S(t,\delta)f - S_0(t,\delta)f_j\| \le c_1 \|f - f_j\| \ ,$$

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

$$\lim_{j\to\infty} \lim_{\delta\to 0} S_0(t,\delta) f_j = \lim_{\delta\to 0} \lim_{j\to\infty} S_0(t,\delta) f_j \ . \tag{22}$$

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

$$\|v(\cdot,t,\delta)\| \le c_3 \|g\| = c_3 \|g^I\| \ .$$

Furthermore, by Theorem 1.4 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 (8) with $\breve{v}^I(x,0) = g^I(x)$. This implies

$$\|\breve{v}^I(\cdot,t)\| \le c_3 \|g^I\| \ .$$

Since $g^I \in M_0$ and $t > 0$ were arbitrary, it follows that the Cauchy problem for (8) is well–posed and consequently (by definition) (8) is strongly hyperbolic. As a consequence, the L_2–solution operator for (8) is well–defined.

Let $S_0^E(t)$ denote the M_0–solution operator of the equilibrium system (8) 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 1.4,

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

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 (22) 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 (8) satisfies the subcharacteristic condition. But this follows quite easily from the convergence result that has been established. Let $\tilde{\lambda}_M$ denote the maximal eigenvalue of $\tilde{A}_{11}$ and let λ_M denote the maximal eigenvalue of A. Suppose that $\tilde{\lambda}_M > \lambda_M$. Then we choose a function $\phi : \mathcal{R} \to \mathcal{R}$ with $\phi \in C_0^\infty, \phi(x) > 0$ for $-1 < x < 0$ and $\phi(x) = 0$ otherwise. Let $\tilde{A}_{11} w = \tilde{\lambda}_M w, w \neq 0$. We consider initial data

$$g(x) = \begin{pmatrix} \phi(x)w \\ 0 \end{pmatrix}, \quad f(x) = Tg(x) \ .$$

If $u(\cdot, t, \delta) = S(t, \delta)f$, then we have

$$u(x, t, \delta) = 0 \quad \text{for} \quad x \geq \lambda_M t, \quad t \geq 0 \ ,$$

because λ_M is the maximal speed of propagation in (15). On the other hand, the solution $\breve{v}^I$ of

$$\breve{v}^I_t + \tilde{A}_{11}\breve{v}^I_x = 0, \quad \breve{v}^I(x, 0) = \phi(x)w$$

reads

$$\breve{v}^I(x, t) = \phi(x - \tilde{\lambda}_M t)w \ .$$

Clearly, $\breve{v}^I$ is positive in a subregion of the set $x \geq \lambda_M t$ if $\tilde{\lambda}_M > \lambda_M$. This contradicts our convergence result

$$T^{-1}u(\cdot, t, \delta) \to \begin{pmatrix} \breve{v}^I(\cdot, t) \\ 0 \end{pmatrix} \quad \text{as} \quad \delta \to 0 \ .$$

Proof of Theorem 1.3: Under our assumptions, the system can be transformed to block form

$$v_t + \widetilde{A}v_x = \frac{1}{\delta}\widetilde{B}v,$$

where

$$\widetilde{B} = T^{-1}BT = \begin{pmatrix} 0 & 0 \\ 0 & -b \end{pmatrix}, \quad \operatorname{Re} b > 0 \ ,$$

$\widetilde{A} = T^{-1}AT$ and $v = T^{-1}u$. Without loss of generality we can assume that the original system is already blocked,

$$A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 0 \\ 0 & -b \end{pmatrix}, \quad \operatorname{Re} b > 0.$$

We show first that $0 \leq a_{12}a_{21} \in \mathcal{R}$. The subcharacteristic condition states $a_{11} \in \mathcal{R}$ and

$$\lambda_{min}(A) \leq a_{11} \leq \lambda_{max}(A).$$

The eigenvalues of A are

$$\lambda_{\pm}(A) = \frac{1}{2}\left(a_{11} + a_{22} \pm \sqrt{(a_{11} - a_{22})^2 + 4a_{12}a_{21}}\right).$$

Since the system is assumed to be hyperbolic, we have $\lambda_{\pm}(A) \in \mathcal{R}$. Therefore, $a_{22} \in \mathcal{R}$ and $0 \leq (a_{11} - a_{22})^2 + 4a_{12}a_{21} \in \mathcal{R}$. Consequently, $a_{12}a_{21} \in \mathcal{R}$.

Now suppose that $a_{12}a_{21} < 0$, thus $a_{11} \neq a_{22}$. If $a_{11} > a_{22}$ we have

$$\lambda_{-}(A) \leq \lambda_{+}(A) < \frac{1}{2}\left(a_{11} + a_{22} + \sqrt{(a_{11} - a_{22})^2}\right) = a_{11} \ ,$$

which contradicts the subcharacteristic condition. Similarly, if $a_{11} < a_{22}$ then

$$a_{11} < \lambda_{-}(A) \leq \lambda_{+}(A).$$

Therefore, $a_{12}a_{21} \geq 0$ is proved.

Case 1: First let $a_{12}a_{21} > 0$. Then we can apply Theorem 2.2 with

$$H = \begin{pmatrix} \alpha & 0 \\ 0 & 1 \end{pmatrix} \ , \quad \alpha = \frac{a_{12}a_{21}}{|a_{12}|^2} > 0 \ .$$

Clearly, HA is Hermitian and $HB = B \leq 0$. Thus Theorem 2.2 implies stiff well–posedness.

Case 2: Let $a_{12}a_{21} = 0$, and assume $a_{21} = 0$, say. Then we have

$$C = -\begin{pmatrix} i\omega a_{11} & i\omega a_{12} \\ 0 & b + i\omega a_{22} \end{pmatrix} \ .$$

By Lemma 6.2 the transformation

$$T = \begin{pmatrix} 1 & s \\ 0 & 1 \end{pmatrix} \ , \quad s = \frac{i\omega a_{12}}{b + i\omega(a_{22} - a_{11})}$$

diagonalizes C, and one obtains $T^{-1}CT =: \Lambda$ where the diagonal entries of Λ are the same as those of C. Therefore, $|e^{\Lambda t}| \leq 1$ for $t \geq 0$. Since s is a bounded function of $\omega \in \mathcal{R}$, we have $|T| + |T^{-1}| \leq const$, and the desired uniform estimate for $|e^{Ct}|$ follows.

4 Examples

4.1 The scalar equation

The simplest example is the scalar equation

$$u_t + a u_x = \frac{1}{\delta} b u, \quad 0 < \delta \ll 1$$

where $a, b \in \mathcal{C}$. Here the properties of solutions can easily be determined from the symbol

$$P(ik, \delta) = b/\delta - ika.$$

Hyperbolicity requires $a \in \mathcal{R}$, and then the exponential $e^{P(ik,\delta)t}$ is bounded independent of $\delta > 0$ and $t \geq 0$ if and only if $\operatorname{Re} b \leq 0$. In other words, *COND1* is equivalent to $\operatorname{Re} b \leq 0$.

When $b \neq 0$ is purely imaginary then $e^{(b/\delta - ika)t}$ does not converge as $\delta \to 0$, confirming that *COND2* is necessary for convergence. Together, *COND1+2* require that either $b = 0$ — then there is no δ–dependence — or $\operatorname{Re} b < 0$. In the latter case, all solutions converge to zero as $\delta \to 0$ for $t > 0$.

4.2 The chromatography system

Our next example is the linear version of a system studied by Tveito and Winther [5],

$$u_t + \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} u_x = \frac{1}{\delta} \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix} u, \quad 0 < \delta \ll 1 \, .$$

This system arises in chromatography, cf. [7]. Here the spectrum of B is $\lambda_1 = -2$, $\lambda_2 = 0$ and obviously *COND2+3* are fulfilled.

The matrix of right eigenvectors of B is

$$T = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad \text{and} \quad T^{-1} B T = \begin{pmatrix} 0 & 0 \\ 0 & -2 \end{pmatrix}.$$

The transformed system reads

$$v_t + \frac{1}{2} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} v_x = \frac{1}{\delta} \begin{pmatrix} 0 & 0 \\ 0 & -2 \end{pmatrix} v, \quad v = T^{-1} u$$

and the equilibrium equation is

$$\breve{v}_t^I + \frac{1}{2}\breve{v}_x^I = 0.$$

The subcharacteristic condition is satisfied since $0 < 1/2 < 1$, and therefore the system is stiffly well–posed by Theorem 1.3.

We remark that even a fully nonlinear version of this system is well understood [5]. Due to a special monotonicity property of the source term, the total variation of the entropy solution is bounded independently of δ and this bound implies further interesting results like error estimates for finite difference schemes [4].

4.3 The linear p–system

Another interesting example is the linear p–system, depending on two parameters a and δ:

$$u_t + \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} u_x = \frac{1}{\delta}\begin{pmatrix} 0 & 0 \\ a & -1 \end{pmatrix} u, \quad 0 < \delta \ll 1, \quad a \in \mathcal{R} \tag{23}$$

LeVeque and Wang [3] observed strange oscillations in the solution when $|a| > 1$ and δ is small. In the above example, the eigenvalues of B are $\lambda_1(B) = 0$, $\lambda_2(B) = -1$ and *COND2+3* are fulfilled. The matrix of right eigenvectors of B is

$$T = \begin{pmatrix} 1 & 0 \\ a & 1 \end{pmatrix}$$

and the transformed system reads

$$v_t + \begin{pmatrix} a & 1 \\ 1-a^2 & -a \end{pmatrix} v_x = \frac{1}{\delta}\begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix} v, \quad v = T^{-1}u.$$

Therefore, the equilibrium equation is the linear advection equation

$$\breve{v}_t^I + a\breve{v}_x^I = 0.$$

Since the eigenvalues of A are ± 1, the subcharacteristic condition reads

$$-1 \leq a \leq 1.$$

Thus Theorem 1.3 implies that the p–system is stiffly well–posed *if and only if* $|a| \leq 1$.

In particular, Theorem 1.3 says that *COND1* does not hold if $|a| > 1$, and consequently, by Theorem 1.1, convergence as $\delta \to 0$ does not hold for general initial data in L_2. On the other hand, since *COND3* holds, M_0–solutions converge for all $a \in \mathcal{C}$ as $\delta \to 0$ (Theorem 1.4). For practical purposes and computations, convergence or non–convergence of L_2–solutions is more important than convergence of M_0–solutions, however. This is reflected by the numerical observations in [3] for $|a| > 1$.

4.4 Subcharacteristic condition is not sufficient

The following system

$$u_t + \begin{pmatrix} a & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix} u_x = \frac{1}{\delta} \begin{pmatrix} 0 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & b & -1 \end{pmatrix} u, \quad a, b \in \mathcal{R},\ 0 < \delta \ll 1 \tag{24}$$

is symmetric hyperbolic, fulfills *COND3* and the subcharacteristic condition holds. Nevertheless, when b is large enough, $B - i\omega A$ has an eigenvalue with positive real part, and therefore the system is not stiffly well–posed.

This example shows that the subcharacteristic condition (together with *COND3*) is not sufficient for stiff well–posedness, and Theorem 1.3 does not generalize to higher dimensions without further assumptions.

4.5 A weakly hyperbolic system

Here is an example showing that the assumption of strong hyperbolicity in Theorem 1.3 is essential:

$$u_t + \begin{pmatrix} -1 & 1 \\ 0 & -1 \end{pmatrix} u_x = \frac{1}{\delta} \begin{pmatrix} 0 & 0 \\ 0 & -1 \end{pmatrix} u, \quad 0 < \delta \ll 1. \tag{25}$$

Obviously *COND3* is satisfied and Theorem 1.4 applies, stating that M_0–solutions converge as $\delta \to 0$. Furthermore, the limit is a solution of the corresponding reduced system. Since B is already in block form, the equilibrium equation is

$$\breve{u}_t^I - \breve{u}_x^I = 0.$$

The characteristic speeds of the full system (25) are $\lambda_{1,2} = -1$, and the subcharacteristic condition is fulfilled. Nevertheless, the full system (25) is not well–posed in L_2 since A is not diagonalizable. In particular, the system is not stiffly well–posed.

4.6 The 3×3 example

As was pointed out in the introduction, the original Chen, Levermore and Liu system (3) is not stiffly well-posed, but the situation is quite different for the modified system (6),

$$u_t + \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & -1 & 0 \end{pmatrix} u_x = \frac{1}{\delta} \begin{pmatrix} -2 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} u, \quad 0 < \delta \ll 1.$$

Here *COND2+3* are obviously satisfied. To show stiff well–posedness, we first transform the system by blocking B. Let

$$T = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 2 & 0 & 0 \end{pmatrix} ,$$

i.e., the columns of T are eigenvectors of B. Then we obtain

$$T^{-1}AT = \begin{pmatrix} 0 & -0.5 & 0 \\ -1 & 0 & 1 \\ 0 & 1.5 & 0 \end{pmatrix} =: \tilde{A} , \qquad T^{-1}BT = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -2 \end{pmatrix} =: \tilde{B} ,$$

and the transformed system is

$$v_t + \tilde{A}v_x = \frac{1}{\delta}\tilde{B}v , \quad v = T^{-1}u .$$

The latter system can be symmetrized without changing $\tilde{B}$ as follows: Setting

$$H = \begin{pmatrix} 3 & 0 & 0 \\ 0 & 1.5 & 0 \\ 0 & 0 & 1 \end{pmatrix} ,$$

one obtains that $H\tilde{A}$ is real symmetric and $H\tilde{B} = \tilde{B} \le 0$. Stiff well–posedness follows from Theorem 2.2.

From the transformed system in v–variables and the matrix $\tilde{A}$ we can read off the equilibrium system,

$$\breve{v}^I_t + \begin{pmatrix} 0 & -0.5 \\ -1 & 0 \end{pmatrix} \breve{v}^I_x = 0.$$

As $\delta \to 0$, we have $v \to \breve{v} = (\breve{v}^I, 0)^T$ for $t > 0$, and transforming back to u–variables it follows that

$$u \to \breve{u} = T\breve{v} = (\breve{v}_1, \breve{v}_2, 2\breve{v}_1)^T \quad \text{as} \quad \delta \to 0, \ t > 0 \ .$$

In particular, $\breve{u}_1 = \breve{v}_1, \breve{u}_2 = \breve{v}_2$, which shows that (5) is the equilibrium system in u–variables. We confirm this below in numerical computations.

Let us further note that the characteristic speeds of the equilibrium system are $\mu_{1,2} = \pm\sqrt{2}/2$. The eigenvalues of A — i.e., the characteristic speed of the original system — are $\lambda_{1,3} = \pm\sqrt{2}$, $\lambda_2 = 0$, and we have $\lambda_1 < \mu_1 < \lambda_2 = 0 < \mu_2 < \lambda_3$. Thus, in this example, the subcharacteristic speeds fall strictly between the characteristic speeds of the original system.

4.7 The VBK system

In the book [6], Vasil'eva, Butuzov, and Kalachev discuss an initial–boundary value problem for the system

$$u_t + \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} u_x = \frac{1}{\delta} \begin{pmatrix} -a & b \\ \gamma a & -\gamma b \end{pmatrix} u, \quad 0 < \delta \ll 1 \tag{26}$$

under the assumptions $\lambda_2 > \lambda_1 > 0$, $a > 0$, $\gamma b > 0$. Let us discuss the results of the present paper assuming first only that $\lambda_{1,2}, a, b, \gamma$ are real. The eigenvalues of B are 0 and $-(a + \gamma b)$. Thus $COND2 + 3$ are satisfied if and only if $a + \gamma b > 0$, a condition which we will henceforth assume. Then the transformation of B to diagonal form, $T^{-1}BT = diag(0, -(a + \gamma b))$, yields the equilibrium equation

$$\breve{v}^I_t + \lambda_3 \breve{v}^I_x = 0 \ , \quad \lambda_3 = \frac{\lambda_1 \gamma b + \lambda_2 a}{a + \gamma b} \ .$$

By Theorem 1.3, stiff well–posedness is equivalent to the subcharacteristic condition, which requires here that λ_3 is a convex combination of λ_1 and λ_2.

We read off that this is the case if and only if either $\lambda_1 = \lambda_2$ or ($\lambda_1 \neq \lambda_2$ and a and γb do not have opposite sign).

Let us remark that the importance of the equilibrium system is not restricted to the Cauchy problem; it also plays an important role if initial–boundary data are given. For the system (26), this clearly follows from [6] and the references given there. In general, one must expect occurance of transition layers along the characteristics of the equilibrium system which enter the domain at $t = 0$. (Here we assume that initial data are given at $t = 0$ and boundary data are given for $t \geq 0$.) In future work, we plan to present a general theory of stiff systems (1) under initial and boundary conditions.

5 Numerical experiments

To illustrate the theoretical results obtained in the previous sections, we plot some special solutions of the 3×3 systems (3) and (6). These solutions are computed by a semi–implicit Lax–Friedrichs scheme. Applied to a system of the form $u_t + Au_x = (1/\delta)Bu$ the scheme reads

$$u_j^{n+1} = \frac{1}{2}(u_{j+1}^n + u_{j-1}^n) - \frac{\Delta t}{2\Delta x}A(u_{j+1}^n - u_{j-1}^n) + \frac{\Delta t}{\delta}Bu_j^{n+1}.$$

As usual, u_j^n is regarded as approximation to the average of $u(\cdot, t_n)$ in a computational cell $[x_{j-1/2}, x_{j+1/2})$, $x_j = j\Delta x$, at a discrete time level $t_n = n\Delta t$. In the computations reported here we have used the mesh sizes $\Delta t = 3.5 \cdot 10^{-3}$ and $\Delta x = 5 \cdot 10^{-3}$. Under this conditions the scheme satisfies the CFL–condition

$$\left| \frac{\lambda \Delta t}{\Delta x} \right| \leq 1, \quad \forall \lambda \in \sigma(A) .$$

The computations are performed for $x \in [0, 2]$ and we applied periodic boundary conditions at $x = 0$ and $x = 2$. The initial data is piecewise constant and in equilibrium $u_3 = 2u_1$:

$$u_1(x,0) = \begin{cases} 0.2 & 0 \leq x < 0.75 \\ 0.0 & 0.75 \leq x < 1.25 \\ 0.4 & 1.25 \leq x < 2 \end{cases} , \quad u_2(x,0) = \begin{cases} 0.1 & 0 \leq x < 0.75 \\ 0.0 & 0.75 \leq x < 1.25 \\ -0.1 & 1.25 \leq x < 2 \end{cases}$$

The results for the not stiffly well–posed system (3) are shown in Figure 1. After 20 time steps the third component of the solution obviously blows up. With decreasing δ the solution becomes more and more oscillatory.

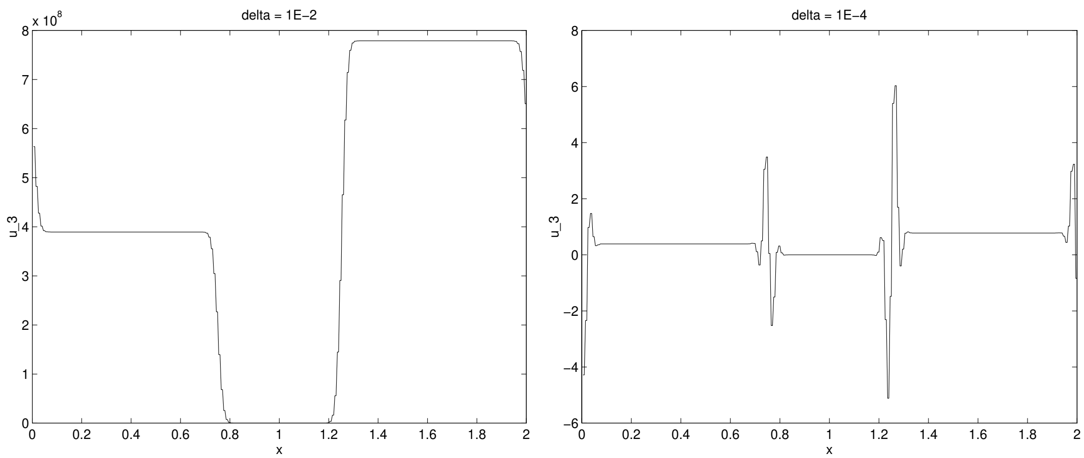

Figure 1: *The third component u_3 of system (3) at $t = 0.07$.*

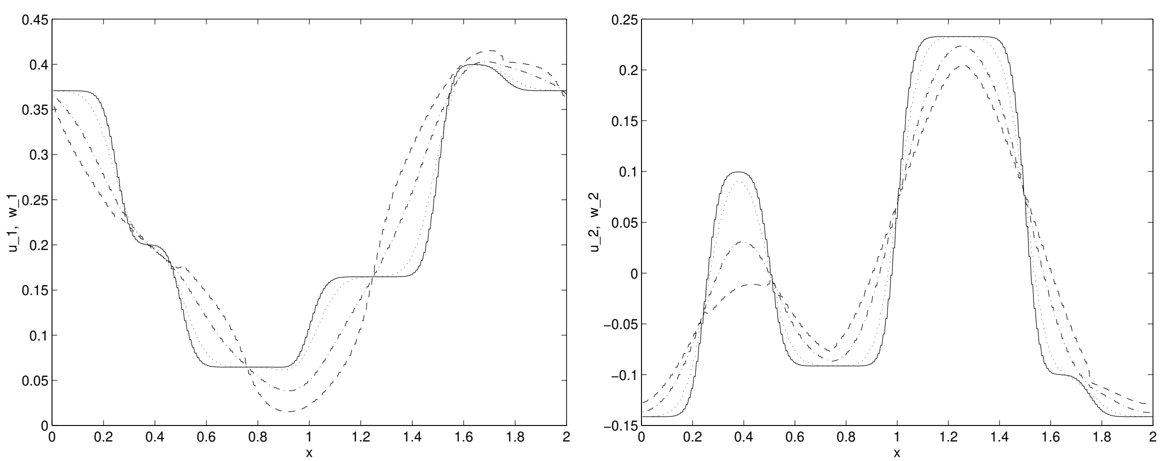

Figure 2: *Solutions of system (6) at $t = 0.35$ for $\delta = 0.1$ $(- - -)$, $\delta = 0.05$ $(- \cdot -)$ and $\delta = 0.01$ $(\cdots)$. The solid line is the solution of the equilibrium system (5).*

For the stiffly well–posed system (6) we computed 100 time steps up to $t = 0.35$. During that time several shock interactions occurred. Nevertheless, we clearly observe convergence towards the solution of the equilibrium system (5). Figure 2 shows the first and second component of u for various values of δ and the corresponding component of the equilibrium solution $\breve{u} = (w_1, w_2, 2w_1)^T$.

6 Appendix

The results given here are known in the linear algebra literature and are included for completeness only.

Lemma 6.1 *Assume the matrix $A \in \mathcal{C}^{n\times n}$ has block form*

$$A = \begin{pmatrix} M_1 & \delta M_{12} \\ \delta M_{21} & M_2 \end{pmatrix}, \quad 0 < \delta \ll 1 \ ,$$

with diagonal blocks $M_1 \in \mathcal{C}^{k\times k}$, $M_2 \in \mathcal{C}^{l\times l}$, $k + l = n$, and small off–diagonal blocks $\delta M_{12} \in \mathcal{C}^{k\times l}$, $\delta M_{21} \in \mathcal{C}^{l\times k}$. If the spectra of M_1 and M_2 are disjoint and δ is small enough, then there exists a transformation $T = I + \mathcal{O}(\delta)$ such that

$$T^{-1}AT = \begin{pmatrix} M_1 + \mathcal{O}(\delta^2) & 0 \\ 0 & M_2 + \mathcal{O}(\delta^2) \end{pmatrix} \ .$$

The transformation T will be constructed in two steps. In the first step the lower off-diagonal block will be eliminated. Here the transformation is determined by a nonlinear equation. The second step, to eliminate the upper nondiagonal block, is a linear argument, which we present first.

Lemma 6.2 *Assume the matrix $B \in \mathcal{C}^{n\times n}$ has block form*

$$B = \begin{pmatrix} M_1 & M_{12} \\ 0 & M_2 \end{pmatrix}, \quad M_1 \in \mathcal{C}^{k\times k}, M_2 \in \mathcal{C}^{l\times l}, M_{12} \in \mathcal{C}^{k\times l}, k + l = n \ .$$

If the spectra of M_1 and M_2 are disjoint, then there is a transformation $T = \begin{pmatrix} I & S \\ 0 & I \end{pmatrix}$ such that

$$T^{-1}BT = \begin{pmatrix} M_1 & 0 \\ 0 & M_2 \end{pmatrix}.$$

Here, $S \in \mathcal{C}^{k \times l}$ is determined as the solution of the regular linear system

$$M_1 S - S M_2 = -M_{12}. \tag{27}$$

Proof: It is sufficient to show that (27) is a regular system for the $k \cdot l$ unknowns $S = (s_{ij})$, $i = 1, 2, \ldots, k$, $j = 1, 2, \ldots, l$. Again, this argument is divided into two steps.

a) First, consider the case where $M_1 = D = diag(d_1, \ldots, d_k)$ and $M_2 = E = diag(e_1, \ldots, e_l)$ are diagonal. Then the homogeneous system $DS - SE = 0$ reads

$$d_i s_{ij} - e_j s_{ij} = (d_i - e_j) s_{ij} = 0$$

for all i and j. Since every eigenvalue of M_1 is distinct from every eigenvalue of M_2, the unique solution is $S = 0$.

b) When M_1 is a full matrix, by Schur's lemma there is a unitary matrix U_1 such that $U_1^* M_1 U_1 = R = (r_{ij})$ is upper triangular with the eigenvalues of M_1 as diagonal entries $r_{ii} = d_i$. Choose $\epsilon > 0$ and set $T_1 = U_1 diag(1, \epsilon, \ldots, \epsilon^{k-1})$, then

$$T_1^{-1} M_1 T_1 = \begin{pmatrix} d_1 & \epsilon r_{12} & \epsilon^2 r_{13} & \cdots \\ & d_2 & \epsilon r_{23} & \cdots \\ & & \ddots & \\ 0 & & & d_k \end{pmatrix} = D + \epsilon \widetilde{D}.$$

Similarly, we have a transformation T_2 such that $T_2^{-1} M_2 T_2 = E + \epsilon \widetilde{E}$. The system $M_1 S - S M_2 = 0$ reads

$$T_1 (D + \epsilon \widetilde{D}) T_1^{-1} S - S T_2 (E + \epsilon \widetilde{E}) T_2^{-1} = 0 \ ,$$

which is equivalent to

$$(D + \epsilon \widetilde{D}) T_1^{-1} S T_2 - T_1^{-1} S T_2 (E + \epsilon \widetilde{E}) = 0. \tag{28}$$

As $D(T_1^{-1} S T_2) - (T_1^{-1} S T_2) E = 0$ is regular by a), so is (28) when ϵ is small. It follows that $T_1^{-1} S T_2 = 0$ and hence $S = 0$. Consequently, (27) is regular and S is well defined.

We are now in a position to prove Lemma 6.1.

Proof of Lemma 6.1: The first step is to find a transformation

$$T_1 = \begin{pmatrix} I & 0 \\ R & I \end{pmatrix}$$

such that

$$T_1^{-1}AT_1 = \begin{pmatrix} M_1 + \delta M_{12}R & \delta M_{12} \\ 0 & M_2 - \delta R M_{12} \end{pmatrix}. \tag{29}$$

The condition for R is

$$RM_1 - M_2R = -\delta RM_{12}R + \delta M_{21}.$$

By Lemma 6.2 the left–hand side is a regular linear operator L acting on R. The right–hand side is a small nonlinearity, and we write

$$LR = \delta F(R), \quad 0 < \delta \ll 1.$$

For small δ the operator $\Phi(R) = \delta L^{-1}F(R)$ is a contraction on the unit ball $D := \{X \in \mathcal{C}^{l\times k} : \|X\| \le 1\}$ and maps D into D. Hence, Φ has a unique fixed point $R \in D$ and $R = \delta L^{-1}F(R) = \mathcal{O}(\delta)$. Therefore,

$$T_1 = \begin{pmatrix} I & 0 \\ R & I \end{pmatrix} = I + \mathcal{O}(\delta)$$

is well defined and $T_1^{-1}AT_1$ has block form (29).

We now apply Lemma 6.2 to $T_1^{-1}AT_1$. Defining S by

$$(M_1 + \delta M_{12}R)S - S(M_2 - \delta RM_{12}) = -\delta M_{12}\ ,$$

we obtain $T_2 = \begin{pmatrix} I & S \\ 0 & I \end{pmatrix} = I + \mathcal{O}(\delta)$. Finally, the product

$$T = T_2T_1 = \begin{pmatrix} I + SR & S \\ R & I \end{pmatrix} = I + \mathcal{O}(\delta)\ ,$$

is the desired transformation.

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 App. Math., XLVII (1994), pp. 787–830.

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

[3] R. J. LeVeque and J. Wang, *A linear hyperbolic system with stiff source terms*, in Notes on Numerical Fuid Mechanics, A. Donato and F. Oliveri, eds., vol. 43, Vieweg, 1992, pp. 401–408.

[4] H. J. Schroll, A. Tveito, and R. Winther, *An L^1–error bound for a semiimplicit difference scheme applied to a stiff system of conservation laws.* Preprint, Dep. of Informatics, Univ. of Oslo, Norway, to appear in SIAM J. Numer. Anal., 1994.

[5] A. Tveito and R. Winther, *On the rate of convergence to equilibrium for a system of conservation laws including a relaxation term.* Preprint, Dep. of Informatics, Univ. of Oslo, Norway, to appear in SIAM J. Math. Anal., 1994.

[6] A. B. Vasil'eva, V. F. Butuzov, and L. V. Kalachev, *The Boundary Function Method for Singular Perturbation Problems*, SIAM Studies in Applied Mathematics, SIAM, Philadelphia, 1995.

[7] G. B. Whitham, *Linear and nonlinear waves*, John Wiley & Sons, New York, 1973.